ICAD

Washington D.C., USA
June 9–15, 2010

Proceedings of the 16th International Conference on Auditory Display

ICAD-10 was organized by
the U.S. Naval Research Laboratory, VRSonic and
the University of Maryland.

ICAD-10 was supported by

NSF

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON DC

NASA

VRSONIC

OFFICE OF NAVAL RESEARCH S&T ONR
Revolutionary Research . . . Relevant Results

Proceedings of the 16th International Conference on Auditory Display Washington D.C., USA

E. Brazil, Irish Centre for High End Computing

June 9–15, 2010

Published by:
International Community for Auditory Display
http://www.icad.org/icad2010/

ISBN: 0-9670904-3-1

Credits: LaTeX's 'confproc' class by V. Verfaille

Conference Co-Chairs:
Derek Brock, U.S. Naval Research Laboratory & Hesham Fouad, VRSonic & Ramani Duraiswami, University of Maryland
Papers/Posters Chair:
Eoin Brazil, Irish Centre for High End Computing
Demonstrations/Compositions Chair:
Evan Rogers, VRSonic
Sonification Chair:
Kelly Snook, NASA
Concert Co-Chairs:
Douglas Boyce, The George Washington University & Katharina Rosenberger, University of California, San Diego
Web Chair:
Brian McClimens, U.S. Naval Research Laboratory
Technical Chair:
Paul Oehlers, American University
Steering Committee:
Stephen Barrass, University of Canberra & Veronique Larcher, Sennheiser & David Worrall, worrall.avatar.com.au

Reviewers:

Alvaro Barbosa, CITAR - Portugal
Soren Bech, Bang and Olufsen
Derek Brock, Naval Research Laboratory
Samuel Ferguson, University of Technology, Sydney
Christopher Frauenberger, University of Sussex
Matti Gröhn, Finnish Institute of Occupational Health
Brian Gygi, VANCHCS
Daniel Hug, Zurich University of the Arts
Trond Lossius, BEK - Bergen Center for Electronic Arts
Roderick Murray-Smith, University of Glasgow
Flaithri Neff, Limerick Institute of Technology
Markus Noisternig, IRCAM
Michael Qin, Naval Submarine Medical Research Laboratory
Tony Stockman, Queen Mary, University of London
Katharina Vogt, University of Music and Dramatic Arts Graz
Katharina Rosenberger, University of California, San Diego

Douglas Boyce, The George Washington University
Eoin Brazil, Irish Centre for High Performance Computing
Milena Droumeva, Simon Fraser University
Federico Fontana, University of Udine
Fabien Gouyon, INESC Porto
Florian Grond, Bielefeld University, CITEC
Thomas Hermann, Bielefeld University
Andy Hunt, University of York, UK
David McGookin, University of Glasgow
Michael Nees, Georgia Institute of Technology
Evan Rogers, VRSonic
Sandra Pauletto, The University of York
Stefania Serafin, Aalborg University
Yon Visell, McGill University
Hesham Fouad, VRSonic
Kelly Snook, NASA

Washington D.C — June 2010

Welcome from Derek Brock, Conference Chair

Welcome to ICAD 2010!

It is a great pleasure to welcome you to Washington, DC and The George Washington University for the 16th International Conference on Auditory Display (ICAD). ICAD was last held in the United States in Boston in 2003, and has met annually since then in Sydney, Limerick, London, Montreal, Paris, and, most recently, in Copenhagen. It is quite an honor to be able to bring ICAD back to the US after a long absence—and to Washington, DC for the very first time!

Our overall program this year has a somewhat novel format. First, we have scheduled not one, but two pre-conference days. Paul Vickers chairs our student consortium known as the "Think Tank" on Wednesday, 9 June, and on Thursday, Kelly Snook, David Worrall, and others hold a full-day, hands-on workshop/hack session devoted to sonification techniques.

The main conference begins on Friday, 11 June and runs through the following Tuesday. We open with a keynote address from Shihab Shamma, and after a full day of paper and aural presentations and our opening reception in the evening, we pause on Saturday and change things up a bit. We meet that morning for posters and demonstrations and take the afternoon off to give attendees time to relax and explore the city. In the evening, we return for the ICAD Concert. This year's event, programmed by our concert co-chairs, Douglas Boyce and Katharina Rosenberger, focuses on the intellectual, artistic, and technical project of sonification and sound spatialization. Emphasizing the human scale of performance to call attention to the question of agency and judgement in both artistic and scientific sonification, the concert will feature performances of commissioned and submitted compositions by the chamber ensemble counter)induction. The remainder of the conference—Sunday, Monday, and Tuesday—begin each day with keynote addresses from Ryoji Ikeda, Douglas Brungart, and Peter Cariani and continue with additional paper and aural presentations.

Our annual ICAD "Open Mic" session will be held on Monday afternoon and that evening we cross the Potomac to the Torpedo Factory in Alexandria, Virginia for our conference banquet. During our Open Mic, we will hear from György Wersényi about next year's conference, which brings ICAD to Budapest in 2011. At dinner, ICAD's Board will present conference awards for best paper, best poster, and best auditory work, and additionally, the Board will honor the recipient of a newly instituted award for outstanding contributions to ICAD, which is proudly named after our founder, Gregory Kramer.

In keeping with our conference theme, "Sonic Discourse—Expression through Sound", a new aural submission category was created this year to provide the community with an opportunity to present and publish extended instances of auditory work that make a contribution to, and/or exemplify, important informational and listening practices in the field of auditory display. Recognizing that systematic aural displays of data may justifiably serve different objectives—specifically, informational purposes versus aesthetic or artistic purposes— work was solicited in two subcategories designated "Sonifications" and "Compositions". Kelly Snook and Evan Rogers handled our reviewing process and the result is a selection of twelve submissions that range in various ways across this expressional spectrum. Work in the Sonification category is spread throughout the conference's four days of technical sessions and Compositional work is presented in Saturday's concert. It is our sincere hope that this new aural class of peer-reviewed conference submissions will be actively embraced by the community and become an established and relevant form of publication at future conferences.

Spatialization is a critical informational dimension for many classes of auditory display and this year we have the exceptionally good fortune to have partnered with VRSonic who, in addition to time, staff, and technical expertise, has provided the conference with free access to their state-of-the-art VibeStudio audio design suite and a 7.1 sound system for use during our technical sessions and in our concert.

We would also like to express our deep appreciation for generous grants of financial support from the National Science Foundation for the Think Tank, the Office of Naval Research for assistance with our facilities costs, and the National Aeronautics and Space Administration for a range of additional expenses.

Finally, together with our Paper and Posters Chair, Eoin Brazil, our Sonification Chair, Kelly Snook, and our Demonstrations and Compositions Chair, Evan Rogers, we would like to express a profound note of appreciation for this year's distinguished review committee. Reviewing is often taken for granted, but without this crucial participatory dimension in our community, ICAD would be a far less important and vital institution.

Once again, welcome to Washington, DC! On behalf of our entire organizing committee, we earnestly hope you enjoy the exciting program we have assembled and the opportunity to exchange ideas and spend time with your colleagues!

Derek Brock, Hesham Fouad, and Ramani Duraiswami

9-15 June 2010 Conference Co-Chairs

Conference Program

Day 1

Session 1 Accessibility

Session 2 Applications

Aural presentation 1

Session 3 3D and Spatial Audio I

Session 4 3D and Spatial Audio II

Aural presentation 2

Day 2

Session 5 Posters

Session 6 Demos

Compositions

Day 3

Session 7 Mapping and Sonification

Session 8 Cognition

Aural presentation 3

Session 9 Auditory Menus

Session 10 Usability

Aural presentation 4

Day 4

Session 11 Design

Session 12 Aesthetics, Philosophy, and Culture of Auditory Displays

Aural presentation 5

Day 5

Session 13 Auditory Scene Design

Session 14 Interaction I

Aural presentation 6

Session 15 Interaction II

Aural presentation 7

TANGIBLE ACTIVE OBJECTS AND INTERACTIVE SONIFICATION AS A SCATTER PLOT ALTERNATIVE FOR THE VISUALLY IMPAIRED

Eckard Riedenklau, Thomas Hermann, Helge Ritter

Ambient Intelligence Group / Neuroinformatics Group
CITEC – Center of Excellence in “Cognitive Interaction Technology”
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
[eriedenk,thermann,helge]@techfak.uni-bielefeld.de

ABSTRACT

In this paper we present an approach that enables visually impaired people to explore multivariate data through scatter plots. Our approach combines Tangible Active Objects (TAOs) [1] and *Interactive Sonification* [2] into a non-visual multi-modal data exploration interface and thereby translates the visual experience of scatter plots into the audio-haptic domain. Our system and the developed sonification techniques are explained in this paper and a first user study is presented.

Index Terms— Exploratory Data Analysis, Scatter Plots, Sonification, Tangible Active Objects, Tangible Interaction

1. INTRODUCTION

Nowadays in the information age, data visualizations are omnipresent. Not only scientists use visualizations, also in everyday life visualizations become more and more important to convey complex information. A variety of visualization techniques have been developed to provide possibilities to explore and understand all kinds of data. Many of these methods address our visual modality, e.g. graphs, diagrams, and plots. Obviously, the visually impaired cannot use these visualization techniques. Therefore different methods have to be developed to allow alternative data exploration, e.g. by using auditory or haptic displays.

Visualization refers not only to visual data representations, but more generally to an abstract concept to convey information about data in different modalities or even in multi-modal manner. Thereby also haptic, auditory approaches and methods using other modalities or combinations of them are visualizations, too. Generally visualization comprehends methods that create a mental representation of the visualized data in the experiencing persons mind [3]. However “perceptionalization” is a term that makes this breadth more obvious.

Because haptic and auditory display methods require a longer time for the user to actively acquire an overview of the represented data, interaction with the data representation system is important. The possibility of actively changing the rendering parameters through interaction allows the users to get an all-embracing overview of the underlying data. Interacting with data representations allows users to explore the kind and characteristics of data. Tangible User Interfaces (TUIs) specifically use physical objects as representatives to give the user that data at hand. This allows to explore the data with our everyday manual interaction skills in a bi-manual and parallel manner. Our Tangible Desk (tDesk) (formerly known as the Gesture Desk [4]) is a typical table-top TUI. In different applications, such as AudioDB [5], TI-Son [6], Tangible Data Scanning (TDS) [7], and AmbiD [8], etc. passive Tangible User Interface Objects (TUIOs) were used. The TUIOs, used in these applications are optically tracked by a camera. By interacting with these TUIOs, the user can directly interact with the represented data.

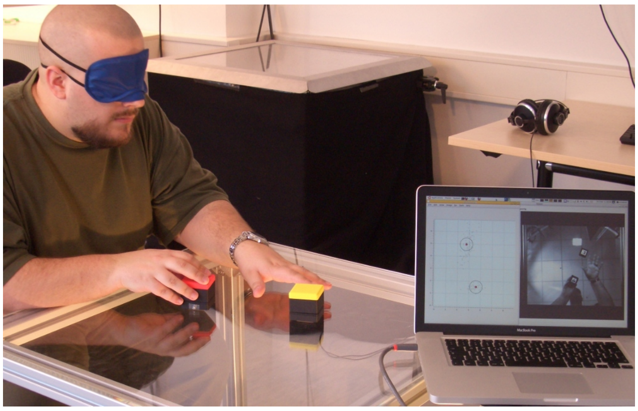


Figure 1: Blindfolded user interacting with the IAS at the tDesk. The laptop shows the computer vision module and the clustering of the data (here the Iris Dataset [21] was clustered in two clusters) with a TAO for each cluster prototype.

We here start with such a Tangible Interaction system and extend it with active feedback capabilities, allowing the objects to move actively on the table surface. In addition we use Interactive Sonification [2], defined “as the use of sound within a tightly closed human–computer interface where the auditory signal provides information about data under analysis, or about the interaction itself, which is useful for refining the activity.” In our approach, we combined TAOs and Interactive Sonification to create a novel data exploration interface for the visually impaired.

Compared to the research field of visual data analysis techniques, the field of non-visual techniques for multivariate data analysis is still quite sparse. Most multivariate data are collected in tables of numbers which are often visualized using scatter plots.

This work has been supported by the Cluster of Excellence 277 Cognitive Interaction Technology funded in the framework of the German Excellence Initiative

1.1. State of the art

We briefly review different non-visual methods to represent data, focusing on haptic, auditory and combined perceptionalization approaches. The Sonic Scatter Plots [9] are an approach to sonify multivariate data. Since “notes can also be plotted in time”, here the data are seen as a score, where one dimension is scaled and interpreted as time. Then every data-point is interpreted as one note and the different characteristics of the notes are controlled by data values. The work of Sarah Bly [10] and John Flowers [11] on multivariate data mappings and auditory scatter plots are seminal to the topic. Another related application is the TDS [7]. “A sonification model following the Modelbased Sonification approach that allows to scan high-dimensional data distributions by means of a physical object in the hand of the user” is developed in this paper. Therefore a virtual plane is linked to one TUIO, that can be moved through the data space. A Model-Based Sonification (MBS) is triggered each time the plane crosses a data point, whereby the user can interactively explore the data distribution.

Panëels and Roberts [12] provide a comprehensive review of designs for Haptic Data Visualization (HDV). They propose a taxonomy of seven categories: Charts, Maps, Signs, Networks, Diagrams, Images and Tables. Unfortunately the authors did not find a haptic translation for scatter plots, but they present techniques, that could be used for scatter plots, as well. Interpreting scatter plots as height-fields, e.g. the Nanomanipulator [13] could be adapted. Further more image translation techniques, such as proposed in [14] can be used to transfer scatter plots into the haptic domain.

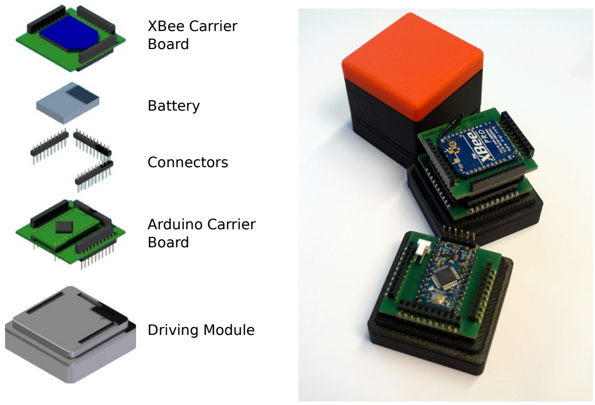
We introduce the combination of TAOs and Interactive Sonification as an alternative to visual scatter plots. The paper is organized as followed: We give an introduction to the TAOs, our novel tangible interfaces used in our system. We describe two different new sonification approaches and their implementation. Furthermore we present the system design, the hardware components and basic ways of interaction. A short study and evaluation is presented, followed by the discussion of results and our conclusion.

2. TANGIBLE ACTIVE OBJECTS

Most Tangible User Interfaces (TUIs) use passive objects that offer no active feedback. Active feedback refers to the ability to actively influence the interaction by e.g. changing the object’s position or orientation, or multi-modal feedback via the haptic, auditory, or visual modality, etc. In our present work we develop a swarm of TAOs, capable of different kinds of feedbacks [1].

2.1. Hardware

The hardware assembly of our Tangible Active Objects (TAOs) is depicted in Fig. 2. The TAOs are built in different modular Printed Circuit Boards (PCBs) that fit into custom layers, compatible to TUImod [15], modular building blocks for TUIOs. These PCBs are connected over simple vertical buses to make them flexibly extensible. Our current TAOs are configured as small mobile robots, but it is also possible, to equip them with a display, buttons, or loud speakers. The mobile configuration consists of a driving module, the control module with additional connectors and batteries, and a wireless communication module. The driving module is a simple differential drive, which means that two motors drive two wheels on the same rotational axis independently. Thereby it is possible to drive the TAO continuously from in-place rotation to straight forward or backward linear movement. The control module is the core of each TAO. It holds an Arduino pro mini [16], a community-based rapid prototyping microcontroller platform, often used for physical computing approaches. This microcontroller is programmed with the SerialControl firmware [17], which allows to receive commands over the wireless communication module and to control the connected in- and output componentes, in this case the driving module. The wireless communication module is based on an XBee module, which is configured to work in a star-network together with other modules. One XBee module is serially connected to the host computer and spreads the remote control commands into the network. Each TAO has its own ID and only reacts on commands starting with this ID. This allows to control each TAO independently. Additionally, each TAO is equipped with a visual fiducial marker, which is based on the tracking algorithm of the Reactable [18], but we use a new marker set which was especially designed for the TAOs.



(a) Exploded assembly drawing (b) Manufactured devices at different stages of assembly (without batteries)

Figure 2: Hardware architecture

2.2. Software architecture

The TAOs are remote controlled by a host computer, where a dynamically extensible software framework controls each TAO and the sonification of the IAS. Fig. 3 depicts the different cooperating modules. All modules are implemented as stand-alone processes which communicate over the XML enabled Communication Framework (XCF) [19]. The Computer Vision module (1) tracks the TAOs’ position and orientation in the camera image. These data are spread into an ActiveMemory server, which is part of the XCF and runs transparently in the background. Several other modules subscribe on the content of this information stream, such as the Path Planner module (2) and the IAS application (3). The Path Planner module reacts also to XCF messages that invoke navigation tasks. Based on the position and orientation of the TAOs and the new target positions from the navigation query, the Path Planner calculates trajectories for each moving TAO, based on a potential field approach [20] and transmits control commands that make the TAOs move to the new targets. To transmit these commands, the XCF2Serial module (4) listens for these commands in the XCF stream and relays them to the serial port of the host computer where an XBee module (5) broadcasts the commands to the wireless network and to the TAOs (6).

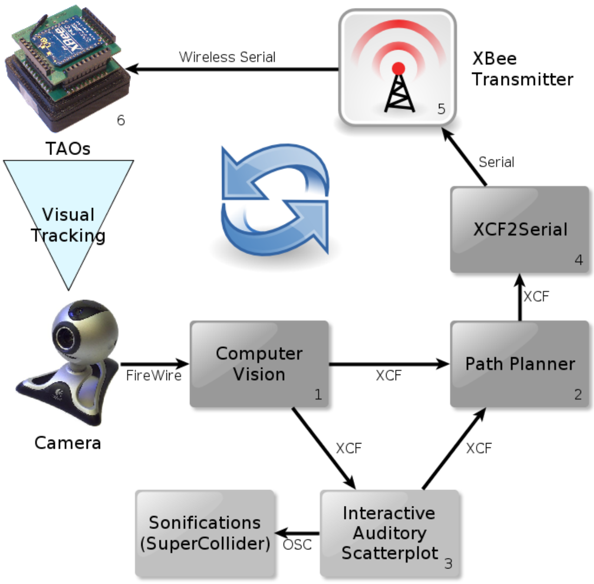


Figure 3: Software architecture. Modules (1) - (6) are explained in 2.2

3. DESIGN: INTERACTIVE AUDITORY SCATTER PLOT

Clustering is the most basic structure of data distributions. Clusters are groupings of data at certain locations in data space as shown in Fig. 4. The TAO-based Interactive Auditory Scatter Plot (IAS) is currently designed to enable and assist the understanding of clustering structure without the need of any visual display. It is implemented as a special application module for the TAO architecture.

3.1. Ideas and concepts

We created a direct two-dimensional transformation of the spatially distributed data into the audio-haptic domain to allow visually impaired people the exploration of scatter plots. Since we have a limited number of K TAOs, which should become graspable anchors of the data, we first need a method to find K representative locations. Vector quantization with the K-Means algorithm is an appropriate starting point for this.The TAOs move autonomously on the table to the cluster centers and represent them as prototype objects. Thereby the visually impaired user is enabled to explore roughly how the data is organized. Each TAO can then be used to examine how the data are distributed in detail: By moving a TAO, a sonification is excited that perceptualizes local characteristics of the data distribution. Furthermore, releasing an object triggers a local data sonogram after which the object moves back to its anchor position, as explained in detail in Sec. 4. During interaction, the user can thereby construct a mental model of the spatial data distribution and clustering structure.

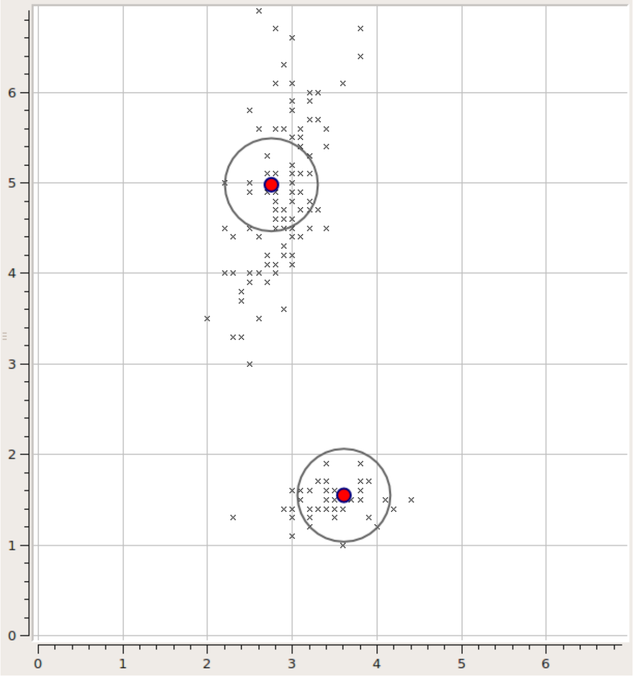


Figure 4: Two TAOs representing prototypes of clusters (detail screenshot from the application module). Here the Iris Dataset [21] was used for the visualization.

3.2. Hardware setup

The hardware basis of our system is the tDesk, which is a redesign of the Gesture Desk, introduced in [4]. On a 70 × 70 × 70 cm cube of aluminum profiles lies a glass surface. A FireWire camera which is mounted underneath this surface looks upwards to the table's surface to track the TAOs position on the table. The tracking area on the table is marked with black tape. The speakers with its sound interface and the camera are connected to the host computer together with the XBee transmitter for the wireless communication.

3.3. Interaction design

As depicted in Fig. 5, we subdivide interaction into three different parts: The first part uses the TAOs as haptic representations of data clusters. By touching the table surface and the TAOs, the user can gain a rough idea of where interesting data are located. The user can then move any TAO, which excites a sonification of the local density and can thereby continuously explore how the data is distributed in the vicinity of a cluster. The third interaction type is to release a TAO. This triggers a local data sonogram, yielding an audible spherical sweep through the data space at the location of the TAO. Afterwards the TAO moves back to its cluster center.

4. SONIFICATION METHODS

We integrated two different sonification approaches into our system using SuperCollider for the implementation [22]. The first is a parameter mapping-based approach, where the local density of the data is mapped directly to parameters of a continuous sonic stream.

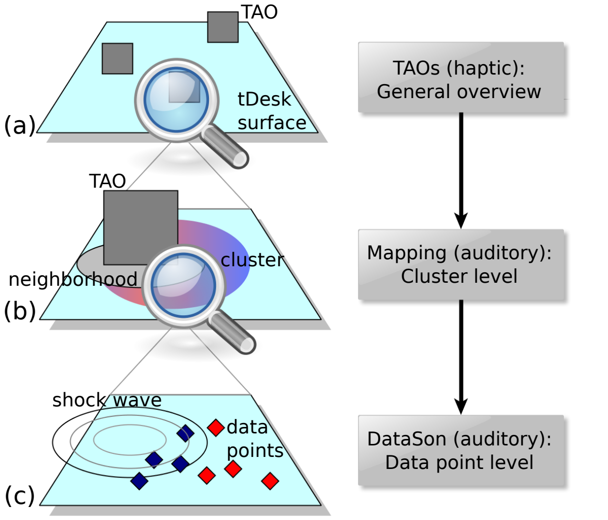

TAO
tDesk surface
(a)
TAO
neighborhood
cluster
(b)
shock wave
data points
(c)
TAOs (haptic): General overview
Mapping (auditory): Cluster level
DataSon (auditory): Data point level

Figure 5: Levels of detail in the IAS (The stylized cluster and data points are only depicted for better understanding of the picture; They are not visible to the user of our system.)

The second approach uses a Model-Based Sonification (MBS) approach to communicate more detailed characteristics of the underlying data [23, 24, 25].

4.1. Parameter mapping-based sonification for IAS

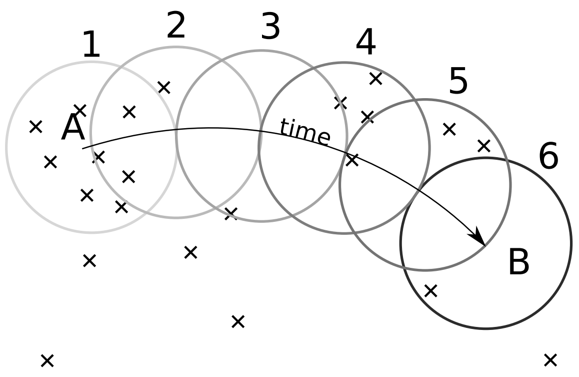


Figure 6: Mapping: The circles depict the neighborhood of TAOs. Moving a TAO from A to B results in a continuous sonification of the data density as pitch of a continuous sound stream.

In our first sonification approach, a simple mapping of the local data density controls a continuous sonic stream. When moving the TAO at position $\vec{x}$, the number of data points N in the neighborhood of an adjustable radius r around the TAO is mapped to the frequency of an additive synthesis using

$$f[Hz] = f_0 \cdot 2^{\alpha N(\vec{x},r)}. \qquad (1)$$

This leads to a pitch increase of an octave if $N \rightarrow N \cdot \frac{1}{\alpha}$. Fig. 6 explains this simple mapping approach. This sonification is automatically activated whenever a TAO is moved by the user. The sound is generated at constant amplitude. At the moment of releasing the TAO, the data sonogram sonification is triggered as explained next.

4.2. Local data sonograms

Releasing a TAO after moving it around excites a local data sonogram [23] at the TAOs location to provides a detailed inspection into the spatial data distribution. For this a virtual 'shock wave' emanates from the the TAO's location to the border of the neighborhood. Whenever this wave crosses a data point, a virtual spring connected to the data point is excited to oscillate, which generates audible sound, as depicted in Fig. 7. This local data sonogram approach was introduced in [23] and generalized to Multi-touch interactions in [26]. However multi-touch enabled visual display are unfortunately unsuitable for the visually impaired so that our extension to graspable interfaces makes data sonograms for the first time usable for visually impaired users.

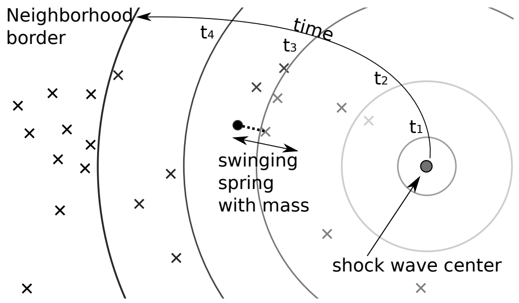


Figure 7: Data sonograms: A virtual shock wave is evoked at the location where the TAO is released and expands in circles until it reaches the border of the TAO's neighborhood. Data points are excited by the shock wave front and thereby contribute to a spacial sweep.

5. INTERACTION EXAMPLES AND FIRST EXPERIMENTS

A basic demonstration of our system is provided at our website[1]. The introduction video shows a user interacting with the system and presenting each of the three interaction stages. The video shows that the system basically works as intended. Our interpretation is that this system is well capable of allowing visually impaired or blindfolded people to understand scatter plots without seeing. In the paper we present results of a first qualitative and quantitative study.

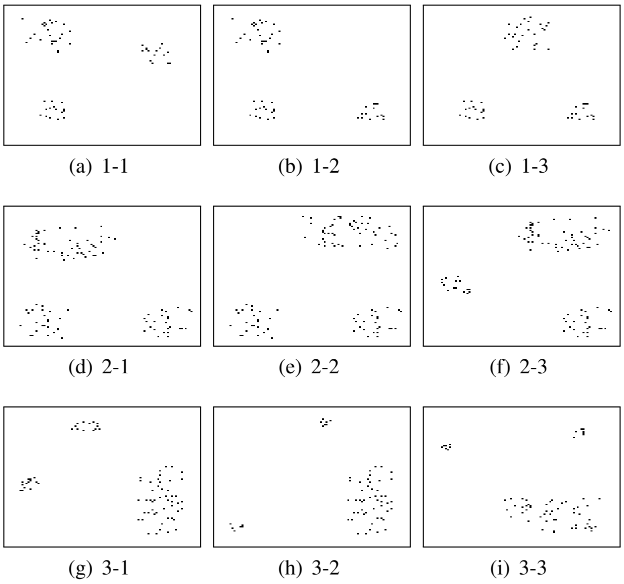
5.1. First experiment: the blind herder

In our first study we wanted to learn in how far our approach can be used as an alternative to the classical scatter plot. Basically we wanted to know if IAS users are able to recognize the same visual plot from a list of slightly different candidates. This can be tested by showing to the users the pictures of different classical scatter plots including the one, the user just explored with the IAS and asking to choose the one they think they just have explored.

[1] see `http://www.techfak.uni-bielefeld.de/ags/ami/publications/RHR2010-TAO/` for the video

Because not all subjects were familiar with the concept of scatter plots and clusters, we created a metaphoric story called “the blind herder”: The subjects were told to think of the tDesk’s surface as the grazing land of three herds of animals. The subjects had to discover the distribution of the animals by interacting with the system.

The subjects have to complete different tasks with the system. In the first task, the sighted but blindfolded subjects have to explore a certain IAS until they subjectively have acquired an understanding of the data distribution. For this task there is no time limit, but the time needed is recorded for later evaluation. As test distributions simple synthetic datasets were created that all show three easily distinguishable clusters of data points at different positions, and of different shapes and sizes. Three sets with three datasets were generated (see Fig. 8). Every trial one row with three datasets is chosen randomly. One dataset is randomly chosen from the row and presented in the IAS. Before the subjects are asked to select the corresponding visualized classical scatter plot from a set of three different versions shown, they had to sketch their mental image of the data distribution with pen and paper.



(a) 1-1 (b) 1-2 (c) 1-3

(d) 2-1 (e) 2-2 (f) 2-3

(g) 3-1 (h) 3-2 (i) 3-3

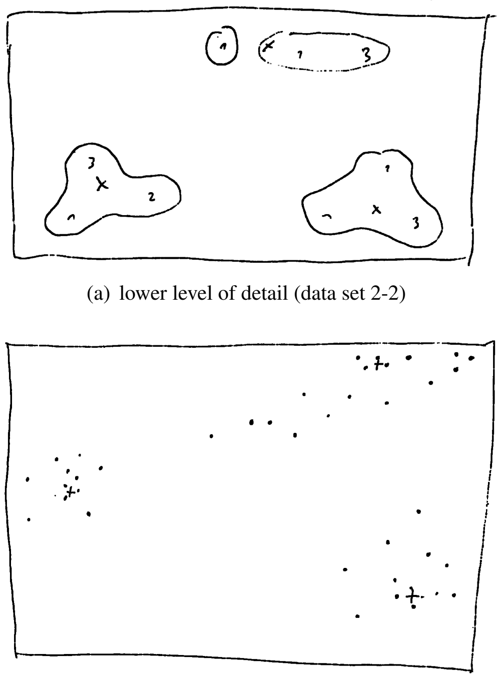
Figure 8: Synthetic datasets used in the study.

5.2. Subjects

Nine untrained subjects participated in this first study and performed 13 trials. One subject conducted three trials, two conducted two trials, all other subjects were tested only once. The age of the subjects ranged from 24 to 67, but most of the subjects were younger, so that the mean age of the subjects was 33. Most of the younger subjects are students.

5.3. First results an observations

77% of the trials were successful, 23% were not. 67% of the subjects were able to successfully recognize the explored dataset, 28% were not. As depicted in Fig. 11 the subjects were allowed to explore an IAS as long as they wanted to. The duration ranges from 51 seconds to 18 minutes for a single plot. In the mean every subject used about 9 minutes. Furthermore the figure shows that even the short period of 51 seconds was enough to recognize the explored dataset in the plots. On the other hand the subjects that took much more time were able to tell quite impressively how dense the data points were distributed, including holes in the clusters (see datasets 2-1 to 2-3) and single data points at the border of the clusters. Fig.9 depicts two hand-drawn examples of what the subjects had explored. The level of detail ranged from single closed curves representing the border of the clusters to very detailed pictures with single data points and the density of their distribution.



(a) lower level of detail (data set 2-2)

(b) high level of detail (data set 2-3)

Figure 9: Examples of subject’s hand drawn plots of the explored data sets. Corresponding underlying datasets are plotted in Fig. 8

Finally the subjects are asked to answer a questionnaire. All subjects stated that they can work with the system and that they think that practice can improve the understanding of IASs. Most of the subjects answered that working with the system is fun. The parameter mapping-based sonification was regarded as useful by all subjects, where as the data sonograms seemed to be much harder to understand. Only two subjects found this sonification element useful, four found it partly useful and seven subjects were not able to understand it. This may be for technical problems or inefficient explanation. A technical problem was the lag between the tracking and the sonification output and sudden jumps of the tracked markers, caused by the users hands or bright colored clothes in the camera image. Also a differing position of the subject in front of the tDesk was problematic in one case.

After filling out the questionnaire, the subjects had the opportunity to freely state what their impression of the system was and which strategy they used to orientate. One subject was surprised how easy it was to grasp the TAOs without seeing them. A good spatial imagination was regarded as very helpful. The tape bordering around the tracked area was often used to measure distances to the border of the interaction area (see Fig. 10).

After a short phase of getting used to the system, many subjects developed individual and interesting strategies to discover the borders of clusters. By scanning the cluster with the clusters TAO vertically and horizontally, a first rough idea of the clusters size and shape was gathered. some subjects also tried to trace the border of the cluster by moving the cluster's TAO over the border in zig-zag-patterns.

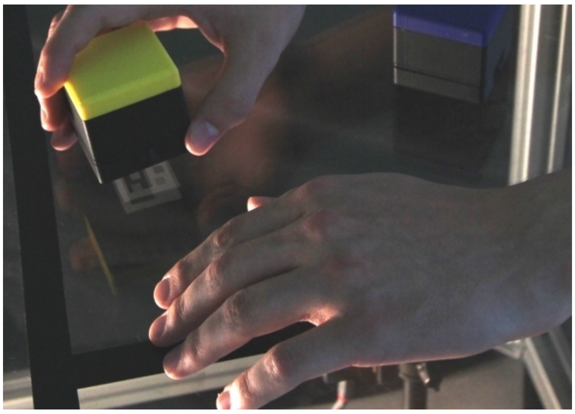


Figure 10: Moving a TAO with the right hand and staying in touch with the taped border with the left hand simultaneously for better orientation.

6. CONCLUSION

In this paper we presented a novel approach for combined auditory and haptic interactive rendering of scatter plots. Through Interactive Sonification and TAOs, it was possible to create a rich exploratory data analysis interface for the visually impaired. As a novel contribution we introduced a hybrid (subdivided) interaction schema where continuous density sonification and a model-based sonification using data sonograms are tightly interwoven to create a rich repertoire for exploratory interactions to create a rich multi-modal user interface. This system was evaluated in a first user study and was proved successfully to enable non-visual exploration of scatter plots.

6.1. Future developments

This application is still under active development, further additions are considered for future developments. For a multi-modal exploratory data analysis interface that can be used both by sighted and visually impaired users simultaneously, we plan to overlay the audio-haptic rendering interface with a visual projection of the scatter plot.

The next step is to extend this case study to a empirical study. Here we want to analyze how well users perform in specific tasks, such as defining different cluster characteristics, e.g. cluster size, density, and shape, etc. We also consider experiments for visually impaired people, since their spatial imagination may differ from those of sighted people. Furthermore we plan to spatialize the data sonograms to enrich the display and the interaction further and to enhance the multi-modal rendering. In summary, the IAS opens attractive new interaction steps and auditory inspection metaphors to support navigation and examination of scatter plots, particularly for visually impaired users.

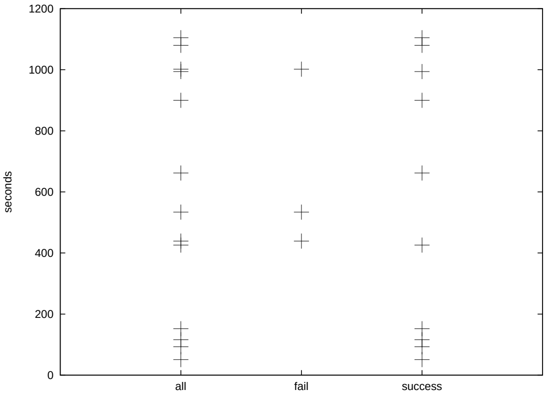


Figure 11: Results: exploration time (one cross for each single trial) in seconds for all trials (left), failed (middle) and succeded (right)

7. REFERENCES

[1] E. Riedenklau, "TAOs - Tangible Active Objects for Tabletop Interaction," Diplomarbeit, Bielefeld University, Bielefeld, Germany, June 2009.

[2] T. Hermann and A. Hunt, "An introduction to interactive sonification," *IEEE multimedia*, vol. 12, no. 2, pp. 20–24, 2005.

[3] J. C. Roberts, "Visualization display models-ways to classify visual representations," *Int. J. of Computer Integrated Design and Construction*, pp. 1–10, 2000.

[4] T. Hermann, T. Henning, and H. Ritter, "Gesture Desk - An Integrated Multi-modal Gestural Workplace for Sonification," in *Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop, GW 2003 Genova, Italy, April 15-17, 2003, Selected Revised Papers*, ser. Lecture Notes in Computer Science, A. Camurri and G. Volpe, Eds., vol. 2915/2004, Gesture Workshop. Berlin, Heidelberg: Springer, 2004, pp. 369–379.

[5] T. Bovermann, C. Elbrechter, T. Hermann, and H. Ritter, "AudioDB: Get in Touch with Sounds," in *Proc. of the Int. Conf. on Auditory Display 2008*, 2008.

[6] T. Hermann, T. Bovermann, E. Riedenklau, and H. Ritter, "Tangible Computing for Iinteractive Sonification of Multivariate Data," *Proceedings of the 2nd International Workshop on Interactive Sonification, York, UK February 3, 2007*, 2007.

[7] T. Bovermann, T. Hermann, and H. Ritter, "Tangible Data Scanning Sonification Model," in *Proceedings of the International Conference on Auditory Display (ICAD 2006)*,

T. Stockman, Ed., International Community for Auditory Display (ICAD). London, UK: Department of Computer Science, Queen Mary, University of London, 06 2006, pp. 77–82.

[8] T. Bovermann, T. Hermann, and h. Ritter, "A Tangible Environment for Ambient Data Representation," in *First International Workshop on Haptic and Audio Interaction Design*, D. McGookin and S. Brewster, Eds., vol. 2. www.multivis.org, 08 2006, pp. 26–30.

[9] T. Madhyastha and D. Reed, "A framework for sonification design," in *Santa Fe Institure Studies In The Sciences Of Complexity-Proceedings Volume-*, vol. 18. Addison-Wesley Publishing Co, 1994, pp. 267–267.

[10] G. H. Kramer, *Auditory display*, ser. Santa Fe Institute studies in the sciences of complexity : Proceedings ; 18. Addison-Wesley, 1994.

[11] J. Flowers, D. Buhman, and K. Turnage, "Cross-modal equivalence of visual and auditory scatterplots for exploring bivariate data samples," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 39, no. 3, pp. 341–351, 1997.

[12] S. Panëels and J. C. Roberts, "Review of designs for haptic data visualization," *IEEE Transactions on Haptics*, vol. 99, no. PrePrints, 2009.

[13] R. M. Taylor, W. Robinett, V. L. Chi, F. P. Brooks, Jr., W. V. Wright, R. S. Williams, and E. J. Snyder, "The nanomanipulator: a virtual-reality interface for a scanning tunneling microscope," in *SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM, 1993, pp. 127–134.

[14] T. P. Way and K. E. Barner, "Automatic visual to tactile translation. I. Human factors, access methods and image manipulation," *Rehabilitation Engineering, IEEE Transactions on*, vol. 5, no. 1, pp. 81–94, March 1997.

[15] T. Bovermann, R. Koiva, T. Hermann, and H. Ritter, "TUImod: Modular objects for tangible user interfaces," in *Proceedings of the 2008 Conference on Pervasive Computing*, 2008.

[16] "Arduino - ArduinoBoardProMini." [Online]. Available: http://www.arduino.cc/en/Main/ArduinoBoardProMini

[17] "Arduino playground - SerialControl." [Online]. Available: http://www.arduino.cc/playground/Code/SerialControl

[18] S. Jorda, M. Kaltenbrunner, G. Geiger, and R. Bencina, "The reactable*," in *Proceedings of the International Computer Music Conference (ICMC 2005), Barcelona, Spain*, 2005, pp. 579–582.

[19] J. Fritsch and S. Wrede, *An Integration Framework for Developing Interactive Robots*, ser. Springer Tracts in Advanced Robotics, D. Brugali, Ed. Berlin: Springer, 2007, vol. 30.

[20] J.-C. Latombe, *Robot motion planning*, 3rd ed., ser. The Kluwer international series in engineering and computer s ci. Boston [u.a.]: Kluwer Acad. Publ., 1993.

[21] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://archive.ics.uci.edu/ml/index.html

[22] "SuperCollider - real-time audio synthesis and algorithmic composition." [Online]. Available: http://supercollider.sourceforge.net/

[23] T. Hermann and H. Ritter, "Listen to your Data: Model-Based Sonification for Data Analysis," in *Advances in intelligent computing and multimedia systems*, G. E. Lasker, Ed. Baden-Baden, Germany: Int. Inst. for Advanced Studies in System research and cybernetics, 08 1999, pp. 189–194.

[24] T. Hermann, "Sound and Meaning in Auditory Display," 09 2001, position statement in Proc. Int. Workshop on Supervision and Control in Engineering and Music, Kassel.

[25] T. Hermann, "Sonification for Exploratory Data Analysis," Ph.D. dissertation, Bielefeld University, Bielefeld, Germany, 02 2002.

[26] R. Tünnermann and T. Hermann, "Multi-touch interactions for model-based sonification," M. Aramaki, R. Kronland-Martinet, S. Ystad, and K. Jensen, Eds., Re:New – Digital Arts Forum. Copenhagen, Denmark: Re:New – Digital Arts Forum, 2009.

SONIFICATION OF COLOR AND DEPTH IN A MOBILITY AID FOR BLIND PEOPLE

Guido Bologna

University of Geneva,
Computer Science Department,
Route de Drize 7, 1227 Carouge,
Switzerland
Guido.Bologna@unige.ch

Benoît Deville

University of Geneva,
Computer Science Department,
Route de Drize 7, 1227
Carouge, Switzerland
Benoit.Deville@unige.ch

Thierry Pun

University of Geneva,
Computer Science Department,
Route de Drize 7, 1227
Carouge, Switzerland
Thierry.Pun@unige.ch

ABSTRACT

The *See Color* interface transforms a small portion of a colored video image into sound sources represented by spatialized musical instruments. Basically, the conversion of colors into sounds is achieved by quantization of the HSL color system. Our purpose is to provide visually impaired individuals with a capability of perception of the environment in real time. In this work the novelty is the simultaneous sonification of color and depth, depth being coded by sound rhythm. Our sonification model is illustrated by several experiments, such as: (1) detecting an open door in order to go out from the office; (2) walking in a hallway and looking for a blue cabinet; (3) walking in a hallway and looking for a red tee shirt; (4) moving outside and avoiding a parked car. Videos with sounds of experiments are available on *http://www.youtube.com/guidobologna.*

1. INTRODUCTION

This paper presents on-going work of the See ColOr project, which aims at improving the perception and the mobility of blind individuals. In previous works we introduced the sonification of colors by means of instrument sounds, as well as experiments related to image comprehension, recognition of colored objects and mobility [1].

In this work the novelty is the use of depth in the sound code. Generally, distance to objects is a crucial parameter for an individual evolving in a given environment. In our experiments, depth is captured by a stereoscopic camera. We perform experiments for which a well trained blindfolded individual takes advantage of depth sonification in a number of situations like: (1) detecting an open door in order to go in and out; (2) walking in a hallway and looking for a blue cabinet; (3) walking in a hallway and looking for a red tee shirt; (4) moving outside and avoiding a parked car. The resultant videos are available on *http://www.youtube.com/guidobologna*. In the following sections, section 2 describes mobility aids without the sonification of color, section 3 presents several works that proposed color sonification, section 4 summarizes our previous experiments, section 5 explains our model of color and depth sonification, section 6 illustrates several experiments, followed by the conclusion.

2. MOBILITY AIDS WITHOUT COLOR SONIFICATION

Several authors proposed special devices for visual substitution by the auditory pathway in the context of real time navigation. The “K Sonar-Cane” combines a cane and a torch with ultrasounds [2]. Note that with this special cane, it is possible to perceive the environment by listening to a sound coding depth.

“TheVoice” is another experimental vision substitution system that uses auditory feedback. An image is represented by 64 columns of 64 pixels [3]. Every image is processed from left to right and each column is listened to for about 15 ms. In particular, every pixel gray level in a column is represented by a sinusoidal wave sound with a distinct frequency. High frequencies are at the top of the column and low frequencies are at the bottom.

Capelle et al. proposed the implementation of a crude model of the primary visual system [4]. The implemented device provides two resolution levels corresponding to an artificial central retina and an artificial peripheral retina, as in the real visual system. The auditory representation of an image is similar to that used in “TheVoice” with distinct sinusoidal waves for each pixel in a column and each column being presented sequentially to the listener.

Gonzalez-Mora et al. developed a prototype using the spatialisation of sound in the three dimensional space [5]. The sound is perceived as coming from somewhere in front of the user by means of head related transfer functions (HRTFs). The first device they achieved was capable of producing a virtual acoustic space of 17*9*8 gray level pixels covering a distance of up to 4.5 meters.

3. COLOR SONIFICATION

3.1. State of the art

Recently, the research domain of color sonification has started to grow [6], [7], [8]. A number of authors defined sound/color associations with respect to the HSL color system. HSL (Hue, Saturation, Luminosity) is a symmetric double cone symmetrical to lightness and darkness. HSL mimics the painter way of thinking with the use of a painter tablet for adjusting the purity of colors. The H variable represents hue from red to purple (red, orange, yellow, green, cyan, blue, purple), the second one is saturation, which represents the purity of the related color and the third variable represents luminosity. The H, S, and L variables are defined between 0 and 1.

Doel defined color/sound associations based on the HSL color system [6]. In this sonification model, sound depends on the color of the image at a particular location, as well as the speed of the pointer motion. Sound generation is achieved by subtractive synthesis. Specifically, the sound for grayscale colors is produced by filtering a white noise source with a low pass filter with a cutoff frequency that depends on the brightness. Color is added by a second filter, which is parameterized by hue and saturation.

Rossi et al. presented the "Col.diesis" project [7]. Here the basic idea is to associate colors to a melody played by an instrument. For a given color, darker colors are produced by lower pitch frequencies. Interestingly, based on the statistics of more than 700 people, they produced a table, which summarizes how individuals associate colors to musical instruments. It turned out that the mapping is: yellow for vibraphone or flute; geen for flute; orange for banjo or marimba; purple for cello or organ; blue for piano, trumpet or clarinet; red for guitar or electric guitar.

Capalbo and Glenney introduced the "KromoPhone" [8]. Their prototype can be used either in RGB mode or HSL mode. Using HSL, hue is sonified by sinusoidal sound pitch, saturation is associated to sound panning and luminosity is related to sound volume. The authors stated that only those individuals with perfect pitch perform well. In RGB mode the mapping of colors to sounds are defined by pan, pitch and volume. For instance, the gray scale from black to white is panned to the centre, with black being associated to the lowest pitch sound. Blue and yellow are mapped to the left, with blue being associated to lower pitch than yellow. Similarly, green and red are related to sounds listened to the right. Finally, the intensity of each color is mapped to the volume of the sound it produces.

In one of their experiments, Capalbo and Glenney illustrated that the use of color information in a recognition task outperformed the performance of "TheVoice" (cf. section 2) [8]. Specifically, the purpose was to pick certain fruits and vegetables known to correlate with certain colors. One of the results was that none of the three subjects trained with "TheVoice" could identify any of the fruit, either by the shape contours or luminance.

3.2. Color sonifcation in See ColOr

Relative to the HSL color system, we represent the Hue variable by instrument timbre, because it is well accepted in the musical community that the color of music lives in the timbre of performing instruments. Moreover, learning to associate instrument timbres to colors is easier than learning to associate for instance, pitch frequencies. The saturation variable S representing the degree of purity of hue is rendered by sound pitch, while luminosity is represented by double bass when it is rather dark and a singing voice when it is relatively bright.

With respect to the hue variable, the corresponding musical instruments are based on an empirical choice:

1. oboe for red ($0 \le H < 1/12$);
2. viola for orange ($1/12 \le H < 1/6$);
3. pizzicato violin for yellow ($1/6 \le H < 1/3$);
4. flute for green ($1/3 \le H < 1/2$);
5. trumpet for cyan ($1/2 \le H < 2/3$);
6. piano for blue ($2/3 \le H < 5/6$);
7. saxophone for purple ($5/6 \le H \le 1$).

Note that for a sonified pixel, when the hue variable is exactly between two predefined hues, such as for instance between yellow and green, the resulting sound instrument mix is an equal proportion of the two corresponding instruments. More generally, hue values are rendered by two sound timbres whose gain depends on the proximity of the two closest hues.

The audio representation h_h of a hue pixel value h is

$$h_h = g \cdot h_a + (1-g) \cdot h_b \qquad (1)$$

with g representing the gain defined by

$$g = \frac{h_b - H}{h_b - h_a} \qquad (2)$$

with $h_a \le H \le h_b$, and h_a, h_b representing two successive hue values among red, orange, yellow, green, cyan, blue, and purple (the successor of purple is red). In this way, the transition between two successive hues is smooth.

The pitch of a selected instrument depends on the saturation value. We use four different saturation values by means of four different notes:

1. C for ($0 \le S < 0.25$);
2. G for ($0.25 \le S < 0.5$);
3. B flat for ($0.5 \le S < 0.75$);
4. E for ($0.75 \le S \le 1$);

When the luminance L is rather dark (i.e. less than 0.5) we mix the sound resulting from the H and S variables with a double bass using four possible notes (C, G, B flat, and E), depending on luminance level. A singing voice with also four different pitches (the same used for the double bass) is used with bright luminance (i.e. luminance above 0.5). Moreover, if luminance is close to zero, the perceived color is black and we discard in the final audio mix the musical instruments corresponding to the H and S variables. Similarly, if luminance is close to one, thus the perceived color is white we only retain in the final mix a singing voice. Note that with luminance close to 0.5 the final mix has just the hue and saturation components.

The sonified part of a captured image is a row of 25 pixels in the central part of the picture. We take into account a single row, as the encoding of several rows would need the use of 3D spatialization, instead of simple 2D spatializazion. It is well known that rendering elevation is much more complicated than lateralization [9]. On the other hand, in case of 3D spatialization it is very likely that too many sound sources would be difficult to be analyzed by a common user.

Two-dimensional spatialization is achieved by the convolution of mono aural instrument sounds with filters encompassing typical lateral cues, such as interaural time delay and interaural intensity difference. In this work we reproduce spatial lateralization with the use of the CIPIC database [10].

4. OUR PREVIOUS EXPERIMENTS

In the first step of the See ColOr project, we performed several experiments with six blindfolded persons who were trained to associate colors with musical instrument sounds [1]. As shown by figure 1, the participants were asked to identify major components of static pictures presented on a special paper lying on a T3 tactile tablet (http://www.rncb.ac.uk/t3/index.htm) representing pictures with embossed edges. When one touched the paper lying on the tablet, a small region below the finger was sonified and provided to the user. Color was helpful for the interpretation of image scenes, as it lessened ambiguity. As an example, if a large region "sounded" cyan at the top of the picture it was likely to be the sky. Finally, all participants to the experiments were successful when asked to find a bright red door in a picture representing a churchyard with trees, grass and a house.

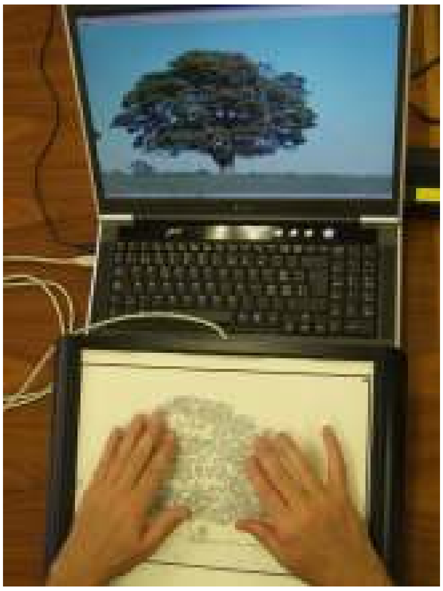


Figure 1: Example of embossed picture on the T3 tactile tablet.

The work described in [11] introduced an experiment during which ten blindfolded individuals participants tried to match pairs of uniform colored socks by pointing a head mounted camera and by listening to the generated sounds. Figure 2 illustrates an experiment participant observing a blue socket. The results of this experiment demonstrated that matching similar colors through the use of a perceptual (auditory) language, such as that represented by instrument sounds can be successfully accomplished.

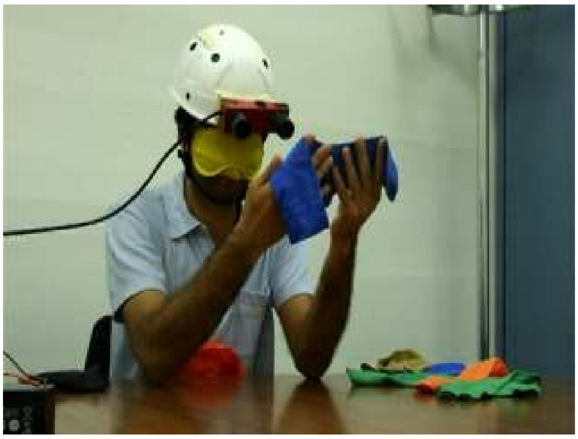


Figure 2: A blindfolded subject observing a blue socket.

In [12] the purpose was to validate the hypothesis that navigation in an outdoor environment can be performed by "listening" to a colored path. We introduced an experiment during which ten blindfolded participants and a blind person were asked to point the camera toward a red sinuous path painted on the ground and to follow it for more than 80 meters. Results demonstrated that following a sinuous colored path through the use of our auditory perceptual language was successful. A video entitled "The See ColOr project" illustrates several experiments on *http://www.youtube.com/guidobologna*.

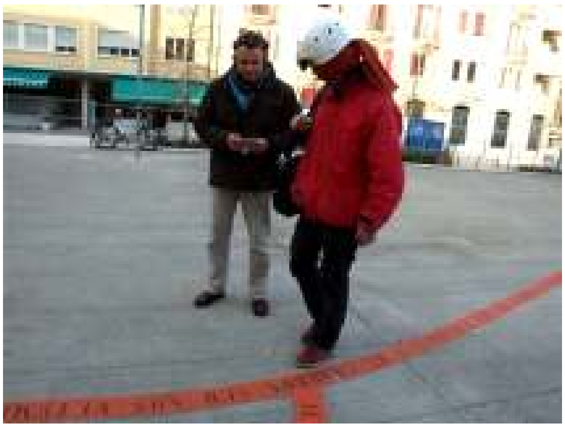


Figure 3: A blindfolded individual following a red sinuous path.

5. SONIFICATION OF COLOR AND DEPTH

We use a stereoscopic color camera denoted STH-MDCS2 (SRI International: http://www.videredesign.com/) and the "Bumblebee" (Point Grey: http://www.ptgrey.com/). An algorithm for depth calculation based on epipolar geometry is embedded within both the stereoscopic cameras. The resolution of images is 320x240 pixels with a maximum frame rate of 30 images per second.

Our See ColOr prototype presents two sonification modes that render color and depth. The first replicates a crude model of the human visual system. Pixels near the center of the sonified row have high resolution, while pixels close to the left and right borders have low resolution. This is achieved by considering a sonification mask indicating the number of pixel values to skip. As shown below, starting from the middle point (in bold), the following vector of 25 points represents the number of skipped pixels:

[15 12 9 7 5 3 3 2 2 1 1 1 **1** 1 1 1 2 2 3 3 5 7 9 12 15]

In the first mode, depth is represented by sound duration. The mapping for depth D is given by :

- 90 ms for undetermined depth;
- 160 ms for ($0 \leq D < 1$);
- 207 ms for ($1 \leq D < 2$);
- 254 ms for ($2 \leq D < 3$);
- 300 ms for D > 3

The second mode sonifies only a pixel of a particular area of 25 adjacent points in the middle of the image. Specifically, we first determine among these 25 points the greatest number of contiguous points labelled with the same hue. Then, we calculate the centroid of this area and the average depth. It is possible to have points of undetermined depth, especially in homogeneous areas like walls, for which the depth algorithm is unable to determine landmark points related to the calculation of the disparity between the left and right images. Points with undetermined depth are not considered in the average depth calculation. The final sonification presents only a spatialized sound source representing the average color and the average depth.

In the second mode, depth between one and four meters is sonified by sound duration (the same sonification scheme explained above), while after four meters the volume V starts to decrease by following a negative exponential function given by

$$\mathrm{f(V) = V * exp(-k * D)} \qquad (3)$$

with k a positive small constant.

6. PRELIMINARY EXPERIMENTS

The experiments were performed by a very well trained blindfolded individual, who is very familiar with this color sonification model, but not with depth sonification. Although in the long term we will aim at complementing the white cane of blind people by a miniaturized version of our prototype, this person relied only on the See ColOr interface. The reason is that we wanted to be sure that our prototype represented the only sensing tool.

In the first video entitled “Going out from the office” and in the second video entitled “Going into the office” we aim at demonstrating that it is possible to perceive an open door and to pass through it. Figure 4 illustrates a picture of this experiment, which is performed with the second sonification mode. The brown door is sonified by a viola and the rhythm is fast when the user is close to it. Note also that the user decided to move when slow sound rhythms or low volume sounds were discerned, indicating distant obstacles.

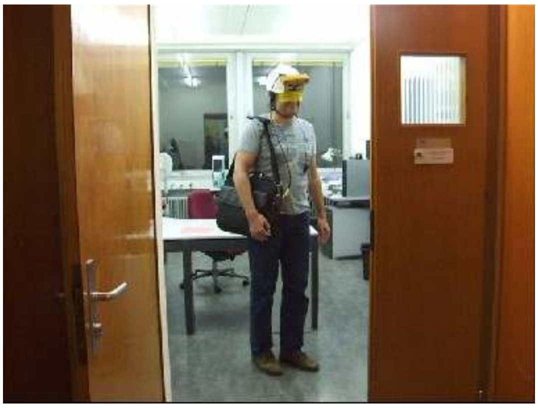


Figure 4: A blindfolded individual looking for an open door.

The third video entitled “Find a red tee shirt with sounds of musical instruments” illustrates the same individual in a successful search task. It is worth noting that here depth is often undetermined when the camera is pointed toward the floor or the white walls. Note also that the user trusted the depth information related to the trumpet sound representing the blue-cyan cabinets. The red tee short is sonified by oboe and when the user was close to it the rhythm frequency increased.

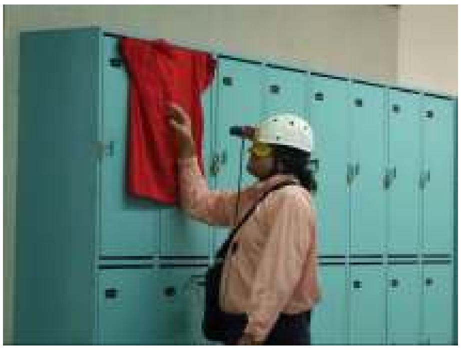


Figure 5: A blindfolded individual looking for a red tee shirt.

In the fourth video entitled “The blue cabinet” the user switched to the first sonification mode (with all 25 points sonified by color and depth). Here the goal was to find a blue cabinet sonified by a piano playing a medium pitched tone. This mode is more complex than the previous, since more than one color can be present in the current sonified frame. Here the distance to the floor is defined, as the floor is textured. Note also that the brown doors are sonified by viola sounds. From time to time, our experiment participant wished to ask to the computer the depth of the middle point of the sonified row. With the use of a mouse button the computer answered with a voice saying numbers in French. “One” means distance between zero and one meter; “two” means distance between one and two meters, etc. At the end of the video the user reached and indicated the cabinet.

In the last video entitled "Walking outside" the user walked outside. He switched again to the second mode with only a sonified sound. The sound of the ground is rendered by a singing voice or a double bass, depending on its gray level. Suddenly, the user found in his trajectory a parked car and he avoided it.

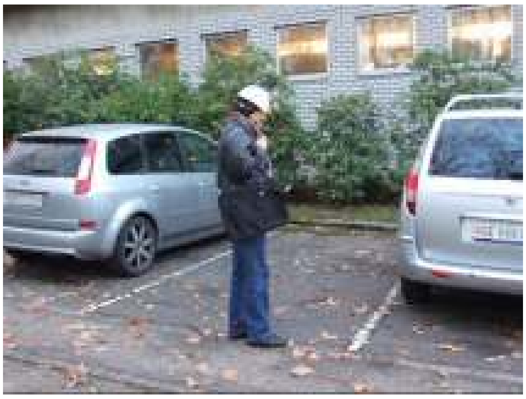


Figure 6: A blindfolded individual walking outside and avoiding parked cars.

After the experiments the blindfolded person was asked to give his impressions about the two different modes. The first impression is that the second mode (with the decreasing volume) is felt as "relaxing" compared to the first mode. The second mode is valuable in large areas (for instance, outside). Moreover, in some situations, it will be very useful to switch from the second mode to the first, as the first mode gives more precision and to some extent, peripheral view. A sonified compass could be also very useful, as it is very easy to loose orientation. Finally, while the first mode provides to some extent limited global information, a "global module" would be helpful in order to get a clear picture of the close environment geometry.

7. CONCLUSION

We presented the color and depth sonification model of the See ColOr mobility aid. A See ColOr prototype was tested by a well trained individual. He successfully (1) detected an open door in order to go in and out; (2) walked in a corridor with the purpose to find a blue cabinet; (3) moved in a hallway with the purpose to locate a red tee shirt; (4) walked outside and avoided a parked car. In the future, we would like to measure in a more systematic way whether the use of our prototype allows users to locate objects and to avoid obstacles of different sizes. Thus, we will perform experiments with more participants, in order to obtain more robust statistics.

8. REFERENCES

[1] G. Bologna, B. Deville, T. Pun, M. Vinckenbosch, "Transforming 3D coloured pixels into musical instrument notes for vision substitution applications", *J. of Image and Video Processing,* A. Caplier, T. Pun, D. Tzovaras, Guest Eds., Article ID 76204, 14 pages (Open access article), 2007.

[2] L. Kay, "A sonar aid to enhance spatial perception of the blind: engineering design and evaluation", *The Radio and Electronic Engineer,* vol. 44, pp. 605–627, 1974.

[3] P.B.L. Meijer, "An experimental system for auditory image representations", *IEEE Trans. Bio. Eng.,* vol. 39, no. 2, pp. 112–121, 1992.

[4] C. Capelle, C. Trullemans, P. Arno, and C. Veraart, "A real time experimental prototype for enhancement of vision rehabilitation using auditory substitution", *IEEE T. Bio-Med Eng.,* vol. 45, pp. 1279–1293, 1998.

[5] J.L. Gonzalez-Mora, A. Rodriguez-Hernandez, L.F. Rodriguez-Ramos, L. Dfaz-Saco, and N. Sosa, "Development of a new space perception system for blind people, based on the creation of a virtual acoustic space", in *Proc. IWANN'99*, pp. 321–330, 1999.

[6] K. Doel, "Soundview: Sensing Color Images by Kinesthetic Audio", *Proc. of the International Conference on Auditory Display,* Boston, MA, USA, July 6-9, 2003.

[7] J. Rossi, FJ. Perales, J. Varona, M. Roca, « Col.diesis: transforming colour into melody and implementing the result in a colour sensor device », *Second International Conference in Visualisation,* Barcelona, Spain, July 15-17 2009.

[8] Z. Capalbo, B. Glenney. "Hearing Color: Radical Plurastic Realism and SSDs", *Fifth Asia-Pacific Computing and Philosophy conference (AP-CAP 2009),* October 1-2 2009, Tokyo, Japan.

[9] R. Begault, *3-D Sound for Virtual Reality and Multimedia,* Boston A.P. Professional, ISBN: 0120847353, 1994.

[10] V.R. Algazi, R.O. Duda, D.P Thompson, C. Avendano, "The CIPIC HRTF Database", *In IEEE Proc. Workshop on Applications of Signal Processing to Audio and Acoustics,* Mohonk Mountain House, (WASPAA'01), New Paltz, NY, 2001.

[11] G. Bologna, B. Deville, M. Vinckenbosch, T. Pun, "A Perceptual Interface for Vision Substitution in a Color Matching Experiment". In *Proc. Int. Joint Conf. Neural Networks, Part of IEEE World Congress on Computational Intelligence,* June 1-6, 2008, Hong Kong.

[12] G. Bologna, B. Deville, T. Pun, "Blind navigation along a sinuous path by means of the See ColOr interface". In *Proceedings of the 3rd international Work-Conference on the interplay between Natural and Artificial Computation: Part II: Bioinspired Applications in Artificial and Natural Computation* (Santiago de Compostela, Spain, June 22 - 26, 2009). J. Mira, J. M. Ferrández, J. R. Álvarez, F. Paz, and F. J. Toledo, Eds. Lecture Notes In Computer Science, vol. 5602. Springer-Verlag, Berlin, Heidelberg.

SONICFUNCTION: EXPERIMENTS WITH A FUNCTION BROWSER FOR THE VISUALLY IMPAIRED

Florian Grond[1], *Trixi Droßard*[2], *Thomas Hermann*[1]

[1] Ambient Intelligence Group, CITEC, Bielefeld University, Germany
[2] Fakultät für Erziehungswissenschaft, University of Hamburg, Germany
fgrond@techfak.uni-bielefeld.de
Trixi.drossard@gmx.de
thermann@techfak.uni-bielefeld.de

ABSTRACT

We present in this paper `SonicFunction`, a prototype for the interactive sonification of mathematical functions. Since many approaches to represent mathematical functions as auditory graphs exist already, we introduce in `SonicFunction` three new aspects related to sound design. Firstly, `SonicFunction` features a hybrid approach of discrete and continuous sonification of the function values $f(x)$. Secondly, the sonification includes information about the derivative of the function. Thirdly, `SonicFunction` includes information about the sign of the function value $f(x)$ within the timbre of the sonification and leaves the auditory graph context free for an acoustic representation of the bounding box. We discuss `SonicFunction` within the context of existing function sonifications, and report the results from an evaluation of the program with 14 partially sighted and blind students.

1. INTRODUCTION

Teaching material for the blind and partially sighted in mathematics is generally tactile using Braille notation or reliefs. When it comes to function analysis, this form of notation has some limitations. One of which is its involved method of production, which is the reason why appropriate teaching material is limited.

Due to the fact that the blind and partially sighted often posses heightened auditory capacities, there have been occasional efforts to develop auditory displays for teaching mathematics. In sonification, a big amount of research has been conducted on auditory graphs. Foundational work was laid in [1] and good overviews over the field can be found in [2] and [3]. In [4] there is an interesting study that contrasts the difference between discrete and continuous auditory graphs, giving evidence that both representation modes serve different purposes. Related work can further be found in [5]. A conceptual model of auditory graph comprehension can be found in [6], where particularly the consideration on the context information in a graph, i.e. axes and their +/- orientation, are relevant for us.

While the results from this field provides a good basis for the development of auditory displays for mathematical functions of one variable $f(x)$, we believe that there are still possibilities for further improvement of the sound design. This is important, first because new concepts, that illustrate how to include information in the sonification of $f(x)$ rather than putting it into the auditory context, extends the usefulness of auditory graphs. Second auditory rich and yet distinguishable information is usually more interesting to listen to, and hence user fatigue can be reduced. Third, in many of the studies above auditory graphs had limited interaction possibilities. However as stated in [7] interaction introduces new and exciting possibilities for a better understanding of sonification in general and also does so for auditory graphs in particular. Particularly questions about continuous and discrete sonifications for mathematical functions must be revisited with respect to new possibilities in interactive sonification.

As a new methodological contribution to the field of function sonification, we here introduce the idea of multi-parameter sonification of mathematical functions, which goes beyond the existing pitch mapping-based strategies in the aspect that they utilize the Taylor expansion of function f at location x as source for a stationary sonic representation. More specifically, we suggest to map the first m terms of the Taylor series $(f(x), f'(x), \dots, f^{(n)}(x))$ at location x as *fingerprint* for the local characteristics of the function and derive a corresponding sonic counterpart.

Depending on the mapping, the main association of $f(x)$ to pitch can be maintained, but be extended to reflect slope for instance as pulse rate, curvature $f''(x)$ as attack time of events or whatever mapping seems appropriate. Since these attributes change systematically while traversing along the x-axis, a sort of recognizable sequenced auditory gestalt builds up when walking towards specific points of interest such as turning points, saddle points or local optima.

In the current empirical study, however, we adapted and restricted this more general sonification approach specifically to fit to the subject group. Second author Trixi Drossard who has a background as math teacher for the partially sighted and the blind conducted the study with pupils. We wanted to evaluate our sonification strategy with blind pupils from the very beginning and we were less interested to evaluate if pupils recognize already learned features of mathematical functions but more in whether sonic function representations work in a teaching situation. This imposed several constraints to the general multivariate representation concept so that in consequence we included only the first derivative, since the concept of higher derivatives are difficult to grasp and not part of the curriculum for pupils of the age of our test subjects.

One important application featuring interactive sonified graphs is the java program MathTrax, [8] which has been adapted to typical requirements for the blind and partially sighted, it works for instance together with screen readers and features shortcuts and hot-keys for efficient navigation. MathTrax presents visual, acoustic and descriptive information about mathematical functions. However, since we were interested in new sonification designs, we decided to implement our own minimal prototype called `SonicFunction`.

2. THE PROGRAM SONICFUNCTION

The program `SonicFunction` is implemented in python and Tcl/Tk for a minimal user interface and the calculations of the mathematical functions. Open Sound Control (OSC) [9] [10] provides the protocol to send the parameters to the `SuperCollider` soundserver [11] available at [12]. As an input device we decided to use the keyboard since it is a very familiar interface for the visually impaired. The user can interact with the program through the following keys:

- The arrow keys up and down control the volume of the sonification. This is important to adapt it to the volume of screen readers, such as Jaws for Windows.
- The arrow keys left and right allow to navigate on the x-axis. If the arrow keys were constantly pressed, the function can be browsed in a constant movement from left to right.
- The keys x,c and v set the step size for the navigation on the x-axis to $^1/_{30}$, $^1/_{10}$ and $^1/_{6}$ respectively. This allows for a quick overview of the function and for detailed inspection.
- The number keys 1 to 6 are the selectors for the test functions, which are described below in detail.
- By hitting the keys h, t, n and a markers for maxima, minima, f(x) = 0 and x = 0 respectively are registered in a protocol file.

While navigating the function on the x-axis the sonification was presented on the corresponding position within the stereo panorama. The interaction for placing markers was included since we wanted to evaluated the sonified function, by recording and analyzing user interaction.

3. DIFFERENCES TO MATHTRAX

As mentioned above, the program MathTrax is a popular reference for function sonification. In `SonicFunction` we try to to include the information that is connected to the function within the acoustic representation of the function itself. In this section, we want to highlight the differences in sound design between `SonicFunction` and MathTrax.

For the distinction of positive and negative function values, MathTrax employs for instance the auditory context by adding a constant level of noise. `SonicFunction` integrates this information within the sound that represents the function value of $f(x)$, by changing its timbre. Thereby leaving the context of the auditory scene free for an acoustic equivalent of a bounding box.

By choosing two noise sources with different center frequencies, this bounding box also helps to indicate, whether the $f(x)$ is currently beyond the upper or lower limit of the bounding box. This is helpful for approximate extrapolation before sounding function values within the box are encountered.

`SonicFunction` also makes use of the derivative as a parameter for for the sonification. This is important to support the exploration around minima and maxima.

`SonicFunction` also combined two sonification approaches using continuous and discrete acoustic representations. Thereby the discrete sonification event is used to give an appropriate feedback for the stepwise interaction when moving along the x-axis. The continuous standing sound that goes with a ramp from one function value to the next emphasizes the dense distribution of real numbers on the x-axis.

4. SOUND DESIGN

As mentioned above the interaction feedback was provided by a discrete sonification, whereas the continuity of the function was represented through a continuous sonification. Examples of the sonifications of all test functions can be found on our website [1].

4.1. The Discrete Sonification

The discrete sonification was played each time the user moved along the x-axis one step. In Figure 1 you find `SuperCollider` code for the synthesis definition of the discrete sonification.

```
SynthDef(\discrete,
  { arg out=0, midinote = 60, pan = 0.0, delay = 0.1,
        duration = 3.5, vol = 0.1, bwf = 1;
    var klank, harm, amp, ring, filter, lfo,
        noise_source, env, freq, sig;
  freq = midinote.midicps;
  harm = Control.names([\harm]).kr([1,2,3,4,5,6,7,8,9]);
  amp = Control.names([\amp]).kr(Array.series(9,1,1).reverse.normalizeSum);
  ring = Control.names([\ring]).kr(Array.fill(9,100.0));
  noise_source =
    EnvGen.ar(Env.new([0,0,1,0],[delay,0.0,duration],-3),1.0,doneAction:0) +
    (EnvGen.ar(Env.new([0,0,1,0],[delay,0.0,duration/20],-3),1.0,doneAction:0)
    * ClipNoise.ar(1));
  env = EnvGen.ar( Env.new([0.0,1.0,0.0],[0.0,delay+0.5],-3),1, doneAction:2);
  klank = Klank.ar(`[harm, amp, ring], noise_source, freq);
  sig = LPF.ar(klank, freq*bwf);
  OffsetOut.ar(0,Pan2.ar(sig*env*AmpComp.kr(freq,40.midicps), pan, vol));
  }  ).load(s);
```

Figure 1: The `SuperCollider` synthesis definition for the discrete sonification.

The sonification was essentially a sound made of subtractive synthesis (the unit generator `Klank.ar`) with a base frequency and a series of overtones of decaying gain. The frequency of the base frequency covered the range from 46,25 to 698,46 Hz, (approx. 4 octaves). The considerably low range was chosen to have enough overhead in the spectrum for the 9 overtones, which helped to identify the pitch of the sound even for low base frequencies.

The excitation of the `Klank` filter was an attack decay envelope with a noise component in the attack phase. The filtered sound was multiplied with an envelope which also had an attack decay characteristic.

The sound was played after a delay, that allowed the continuous sonification to ramp to the target frequency. The discrete sounds were played back within the stereo panorama corresponding to the actual position on the x axis within the bounding box. Basic psychoacoustic amplitude compensation `AmpComp.kr` was additionally implemented.

4.2. The Continuous Sonification and the Derivative

With respect to spectral characteristics, the continuous sonification resembled very much the discrete sonification, except it was implemented as additive synthesis using the unit generator `Klang.ar`.

[1] http://www.techfak.uni-bielefeld.de/ags/ami/publications/GDH2010-SEW/

In Figure 2 you find the corresponding `SuperCollider` synthesis definition.

```
SynthDef(\continuous,
  { arg out=0, midinote = 60, pan = 0.0, lg = 0.1,
        vol = 0.1, bwf = 1, modf = 5, moda = 0.1;
  var klank, harm, amp, phase, freq, sig;
  freq = midinote.midicps;
  harm = Control.names([\harm]).kr([1,2,3,4,5,6,7,8,9]);
  amp = Control.names([\amp]).kr(Array.fill(9,{1}));
  phase = Control.names([\pi]).kr( Array.geom(9,1,9).reverse.normalize);
  klank = DynKlang.ar(
        `[harm.lag(0.1)*freq.lag(0.1),amp.lag(0.1),phase.lag(0.1)]);
  sig = LPF.ar(klank, freq*bwf);
  OffsetOut.ar(0,
              Pan2.ar(sig * AmpComp.kr(freq.lag(0.1), 40.midicps),
              pan,
              vol * SinOsc.kr(modf,0,moda,1)  ));
  }).load(s);
```

Figure 2: The `SuperCollider` synthesis definition for the continuous sonification.

The continuous sonification was also the carrier of the information about the derivative $f(x)/dx$ which was mapped to an Amplitude oscillation, where the oscillation of the amplitude approached 0 if the derivative approached 0. You find the corresponding implementation detail in Figure 2 as `SinOsc.ar(modf,0,moda,1)`.

4.3. The Difference between Positive and Negative $f(x)$

For the distinction between positive and negative function values $f(x)$, the sound was send through a 2nd order Butterworth lowpass filter, `LPF.ar`, that allowed to control the brightness. By controlling the cutoff frequency (5 or 2.5 times the base frequency) two different brightness modes were selected, with the brighter one indicating positive function values.

4.4. The Acoustic Boundig Box

For the upper and the lower limit of the bounding box noise was send through a band pass filter (BPF). The metaphor of upper and lower was mapped to high and low for the center frequency of the BPF. The center frequency for the upper limit was set to 5000, and for the lower 200 Hz. The noise source was played back on the actual x position within the stereo panorama. The left and right bounding box limit was indicated through noise played back on the corresponding stereo channels. For all the functions the bounding box was set from -10 - 10 in x and -5 to 5 for $f(x)$.

We think that the acoustic bounding box is particularly instructive at singularities, where the function graph would first have a ascending frequency, then it would audibly cross the upper bounding limit, then at the singularity the center-frequency would change to low and finally the function is audible again at low frequencies.

4.5. Clicks as Tick-Marks on the x-Axis

In order to indicate tick-marks at each integer value on the x-axis, simple clicks were used. They were synthesized through short envelopes over an additive synthesis of 4 overtones with a base frequency of 1.000 Hz . The tick-marks were played back on the stereo panorama according to their position. The tick-mark at the position $x = 0$ was highlighted by n elevated base frequency of 1.600 Hz.

5. THE EXPERIMENT

Fourteen (7 female, 7 male) blind and partially sighted German students from the age range 17-19 participated in the study. Seven participants were blind, four were partially sighted, and three high-grade partially sighted, as stated by the participants themselves. For eleven of the participants their vision was constantly restricted or absent since their birth. Two of the participants with strongly restricted vision and one blind participants reported a degradation of their vision over the years.

Figure 3: Photo from the experiment: the test subject sits in the foreground on the right following the instructions by coauthor — ————

The experiment was conducted with each student individually in a quiet room in order to avoid acoustic disturbance. The assisting conductor of the study, coauthor ———— instructed the students how to use the program `SonicFunction`. For the acoustic display, regular headphones were used.

During the instruction period the students were encouraged to ask the instructor about the meaning of the sounds and the possibilities of interacting with `SonicFunction`. The instructor made sure that all acoustic features relevant for the tasks were understood.

The participants were browsing a selected function and reported verbally what kind of features they encountered. Each time they reported minima, maxima or values for $x = 0$ or $f(x) = 0$, the conductor marked the finding on the keyboard and the data were recorded in a file.

The students were also asked to guess and describe with words, what kind of function they thought they heard. At the end of the experiment they were asked to give feedback about the program `SonicFunction`. The participants were further asked what kind of learning type they are (visual, auditory or haptic), according to their preferences for learning most efectively. From the statements we could concloude, that ten students are visual learners while the other four are auditory learners.

6. TYPICAL FUNCTIONS AS TEST CASES

The following functions eq. 1 - eq. 6 were selected as test cases for the participants. The choice was primarily motivated by pedagog-

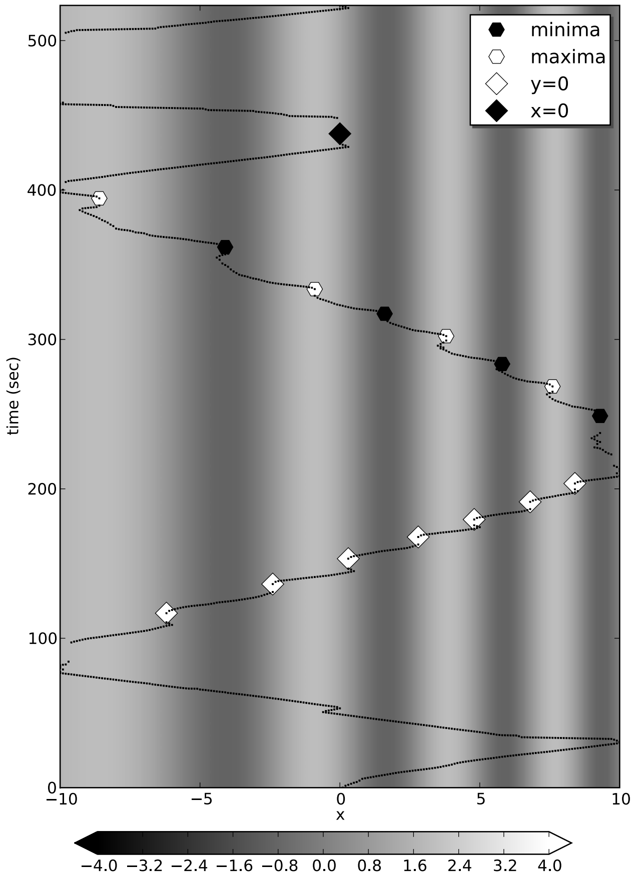


Figure 4: Example of a typical exploration of $f_5(x)$. The function values $f_5(x)$ are encoded in grey. The participant started in the middle and explored the function to both limits of the bounding box. Then $f(x) = 0$, further extrema and finally, $x = 0$ were marked.

ical aspects with regards to function analysis.

$$f_1(x) = {}^{3}\!/_{4}\,(t+1)^2 - 2 \quad (1)$$

$$f_2(x) = 2t + 3 \quad (2)$$

$$f_3(x) = t^2 + 1 \quad (3)$$

$$f_4(x) = 0.5/t \quad (4)$$

$$f_5(x) = \sin((0.2t + 3)^2)1.5 \quad (5)$$

$$f_6(x) = {}^{1}\!/_{\sin(t)} \quad (6)$$

The function from eq. 1 was selected to introduce the test-subjects to all the audible features of the auditory function graph. The values for x and $f(x)$ cover positive and negative values. Hence the test-subject hears the click for $x = 0$ and the change in timbre at the transition from negative to positive function values $f(x) = 0$. The minimum at $x = -1$ makes the LFO oscillation of the base frequency audible, which is controlled by the derivative $df_1(x)/dx$.

The second function eq. 2 was used to verify if the test subjects had understood the concept $x = 0$ as well as the concept of $f(x) = 0$ at $x = -3$.

In the third function, the symmetric parabola from eq. 3, test subjects were asked to identify the minimum and the position with $x = 0$.

With including function 4 we wanted to find out if test-subjects were able to make sense of an acoustically represented singularity.

Function 5 was included because we were interested if and how the precision of the extrema identification depends on the curvature i.e. the acoustic contrast around $df(x)/dx = 0$.

By including function 6 we wanted to find out how the concept of minima and maxima is perceived between singularities. These extrema are located at $\pi/2 \cdot m$ with $m \in | -5, -3 - 1, 1, 3, 5|$.

The test case functions together with the recorded markers for $f(x) = 0$ and $x = 0$ can be found in Figure 5, the markers for minima and maxima in Figure 6.

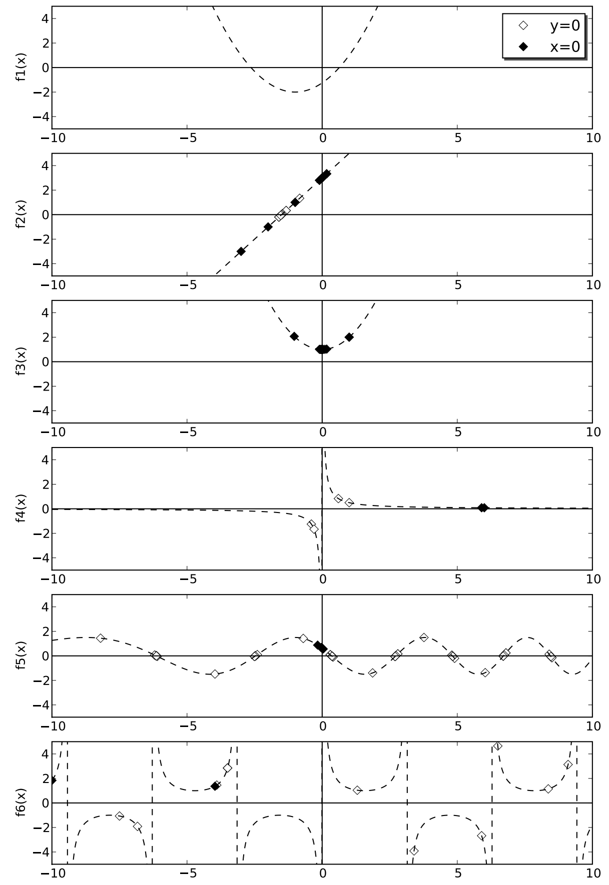


Figure 5: The test case functions with the $f(x) = 0$ and $x = 0$ markers

6.1. Discussion of Figure 5 and 6

6.1.1. Markers for $x = 0$ and $f(x) = 0$ in Figure 5:

1. $f_1(x)$ shows no markers since its sole purpose was to instruct the participants.
2. $f_2(x)$ shows that most of the markers were placed around $x = 0$ and $f(x) = 0$. There were outliers for $x = 0$. It seems that ordinary tick-marks on the x-axis were believed to be the distict tick-mark at $x = 0$.

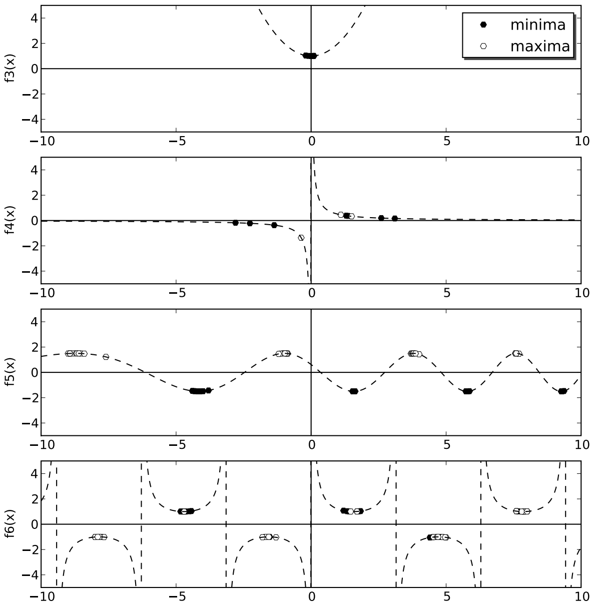


Figure 6: The test case functions with the minima and maxima markers

3. $f_3(x)$, here most of the markers have been placed at $x = 0$, again two outliers are found, which suggest similar problems as in $f_2(x)$.
4. $f_4(x)$ was a real challenge for the participants since none of the concepts $x = 0$ or $f(x) = 0$ were explicitly present. Interestingly some markers were placed approximately where the function has the strongest curvature.
5. $f_5(x)$ shows that most of the participants became familiar with the sonification and identified well the tested position except one person that marked the extrema as $f(x) = 0$.
6. $f_6(x)$ was a similar challenge as $f_4(x)$, and no particular pattern in the positioning of the markers can be found.

6.1.2. Markers for minima and maxima in Figure 5 :

1. $f_3(x)$ the minimum was well identified by all participants.
2. $f_4(x)$ some minima were wrongly identified were the function approached the x-axis.
3. $f_5(x)$ minima and maxima were well identified. note the broader distribution at extrema with lower curvature.
4. $f_6(x)$ minima and maxima were identified however the concept of both was confused.

6.2. A closer look on function $f_5(x)$

For the evaluation of the questions regarding $f_5(x)$ the first and the second derivative was calculated as in eq. 7 and eq. 8 respectively.

$$\frac{df_5(x)}{dx} = \frac{3}{25}\,(15+x)\,\cos(\Big(3+\frac{x}{5}\Big)^2) \qquad (7)$$

$$\frac{df_5(x)}{dx^2} =$$

$$\frac{-3}{625}\,\Big(-25\,\cos((3+\tfrac{x}{5})^2)+2\,(15+x)^2\,\sin((3+\tfrac{x}{5})^2)\Big) \quad (8)$$

By using numerical methods[2] to solve the equation $\frac{f_5(x)}{dx} = 0$, values for x were obtained within the interval from -10 to 10. Those values together with corresponding curvature are compiled in Table 1.

		$f_5(x)/dx = 0$	$f_5(x_i)/dx^2$
x_1	max	-8.733	-0.377
x_2	min	-4.146	1.131
x_3	max	-0.988	-1.885
x_4	min	1.580	2.639
x_5	max	3.799	-3.393
x_6	min	5.784	4.147
x_7	max	7.594	-4.901
x_8	min	9.270	5.655

Table 1: extrema and curvature values for $f_5(x)$

7. STATISTICS OF THE MARKER DISTRIBUTION

We calculated for some of the interesting cases the mean value and the standard deviation for the marker distribution. The results are compiled in Table 2.

fuction	marker	position	numeric value	mean	standard deviation
$f_2(x)$					
	x_0	0.0	0.0	-0.378	0.834
	y_0	$-3/2$	-1.5	-1.448	0.188
$f_3(x)$					
	min	0.0	0.0	5.3^{-15}	0.076
	x_0	0.0	0.0	4.4^{-3}	0.376
$f_5(x)$					
	max		-8.733	-8.640	0.377
	min		-4.146	-4.166	0.165
	max		-0.988	-0.993	0.106
	min		1.580	1.585	0.028
	max		3.799	3.806	0.077
	min		5.784	5.792	0.038
	max		7.594	7.604	0.042
	min		9.270	9.295	0.033
$f_6(x)$					
	min	$-5\pi/2$	-7.854	-7.862	0.112
	max	$-3\pi/2$	-4.712	-4.666	0.132
	min	$-\pi/2$	-1.571	-1.604	0.131
	max	$\pi/2$	1.571	1.566	0.208
	min	$3\pi/2$	4.712	4.710	0.190
	max	$5\pi/2$	7.853	7.830	0.151

Table 2: Results for the mean value and standard deviation for some of the markers in $f_2(x)$, $f_3(x)$, $f_5(x)$ and $f_6(x)$

The high values for the standard-deviation of x_0 for $f_2(x)$ and for $f_3(x)$ are due to the outliers as discussed in 6.1.1. The distribution of the markers around the maxima and minima of $f_6(x)$, which were all treated as extrema, is quite uniform. The function

[2] such as damped Newton's Method, as implemented in the software package Mathematica

$f_5(x)$ is an interesting case. Here we can see how the standard-deviation of the extrema decreases as we go along the x-axis from left to right. This seems to correspond to the increasing absolute curvature of the extrema in Table 1. In Figure 7 a correlation plot of the absolute value of the curvature against the standard-deviation $\sigma$and also against the standard deviation of $(x_i^k - x_0^k)$ denoted as and $\hat{\sigma}$ with x_0^k being the exact position of the extremum can be found.

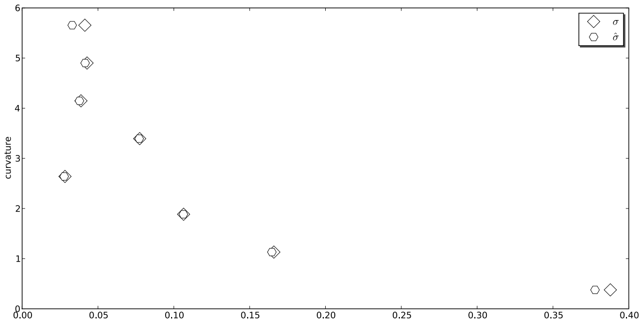


Figure 7: curvature versus the σ and $\hat{\sigma}$. It can be seen that low curvature tends to go with a broader distribution of clicks around the position of the function extrema

	correlation coefficient	p-value
$\lvert curvature\rvert$ versus σ	-0.7381	0.0366
$\lvert curvature\rvert$ versus $\hat{\sigma}$	-0.6905	0.0580

Table 3: Results from the Spearman rank correlation test

In Table 3 you find the results of the Spearman rank correlation coefficient and the two-sided p-value for a hypothesis test whose null hypothesis is that the two sets of data are uncorrelated. If we accept as a threshold for significance of 5% only the correlation with σ is below. None the less, we think that given the low amount of data (8 extrema), a general correlation between the curvature i.e. the acoustic contrast and the precision with which extrema are identified, can be established.

8. DISCUSSION

Looking at the results from the analysis of the markers set by the participants, we need to take into account that some of the mathematical concepts that were tested had maybe not been properly understood. One example is the misunderstanding of the definition of minima and maxima in function f_6. Their confusion in f_6 might be explained by the fact that the function values for maxima and hence their corresponding pitch was lower than the one for minima. Maximum and minimum seems to have been related to the absolute function value at $df_6/dx = 0$ and not to the sign of the curvature at that point.

However the evaluation of function f_5 lead to interesting insights. The precision with which extrema can be localized depends on the acoustic contrast i.e. the curvature around the extremum.

If we quickly summarize what the participants reported verbally about function f_5, none of them reported explicitly the increase of frequency while exploring $f_5(x)$ along the x-axis. In brief the participants said that the function appears as "something sinus like". This is explainable since the functions were all explored interactively and the progression along the x-axis was not necessarily constant. Therefore the change in the frequency of oscillations between the extrema have not been perceived or interpreted by the participants.

9. CONCLUSION

The evaluation of SonicFunction should be considered as a preliminary study of sonified graphs in a real-world teaching situation with partially sighted. From the experience of using SoniFunction in school, we can conclude that particularly for students who are either strongly partially sighted and use media specific for the blind or who are primarily an auditory learning type, the sonified functions are very supportive to grasp important characteristics of a mathematical function. The auditory graphs are especially well suited to be an alternative offer for strongly partially sighted or for students who are in the in the process of loosing sight. In these cases the sense of touch is not yet differentiated enough and those people often cannot handle braille yet. However auditory graphs are not meant to be a replacement for tactile graphs, but rather an addition to them to facilitated understanding for different learning types.

As far as the sonification design is concerned, we can not yet prove nor measure its utility, and the experimental results do not permit to compare our design to other sonification designs. This is mostly due to the heterogeneous population of our test subjects in terms of the restriction of their eyesight. Furthermore, the pupils have initially not been familiar with the idea of acoustic representation so that this was already a challenge and novelty, although it was generally much appreciated.

For future studies we therefore consider two directions: (a) using different sonification designs according to our new Taylor-based multi-parameter mapping concept with subjects that are already familiar with the mathematical background, e.g. math students, and (b) testing the winning design in a longitudinal study together with pupils who learn mathematical functions with the aid of sonification.

From our experience so far we found that the strategy to move information from the context to the sonification itself is promising: the integration of the transition from negative to positive function values as timbre filter leaves the noise stream available for the bounding box information. The successful integration of derivatives into the sonification is particularly important in case of interactive exploration where it is not assured that the user receives a proper overview over the progression of the function along the x-axis with a constant rate and therefore has more challenges to deduce information such as the curvature.

In summary, SonicFunction introduced a new mapping rationale and demonstrated hand-tuned contextual elements for the auditory display of mathematical functions for the visually impaired. We plan to address the open questions in our future research.

10. ACKNOWLEDGMENT

The authors would like to thank the Carl-Strehl-Schule, Marburg (Germany) for co-operation and the 14 students for participating

in the study.

11. REFERENCES

[1] B. u. J. Mansur, “Sound graphs: A numerical data analysis method for the blind,” *Journal of Medical Systems*, vol. 9, no. 3, 1985.

[2] T. Stockman, L. V. Nickerson, and G. Hind, “Auditory graphs: A summary of current experience and towards a research agenda,” E. Brazil, Ed., Department of Computer Science and Information Systems, University of Limerick. Limerick, Ireland: Department of Computer Science and Information Systems, University of Limerick, 2005, pp. 420–422.

[3] T. L. Bonebright, “A suggested agenda for auditory graph research,” E. Brazil, Ed., Department of Computer Science and Information Systems, University of Limerick. Limerick, Ireland: Department of Computer Science and Information Systems, University of Limerick, 2005, pp. 398–402. [Online]. Available: http://www.icad.org/Proceedings/2005/Bonebright2005.pdf

[4] L. Harrar and T. Stockman, “Designing auditory graph overviews: An examination of discrete vs. continuous sound and the influence of presentation speed,” G. P. Scavone, Ed., Schulich School of Music, McGill University. Montreal, Canada: Schulich School of Music, McGill University, 2007, pp. 299–305. [Online]. Available: Proceedings/2007/HarrarStockman2007.pdf

[5] S. Hetzler and R. Tardiff, “Two tools for integrating sonification into calculus instruction,” in *Proceedings of the twelfth International Conference on Auditory Display (ICAD2006)*, 2006, pp. 281–284.

[6] M. A. Nees and B. N. Walker, “Listener, task, and auditory graph: Toward a conceptual model of auditory graph comprehension,” G. P. Scavone, Ed., Schulich School of Music, McGill University. Montreal, Canada: Schulich School of Music, McGill University, 2007, pp. 266–273. [Online]. Available: Proceedings/2007/NeesWalker2007.pdf

[7] T. Hermann and A. Hunt, “An introduction to interactive sonification (guest editors’ introduction),” *IEEE MultiMedia*, vol. 12, no. 2, pp. 20–24, 04 2005.

[8] R. Shelton, S. Smith, T. Hodgson, and D. Dexter. Mathtrax. [Online]. Available: http://prime.jsc.nasa.gov/MathTrax/index.html

[9] CNMAT, Ed., *Proceedings of The Open Sound Control Conference 2004*, vol. 1, Berkeley, CA, USA, 30/07/2007 2003.

[10] M. Wright, A. Freed, and A. Momeni, “Open sound control: State of the art 2003,” in *International Conference on New Interfaces for Musical Expression*, Montreal, 2003, pp. 153–159.

[11] J. McCartney, “Rethinking the computer music language: SuperCollider,” *Computer Music Journal*, vol. 26, pp. 61–68, 2002.

[12] [Online]. Available: http://supercollider.sourceforge.net

CORRECTIVE SONIC FEEDBACK FOR SPEED SKATING: A CASE STUDY

Andrew Godbout

University Of Calgary
Department Of Computer Science
Calgary, Alberta, Canada
agodbout@ucalgary.ca

Jeffrey E. Boyd

University Of Calgary
Department Of Computer Science
Calgary, Alberta, Canada
jboyd@ucalgary.ca

ABSTRACT

We present a system that provides real-time audio feedback to athletes performing repetitive, periodic movements. The system synchronizes the temporal signal from a sensor placed on the athletes body with a model signal. The audio feedback tells the athlete how well they are synchronized with the model, and whether or not they are deviating from the model at critical points in the periodic motion. Because the feedback is continuous and in real-time, the athlete is able to correct their motion in response to the sounds they hear. The system uses simple, inexpensive instrumentation (the entire system costs less than $500) and avoids the uses of expensive and inconvenient motion capture systems. We demonstrate the effectiveness of the system with a case study featuring a speed skater that had developed a significant anomaly in his technique.

1. INTRODUCTION

In sports, certain movements happen so quickly that it is almost impossible for a coach to give instructions while an athlete is in the process of executing the movement. A golfer performing a golf swing, a gymnast performing a flip or a speed skater executing a cross-over are examples of movements that through traditional coaching methods are analyzed only after they have been executed and adjustments made only on the next repetition. Figure 1 illustrates this problem. Advances in wireless technology, computing power and computing portability now allow for analysis of these fast movements to happen in real time in the sporting environment. Sound is the natural communication medium to communicate to an athlete that is already taxing their vision, balance and tactile senses to perform their movement. We have developed a computerized sonic feedback system that improves how these rapid and repetitive movements can be coached. We have focused on the sport of speed skating and a unique opportunity to work with one particular athlete who had lost the ability to perform a proper speed skating cross-over.

We developed a system to aid this athlete using corrective sonic feedback. Using a sensor, we matched the skating stride of our subject to that of a model skater. Using this matching information we were then able to sonify the data relating to differences or imperfections in the subjects movement and communicate it to the subject as it was happening. We were able to provide cues, timing and body position information all in real-time. The sonification we produced allowed the athlete to make corrections and adjustments on the fly, something he and his coach were not able to do through traditional means.

Our system is a cheap and effective alternative to expensive and bulky motion capture systems. The cost of the measurement sensor we employ is on par with a cup of coffee. The whole system costs less than $500. The system is unobtrusive and easily worn by an athlete, and best of all it requires little calibration. We are able to achieve this without expensive measurement apparatus because we match using a relative pattern in the data rather than absolute measurement values.

As a case study with a single subject, we had no controls for validity. Nevertheless, the singular opportunity to test sonic feedback for correcting an athletic movement provides a valuable lesson, and suggests future direction for studies where controls are possible.

2. BACKGROUND

2.1. Sound in Sport

Sound plays an important role in sport, generally providing complimentary information to athletes. Naturally occurring sounds like a skate blade gliding on ice or a golf club impacting a ball are common in sport and can influence an athlete [1]. Advances in computing ability allow for all sorts of sport related information to be analyzed electronically and converted into sound. Running pace [2], rowing boat velocity[3] and karate movements [4] have all been electronically analyzed and used to control or create a sound that is communicated back to the athlete. The ability of a subject to mimic the jumping height of another subject using sonified jumping data [5] shows the potential for sound as a teaching tool.

2.2. Phase Matching

To compare the strides of a skater with model data, we must synchronize two signals, i.e., we must compute the phase shift between the two signals. Measurement of the phase shift between two periodic waveforms is a well-known signal processing technique with cross-correlation being the most frequently used method [6]. A useful variation that accounts for linear transformations of the signals is the normalized cross-correlation. When a signal is known to be a sinusoid, a phase-locked loop is a good alternative for synchronization of a reference oscillator to an input signal [7]. The method we use, described later, is a normalized cross-correlation with modification to allow for signals scaled in time and frequency.

2.3. Speed Skating

In speed skating, athletes race counterclockwise around a 400 meter oval consisting of two 100m straight sections and two 100m

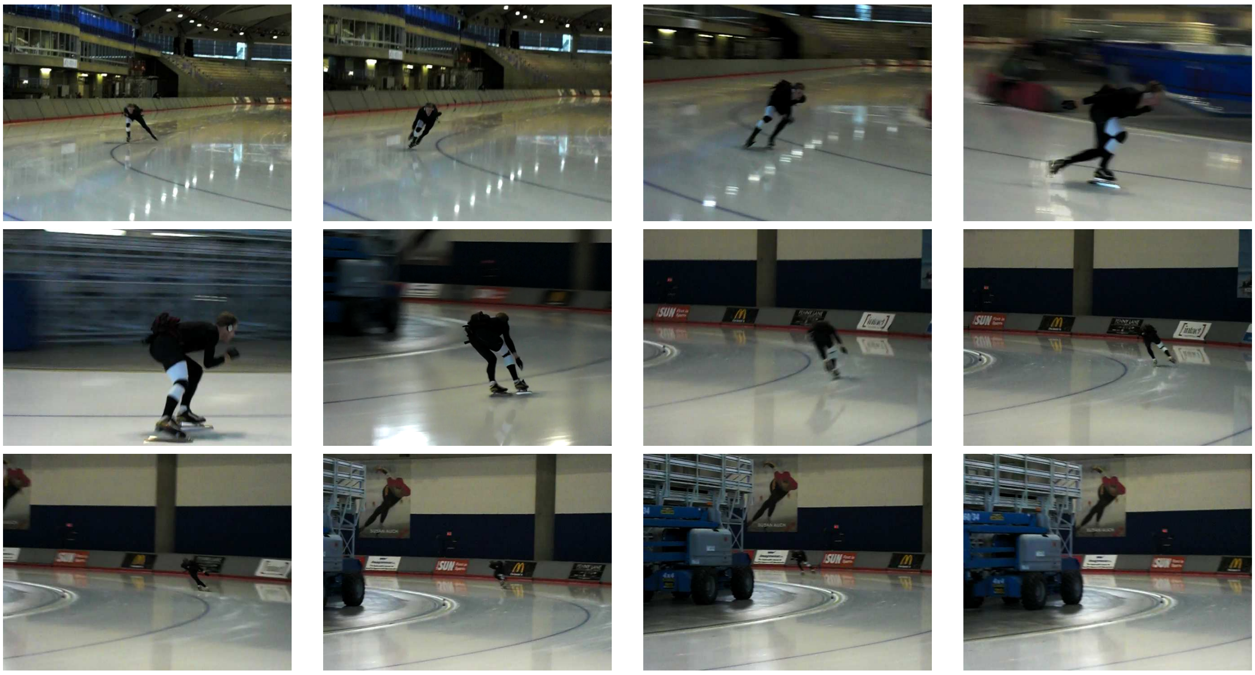


Figure 1: The window for coaching feedback. This sequence shows our subject skater over an interval of $8s$ during which time he travels approximately $80m$. At $10m/s$ ($36km/h$) it is difficult for a coach to see more than one or two single strides and give meaningful verbal feedback.

corners. To execute the corners a skater performs a cross-over, lifting the right skate over top of the left one. A skater will perform between 6 and 9 cross-overs in a span of 6-8 seconds as they navigate the 180 degrees that span the 100 meter corner. Figure 2 shows a plot of right ankle extension versus time in a cross-over from our model skater. The plot is divided into the three components that make up a cross-over:

1. **Right Foot Pushing:** the skate is in contact with the ice as the skater pushes (Figure 2 - A),
2. **Right Foot in Air:** the skater lifts his skate off the ice and moves it across the left skate (Figure 2 - B), and
3. **Right Foot Prepares to Push:** the skate blade contacts the ice (the set-down) as the skater prepares to push again (Figure 2 - C).

Efficient cross-overs are critical to achieve top performance in speed skating.

2.4. Our Subject

Our subject was unable to perform an efficient cross-over. He was able to perform component A and parts of component B without difficulty, however when he wanted to put his right skate back onto the ice at the set-down point he would dig the toe of his skate blade into the ice. This caused loss of speed and risk of crashing. The correct motion is to set the right skate blade down evenly onto the ice.

The athlete had previously been able to perform a cross-over and was a successful, nationally ranked racer. Cross-overs were a routine movement for him but at the start of a recent season he lost the ability to perform a cross-over properly. This condition is often referred to as *”Lost Move Syndrome”* and although uncommon, does occur in elite level sports[8]. Our subject sought help through traditional coaching methods like video analysis, sports therapists, and physiologists and over the course of 14 months was unable to improve the problem.

The subject was a perfect candidate for the sonic feedback system we have developed. He described his problem as not knowing that his toe was about to dig into the ice and feeling like he had a disconnect between what he was feeling and what was actually happening. His coaches were unable to provide feedback about the orientation of the toe of his blade during the process of the cross-over. We hypothesized that corrective sonic feedback would help correct his cross-overs.

We built a system to measure ankle extension and matched the amount of ankle extension during his skating stride to that of a model. In this way we were able to predict whether his toe would dig into the ice or not on any given cross-over and provide sonic feedback to the athlete in advance of the set-down.

3. SYSTEM DESCRIPTION

3.1. Apparatus

We used a single variable-resistance elastic, depicted in Figure 3, attached between the toe and shin of our skater (Figure 4) to measure the amount of ankle extension at any time. As the athlete skated, a netbook computer carried in a backpack measured the elastic’s resistance, R_s at 33 Hz. The plot in Figure 2 was obtained from this apparatus. At less than $500, the cost of our entire system is only a fraction of what other options like video based motion capture or motion capture suits cost. The simplicity of

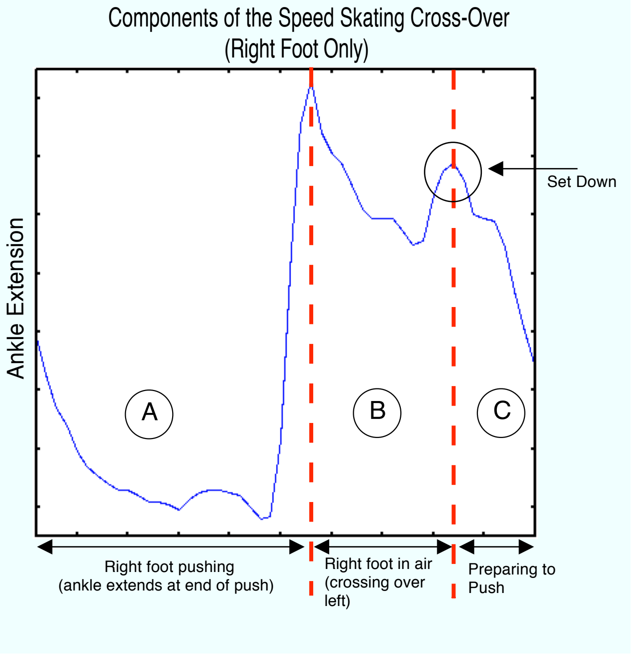


Figure 2: The Speed Skating Cross-over - Ankle Extension versus Time. The Cross-over is divided up into 3 components for the right foot. A) The right foot is pushing. The ankle is compressed with the knee over the toe for the first portion of this component (little ankle extension). During the end of this phase the ankle extends as the calf finishes the push. B) The right foot is in the air crossing over the left. The ankle retreats from its fully extended position. During the second portion of this component the ankle comes back to neutral or level so that the skate can set down flat on the ice. C) The right skate is back on the ice, the knee moves back over the toe and ankle extension reduces in anticipation of the next push component. The set-down when the right skate comes back into contact with the ice is labeled. This plot is from our model skater.

the system and little requirements for calibration or time consuming manual body measurements make this system practical for use with real athletes in the sporting environment.

3.2. Synchronization

The most important aspect of our system, is its ability to accurately synchronize the subject's skating stride to that of our model's stride. The following is a description of our brute-force method to estimate the phase of a speed skating stride from a single sensor stream.

Let g be the model signal of n samples containing a single cycle of data from the sensor. If f is an n-sample segment from on-line sensor data (we use the most recent n samples when synchronizing on-line in real-time), we can use a correlation to compare f to the model signal, g, i.e.,

$$\begin{aligned} h &= f \otimes g, & (1)\\ &= \sum_{i=0}^{n-1} f(i)g(i). & (2) \end{aligned}$$

The magnitude of h is a measure of how well f matches g.

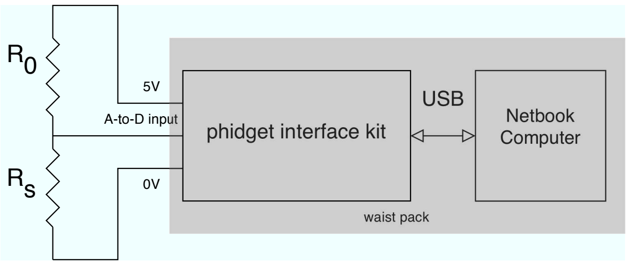


Figure 3: The sensor circuit: The sensor is a variable-resistance elastic (R_s) approximately 20mm in length . R_0 and R_s form a voltage divider. 5V supplied by the Phidget Interface Kit (http://www.phidgets.com) is applied across the voltage divider. An analog-to-digital converter in the Phidget Interface Kit measures the voltage across R_s, thereby measuring the stretch of the elastic. A netbook computer acquires the digital data from the interface kit, making it available for sonification.

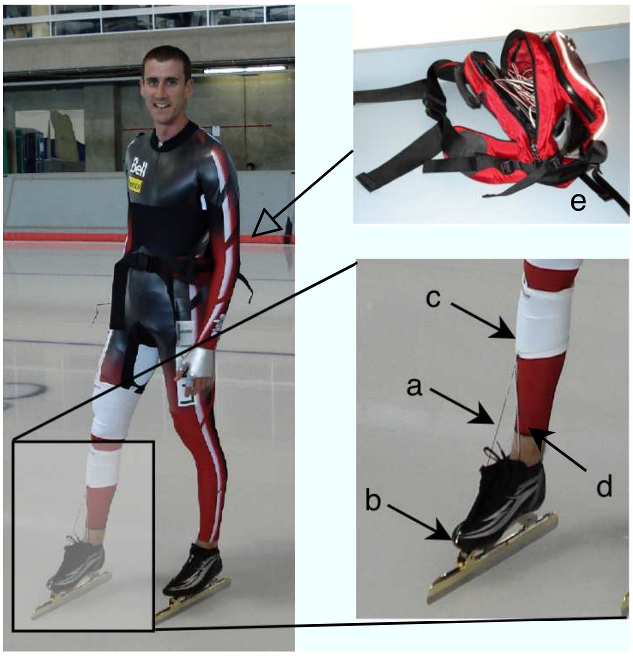


Figure 4: The sensor installed on the model skater: The variable-resistance elastic (a) is connected between a skate lace near the toe (b) and an elastic joint-support band (c) (used only to fasten the sensor). In this configuration, R_s (and therefore the voltage measured by the interface kit) increases with ankle extension. Leads (d) connect the sensor to the phidget interface kit (http://www.phidgets.com) and netbook computer worn by the skater in a waist pack (e). Sound is broadcast through headphones (not shown).

However, f is periodic, and there is no guarantee that the phase of f will match that of g, so we must consider the set of models given by

$$g((i+s) \bmod n), \qquad (3)$$

where $0 \le s < n$ determines the phase shift of the model. Now consider the correlation

$$h(s) = \sum_{i=0}^{n-1} f(i)g((i+s) \bmod n). \tag{4}$$

Therefore, phase, ϕ of f is

$$\phi = \frac{1}{n} \operatorname*{argmax}_s h(s). \tag{5}$$

and

$$\max_s h(s) \tag{6}$$

indicates how well f matches the model. Note that $0 \leq \phi < 1$.

Now suppose that we know the shape of each cycle of the signal, but we do not know the frequency. In this case, we need a set of models, g_n, where the subscript n indicates the number of samples in g_n. Thus n determines the period (and therefore the frequency) of the stride. The matching function becomes

$$h(s,n) = \sum_{i=0}^{n-1} f(i)g_n((i+s) \bmod n). \tag{7}$$

We can determine the correct period of the model, $\hat{n}$ with

$$\hat{n} = \operatorname*{argmax}_n \max_s h(s,n), \tag{8}$$

and the phase with

$$\phi = \frac{1}{\hat{n}} \operatorname*{argmax}_s h(s,\hat{n}). \tag{9}$$

The absolute measurements from the sensor vary with temperature, length of sensor, and where it is mounted on the toe and shin of the athlete. Given that we cannot control these factors, it is essential to *normalize* f and g with a linear transformation such that:

$$\sum_{i=0}^{n-1} g(i) = \sum_{i=0}^{n-1} f(i) = 0 \tag{10}$$

$$\sum_{i=0}^{n-1} g(i)^2 = \sum_{i=0}^{n-1} f(i)^2 = 1 \tag{11}$$

Note that a perfect match between skater and model will yield

$$h(s,n) = 1 \tag{12}$$

when f and g are normalized this way.

3.3. Sonification

Once the phase is matched successfully the stride cycle can be sonified. We worked within the Pure Data (http://puredata.info/) environment to do the sonification. The model stride is divided arbitrarily into four equal sections. When the phase of the subject crosses over one of the boundary lines between a section the system plays a note. Figure 5 shows a depiction of a successful stride matching using Equations (8) and (9). We selected four sine tones from a C major chord as the notes. The frequencies of the four tones are: 261.6 Hz, 329.6 Hz, 391.9 Hz, and 523.2 Hz. This produces an arpeggio, helping the subject to naturally synchronize his stride with that of the model.

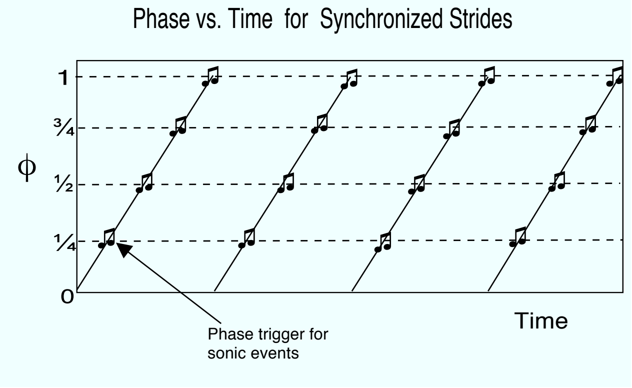


Figure 5: A successful synchronization between model and subject. ϕ ramps from 0 to 1 during each stride. Sound events are triggered as the skater progresses through the stride.

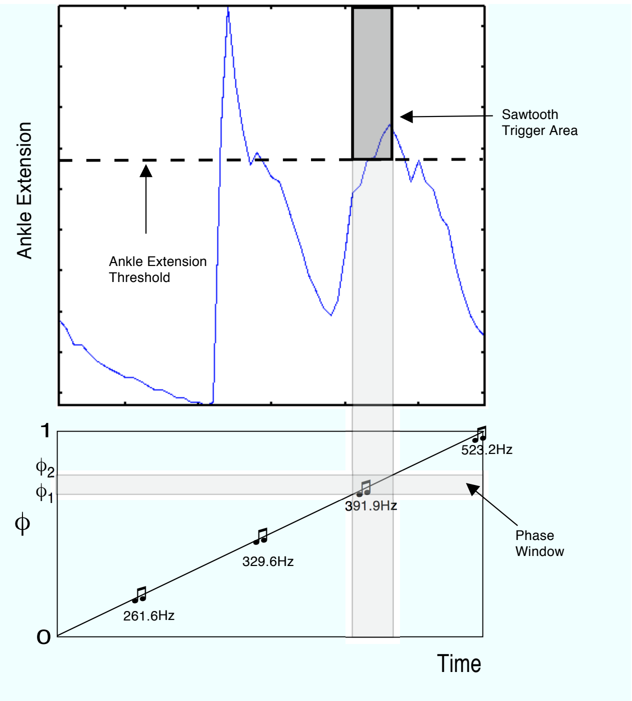


Figure 6: Phase is used to identify the window of time (highlighted rectangle) when we check if the subject is performing an incorrect movement. Phases between ϕ_1 and ϕ_2 form this window of time. Ankle extension exceeding the threshold, occurring during phases between ϕ_1 and ϕ_2 are considered to be incorrect. Incorrect movements trigger a corrective sound in the form of a sawtooth tone. Any sawtooth tone interrupts the background sine tones which are marked on the phase graph.

A reliable and accurate phase matching gives us the ability to focus on any part of the stride. We focus on the problematic area of the stride for our subject, the period of time immediately before the set-down. This is where we singled out variations between the model stride and the subject. We aimed to limit the amount of ankle extension allowed during this period. A threshold is set in the phase interval defined by ϕ_1 and ϕ_2 immediately preceding the set-down. Within this time interval, if the ankle extension of the subject surpasses the threshold, the system produces a corrective sound in the form of a sawtooth tone with harsh harmonics. Figure 6 shows a graphic representation of how phase information combined with ankle extension data is used to trigger corrective sawtooth tones. Our system is designed such that the threshold is adjustable and controllable from a base station via wireless network while the skater is skating.

The intensity of the sawtooth tone is directly related to the amount by which ankle extension surpasses the set threshold. In this way the subject can distinguish small deviations from the expected movement apart from large ones. Because our ankle extension measurements are relative rather than absolute, we scale the intensity of the sawtooth tone in proportion to the maximal readings from the sensor.

The resulting system produces a rhythmic arpeggio of consonant sine tones when the skater matches the model, but changes to a harsh sawtooth tone when the skater deviates at critical points in the stride. The intensity of the harsh sawtooth tone is proportional to the degree of deviation.

4. TESTS

We worked with our skating subject for a period of two months with approximately two one hour training sessions per week. The athlete also conducted his regular training regime and competed in a number of competitions during this time. We aimed to have the athlete use the system for as long a continuous period as was practical during a session. Ultimately we determined that fitting as many 3 - 4 lap repetitions in the one hour ice time was the most practical training method. Four laps last approximately 2.5 minutes total.

Speed Skating is physically demanding and we had to work within the abilities of the skater. Skating for 2.5 minutes and then taking a few minutes rest seemed to work the best. We had hoped to try training for continuous periods in the 10 minute range (15 - 20 laps) but that was not practical for our situation.

We progressed through three different training methods during the two months. We used our observations and feedback from the athlete to make necessary adjustments.

4.1. Corrective Feedback Training

The first training method we used was a corrective feedback set-up as described in detail in Section 3.3. Figure 7 shows the subject's skating stride before any training. We set up a threshold on the amount of ankle extension allowed in the period immediately before set-down as shown in Figure 6. Exceeding the threshold results in a sawtooth tone. The skater was instructed to try to avoid making the sawtooth tone. We began with a modest threshold, slightly less ankle extension than what the subject was already doing. We gradually decreased the threshold allowing less and less ankle extension until we reached a level that would result in a correct cross-over.

4.2. Awareness Feedback Training

The second training method we attempted required no alterations to the hardware or software that was used in the corrective feedback set-up described above. The skater however was given different instructions than what was given during corrective feedback training. We used the system and the sawtooth tone to create awareness about how much ankle extension was allowed. This time the skater was instructed to purposefully create the sawtooth tone. During Component B, labeled in Figure 2, when the right skate is in the air, the skater was instructed to extend his ankle pointing his toe towards the ice (creating the sawtooth tone) and then lift his toes back up until the sound stopped. After this the skater would attempt to finish his stride regularly by proceeding to set his right skate back to the ice. The skater did not skate normally doing this, it was a modified skating stride that allowed him more time with his right skate in the air. The skater went slower and was more upright to allow for this additional movement.

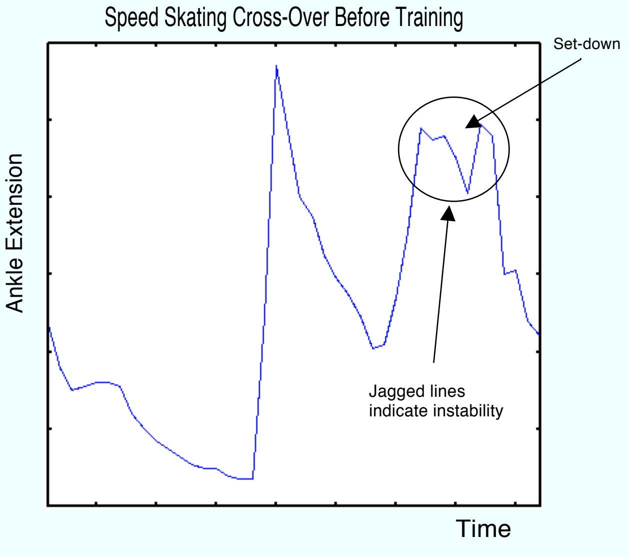


Figure 7: The skating stride of the subject before training. Notice the jagged movements at the set-down resulting from skater instability when the toe of the skate blade digs into the ice. In general the abrupt changes in direction on the graph and lack of smooth curves indicate a lack of flow in the skating stride.

The sawtooth tone was used to increase awareness about the expected movement. As the skater became comfortable with the movement, the threshold on the sawtooth tone was decreased. The aim was to reduce the threshold until the purposeful extension of the ankle was gone and the athlete, using his new found awareness of the correct amount of ankle extension, was left doing correct cross-overs.

4.3. Instruction Based Training

The final training method we used changed from a reactive system to a proactive system. Rather than giving feedback after the ankle extension exceeded the threshold, we provided a prompt telling the skater when we though he should extend his foot to meet the ice. We tried to manufacture the set-down point. The aim here was to not allow the athlete enough time to extend his ankle beyond the threshold. Instead we prompt the athlete to set-down before he has made the incorrect movement. There no longer was a corrective feedback aspect but rather we used the phase matching information to determine when we thought the skater should try to set down his foot.

We produced a bell tone at what we thought was the appropriate moment to start setting the right foot on the ice. The skater was instructed to extend for the ice with his right skate each time he heard the bell. We did not want to allow the skater enough time to extend his ankle pointing his toe to the ice. With enough training the manufactured set-down point would become the athlete's natural movement.

5. OBSERVATIONS

5.1. General Observations

The model stride, which is shown on the graph in Figure 2, from a speed skating perspective, is aesthetically pleasing. Focusing on the area around the set-down the model stride is smooth and devoid of abrupt changes in direction. Comparing this to the subject's stride before training, shown in Figure 7, we see the graph has plenty of abrupt changes in direction indicating inefficient on-ice movements and skater instability.

It is important to note that the subject had problems with his cross-overs during a continous 14 month span. During this time he repeatly executed incorrect cross-overs and that incorrect movement became ingrained into his motor pattern. Knowing that the subject had tried many different possible solutions to this problem without success, we entered into our training with moderate expectations. Contrary to those expectations, the athlete displayed improvements much sooner than we anticipated.

Upon training with the system the aesthetics of the athlete's stride quickly improved. The abrupt and extreme variations in ankle extension were muted and in some cases we achieved flawless set-downs. A flawless set-down was something the athlete was unable to achieve during the previous 14 months.

Common throughout testing were the arpeggio of background sine tones. One of the first improvements we noted with the athlete was an amelioration of the subject's cross-over. When we first started with the subject one of his cross-overs lasted approximately 1.3 seconds (42 samples at 33Hz). The duration of a typical cross-over from our model was closer to 1.5 seconds (50 samples at 33 Hz). The skaters were skating approximately the same speed. The model was covering more ground per cross-over than our subject. Almost immediately the subject modified his skating style to mimic that of the model in terms of stride duration. This improvement was persistent in all training methods. We attribute it to the subject using the arpeggio of background sine tones to maintain the rhythm of the different components of his stride. Our subject agreed: "this device was highly successful in helping me achieve a more efficient and fluid stride pattern while skating".

5.2. Corrective Feedback Training

Corrective feedback training produced a stride that was improved from what the skater was doing previously. However it seemed on par with what the skater was able to achieve using prior methods. During this training, the skater was attempting to pull his toes up to make the sawtooth tone go away. This was similar to his previous attempts at pulling his toes up before setting down. This time however the intensity of the sawtooth tone would predict how badly his toe was about to dig into the ice. He was not able to have a clean set-down but only mitigated the problem. It is evident at the set-down in Figure 8 by the jagged portion of the plot that the set-down was not ideal. It is also evident, by a reduction in the jagged portions of the plot, that this set-down is an improvement over the set-down seen in the untrained stride (Figure 7).

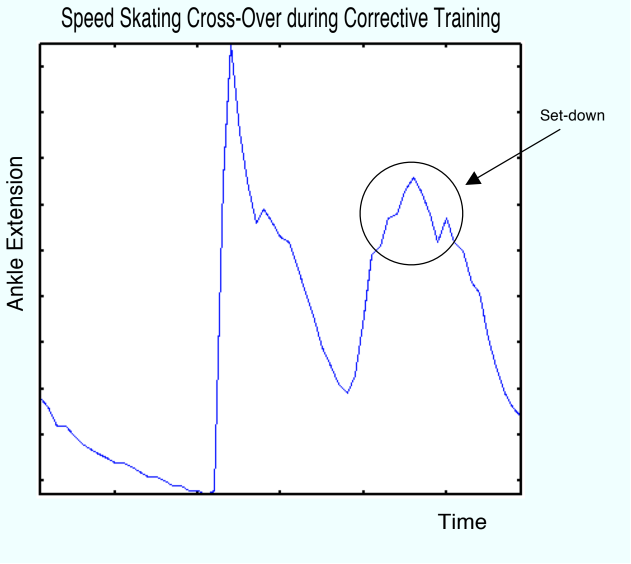


Figure 8: The skating stride of the subject during corrective training. The jagged lines during set-down indicate instability.

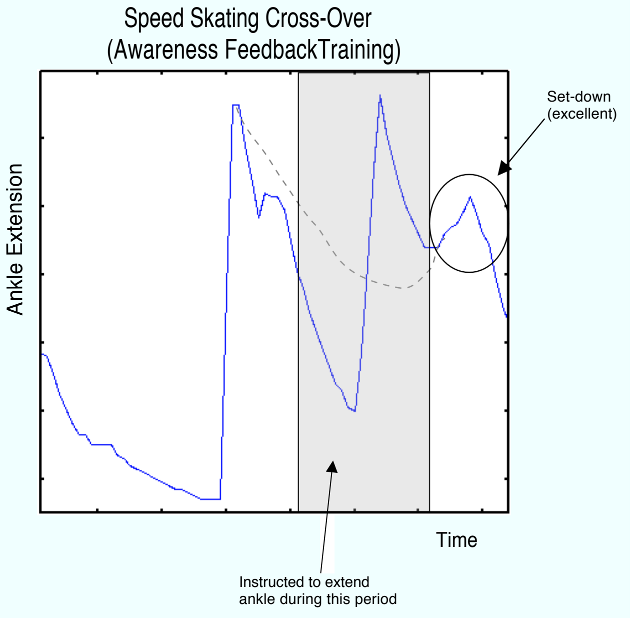


Figure 9: The skating stride of the subject during Awareness Feedback Training. Notice the smooth curves around the set-down indicating a flawless set-down. The dotted grey line represents a correct cross-over and what we hoped to achieve by reducing the purposeful ankle extension.

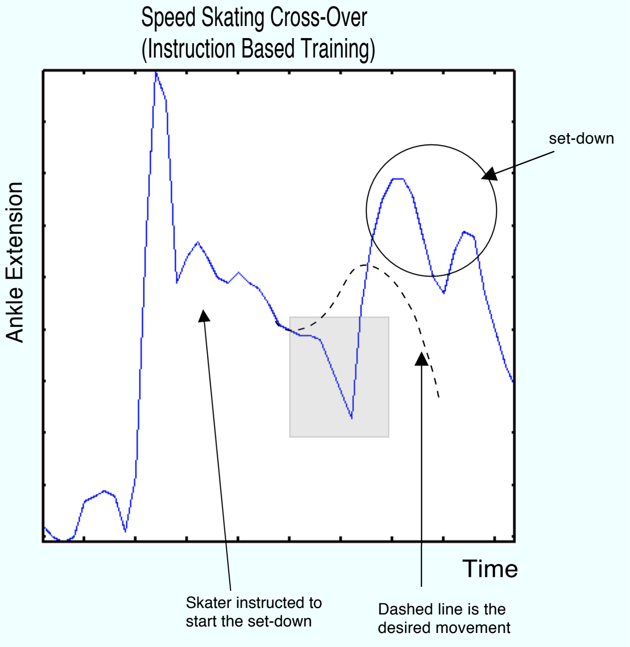


Figure 10: The skating stride of the subject during Instruction Based Training. The athlete is prompted to start the set-down process while the right skate is in the air. Notice that the athlete attempts to do this but reverts back to a more comfortable pattern, shown in the highlighted box. The dashed line indicates our desired pattern of movement. At set-down a change in the direction of the plot indicates the skater was not perfectly stable.

5.3. Awareness Feedback Training

Awareness Feedback training produced promising results. Using this training method the skater achieved some flawless set-downs, as seen in Figure 9. The skater could immediately tell that his set-downs were good and described it as the “first successful set-down in 14 months”. The skater moved at a slower pace during this training to allow time for the deliberate ankle extension. Attempts at having him skate at a faster pace while doing these extraneous motions were unsuccessful. We were also not able to replicate the flawless set-down without first doing the purposeful ankle extension. During this training method two things became clear:

- this training method fixed the problems occurring at the set-down point, and
- reducing the amount of purposeful ankle extension while maintaining a flawless set-down required a long training period.

The introduction of changes into the middle of the cross-over (the purposeful ankle extension), provided awareness to the athlete about proper ankle extension. These additional movements being new to the athlete were hard for him to control. It became obvious that we would need more training time for him to become more comfortable with the extra movement and to eventually eliminate it. Ideally we would have continued to pursue these promising results, but it did not fit the training schedule of the athlete.

5.4. Instruction Based Training

During Instruction Based Training we attempted to manufacture a right foot set-down for the subject. We observed some near flawless set-downs using this training method. This was a very challenging movement for the athlete, as we were asking him to execute a critical part of the cross-over earlier than he was accustomed to doing it. We were asking him to execute the set-down before he felt he was ready to do it. This placed a large stress on the athlete to try to execute the movement when the system wanted but also to make adjustments so that he was able to execute the movement without crashing.

The plots during this movement varied greatly depending on how the athlete was reacting to the system. In some cases like the one shown in Figure 10 we see that the athlete attempted to start the set-down but reverted back to a more comfortable movement. Results were inconsistent during this training which is expected given the drastic changes we were making to the athlete’s movements. More training time was necessary to fully evaluate this method as a solution to our subject’s problem.

5.5. Discussion

Due to time constraints with the athlete, we evaluated and proceeded through the training methods very quickly. The skater had attempted many different solutions to the problem prior to our testing. We were familiar with the level of skating the athlete had achieved with other methods and if we determined our method did not produce better results than what we had previously seen we quickly moved on.

We are confident that if we had continued working with the athlete we would have continued to see improvements in the skater’s cross-overs. The athlete tried many methods to correct his skating and about our system he commented ”This device was the only thing that was able to improve my skating.”

The athlete raced a number of times during the two months we were involved with him. The aesthetic improvements he was making did not show up as improvements in racing time. He did not wear the system during racing but did try to incorporate the same things he was working on during our training sessions. In the days leading up to a race we did not work with the athlete as he was busy with race preparations.

6. REFERENCES

[1] J. Roberts, R. Jones, N. Mansfield, and S. Rothberg, “Evaluation of impact sound on the [‘]feel’ of a golf shot,” *Journal of Sound and Vibration*, vol. 287, no. 4-5, pp. 651 – 666, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/B6WM3-4FCRFFS-2/2/307e13f9bc6a935db24b90e09b607b8f

[2] J. A. Hockman, W. M. M., and I. Fujinaga, “Real-time phase vocoder manipulation by runners pace,” in *NIME ’09: Proceedings of the 2009 conference on New interfaces for musical expression*, 2009.

[3] N. Schaffert, K. Mattes, and A. O. Effenberg, “A sound design for the purposes of movement optimisation in elite sport (using the example of rowing),” M. Aramaki, R. Kronland-Martinet, S. Ystad, and K. Jensen, Eds., Re:New Digital Arts Forum. Copenhagen, Denmark: Re:New

Digital Arts Forum, 18-21 May 2009. [Online]. Available: Proceedings/2009/SchaffertMattes2009.pdf

[4] M. Takahata, K. Shiraki, Y. Sakane, and Y. Takebayashi, “Sound feedback for powerful karate training,” in *NIME ’04: Proceedings of the 2004 conference on New interfaces for musical expression*. Singapore, Singapore: National University of Singapore, 2004, pp. 13–18.

[5] A. Effenberg, “Movement sonification: Effects on perception and action,” *Multimedia, IEEE*, vol. 12, no. 2, pp. 53–59, April-June 2005.

[6] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Upper Saddle River, NJ: Pearson Prentice Hall, 2008.

[7] R. E. Best, *Phase-locked loops design, simulation and applications*. New York: McGraw–Hill, 1999.

[8] M. Day, J. Thatcher, I. Greenlees, and B. Woods, “The causes of and psychological responses to lost move syndrome in national level trampolinists,” *Journal of Applied Sport Psychology*, vol. 18, no. 2, pp. 151–166, June 2006.

WEARABLE SETUP FOR GESTURE AND MOTION BASED CLOSED LOOP AUDIO INTERACTION

Tobias Großhauser

Ambient Intelligence Lab
CITEC - Cognitive Interaction Technology
Bielefeld University
tgrossha@techfak.uni-Bielefeld.de

Thomas Hermann

Ambient Intelligence Lab
CITEC - Cognitive Interaction Technology
Bielefeld University
thermann@techfak.uni-Bielefeld.de

ABSTRACT

The wearable sensor and feedback system presented in this paper is a type of audio-haptic display which contains onboard sensors, embedded sound synthesis, external sensors, and on the feedback side a loudspeaker and several vibrating motors. The so called “embedded sonification” in this case here is an onboard IC, with implemented sound synthesis. These are adjusted directly by the user and/or controlled in realtime by the sensors, which are on the board or fixed on the human body and connected to the board via cable or radio frequency transmission. Direct audio out and tactile feedback closes the loop between the wearable board and the user. In many situations, this setup can serve as a complement to visual output, e.g. exploring data in 3D space or learning motion and gestures in dance, sports or outdoor and every-day activities.

A new metaphor for interactive acoustical augmentation is introduced, the so called “audio loupe”. In this case it means the sonification of minimal movements or state changes, which can sometimes hardly be perceived visually or corporal. This are for example small jitters or deviations of predefined ideal gestures or movements.

Our system is easy to use, it even allows operation without an external computer. In some examples we outline the benefits of our wearable interactive setup in highly skilled motion learning scenarios in dance and sports.

1. INTRODUCTION

This project presents wearable sensing, embedded sonification and vibrotactiles, in the form of a wearable audio-haptic display. The auditory displays, according to Kramer, p. 7-10 [1], provides several advantages and even more in conjunction with other displays, in our case tactile feedback. These benefits are, according to [1] p. 9, just to name a few, “increase in perceived quality, enhanced realism, learning and creativity and lower computational requirements”.

Further the system provides real-time feedback in an acoustic and tactile form by means of closed-loop interactive sonification (see fig. 1) according to Hunt [2], Hermann [3] and haptic feedback according to MacLean [4] and Morris [5]. Information is conveyed acoustically as well as haptically and by useful combinations of both.

Our setup is a new approach and method for movement and posture measurements in 3D-space. The wearable sensing technologies further provides the user the possibility of on-site and real-time measurements of movements and postures and embedded data analysis. Many established motion capturing technologies for complex movements, most are less mobile and not wearable. One further disadvantage of the existing systems is their high complexity, for instance they demand high-speed cameras, multichannel audio systems or the fixation to a special room or laboratory. Some combinations of our system and the before mentioned ones promise interesting synergetic effects, but these are not considered in this paper.

The presented wearable device is simple and robust and very cheap compared to visual screens, projectors, multi-channel audio systems on the output side or video cameras and microphones on the input side. Furthermore it is easy to use and install. The devices can be cascaded to a complex system e.g. attached to different body parts or more than one person. In combination with data processing methods, using external computers or directly implemented in the onboard chip, the wearable multi- sensor device provides new possibilities for research in motion capturing, human comunicaton and manual learning. No complicated external cameras or CAVEs are needed. The lightweight and wearability of our system allows unhindered movements in 3D space and enables applications in many fields, such as sports, arts and multi media to name a few. Similar technology has been first demonstrated in haptic augmentations and sonifications for applications for musicians in [6] and [7], since it (a) doesn’t affect the visual sense, occupied e.g. by the communication between performers or performers and audience, (b) doesn’t disturb in bang sensitive situations such as performances, (c) allows to relate feedback information in the tactile and acoustic medium, so that these important feedback possibilities are extended and trained supportive. Even more, external instructions from the teacher, trainer and the computer or other users can be transmitted directly and unobtrusively in these audio-haptic feedback channels.

Our sensor setup is also designed for motion and gestures in general. Several approaches of interactive gestures exsist. Early approaches by Hermann [8], Verfaille et. al. [9] show first ideas and setups. In our application section we describe the use in dance, sports and ideas about data exploration. In the case of dancers, which are used to coordinate to music, sound and rhythm, sonification can depict complex dependencies between action and reaction. Accordingly these dependencies can be understood easier through listening. Examples of dance training and learning scenarios for teacher to student or self assessment and analysis for dance motor skill learning are shown.

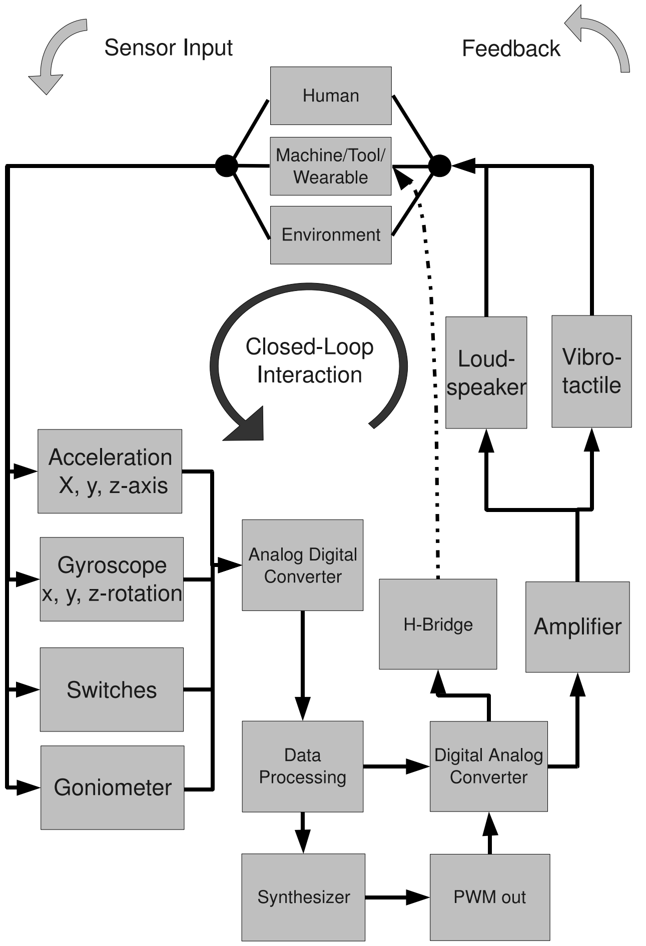


Figure 1: *Overall closed loop feedback*

2. DESCRIPTION OF THE WEARABLE DEVICE

Our wearable sensor setup is similar to the work described in [7] but here the system is not tool-integrated and consists not only of a 5 degrees-of-freedom (DOF) sensor, meaning 3 axis acceleration sensors and 2 axis gyroscopes. But also of different goniometers (see fig. 3 and 4) and shoe integrated foot switches (see fig. 2).

2.1. Sensor Setup

Our easily relocatable flexible sensor based system is divided into the following 3 parts: Input (Sensors), Atmel IC and output (loudspeakers and vibrotactiles). This wearable setup allows simple usage, with or without the need of an external computer. In this contribution different employed sensor types will be presented, to show the possibilities and usage in several application scenarios.

The sensors are:

- 6 DOF sensor, meaning 3 axis acceleration and 3 axis gyroscopes sensors.
- Foot switches integrated into shoes.
- Semi-flexible goniometer.

The whole setup can be fixed simple and situational on the body of any person or dancer and adapted to special training situations and problem statements. Due to the higher scanning frequencies of the sensors compared to most visual sensing-based approaches, such as for instance fast movements like jumps or even pirouettes can be examined at high accuracy.

2.1.1. Foot Switches

Several approaches of sensors in shoes and soles exist, mostly for medical observations and gait analysis like in [10] [11] and [12]. In this paper, two insole soft pad switches in each shoe are used for simple foot position and movement detection. Especially for jumps it is important to know, when the feet leave the floor.

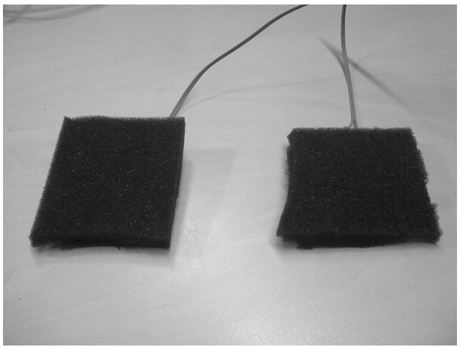


Figure 2: *Insole soft pad switches*

In our setup, two foot switches (see fig. 2) are integrated into each shoe. This allows contact detection and weight distribution of the feet and e.g. investigations on how and when "losing the contact" to the floor during jumps. The dimension of the basic sensors we use are now between 2x2x0.5cm and 5x5x0.5cm, embedded into foam plastic.

2.1.2. Goniometers

A goniometer is an instrument which measures an axis and range of motion, or the angle or rotation of an object precisely about the attached axis between two connected arms or small sticks.

Our self made goniometers are equipped with a potentiometer and used for joint angle measurement. This is a very precise and cheap sensor, easy to fix and install. The goniometers provide a high repeat accuracy, which allows usage for longer periods. Repeat accuracy here means, that the goniometers give the same start and end value before and after a dance figure or jump, without any "drift". They can be mounted directly on the body or into the clothing, depending on how precise the measurement has to be. In our case, the goniometers are fixed to the body and we used them, to investigate the spatiotemporal correlation between different body parts, e.g. foot and knee (see sec. 4.1.3).

2.1.3. Accelerometer and Gyroscopes

Two IDG-300 dual-axis angular rate gyroscopes from InvenSense are used. This allows the measurement of the rotation of the x-, y- and z-axis. Further the ADXL330 acceleration sensor from

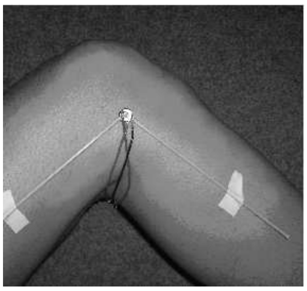


Figure 3: Flexed knee with goniometer

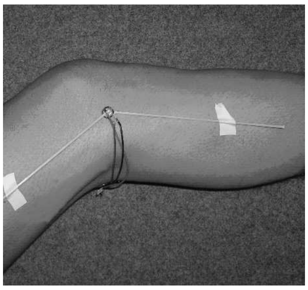


Figure 4: Streched knee with goniometer

InvenSense is used, a small, thin, low power, complete x-, y-, and z-axis accelerometer.

2.1.4. Hardware, Data Transfer and Battery

The basic setup is realized with an Atmel Atmega328 microcontroller with 14 Digital I/O Pins (of which 6 provide PWM output) and 8 analog Input Pins. The dimension is 0.73" x 1.70", (1,8 X 2,5cm) allows a small form factor and makes wearabilty easy.

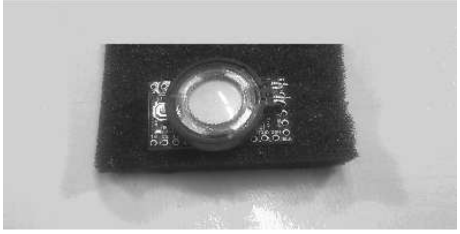


Figure 5: *Wearable PCB board with loudspeaker*

Each sensor-IC node (see fig. 5) works self-sustaining, but additional Bluetooth data transmission is possible and external peripherie like computers or more complex soundsystems can be used.

A small Lithium Polymer (LiPo) battery is directly attached for power supply. The H-Bridge is an integrated electronic circuit, which is in our case used to apply a voltage to the vibration motors and changes the speed.

3. MULTIMODAL OUTPUT AND CLOSED LOOP FEEDBACK

3.1. Sonification and Sound Synthesis

Different sound synthesis models in the area of music technology exist to generate sound and music. Beside the analog sound synthesis, various digital synthesis methods exist. The most common ones are subtractive, additive and frequency modulation synthesis. Further synthesis methods are granular-, wavetable-, phase distortion, sample-based and physical modeling synthesis, just to name a few.

In this paper, the embedded synthesizer (see scheme fig. 6) is using granular synthesis similar to [13], which works on the microsound time scale. Granular synthesis is often used sample based and in analog technology. Samples are split in small pieces of around 1 to 50 ms in length. The wearable embedded synthesizer uses oscillators instead of samples and multiple grains of these are layered on top of each other all playing at various speed, phase, volume, and pitch. Most parameters can be unfluenced with sensor input, so the scope of design is manifold.

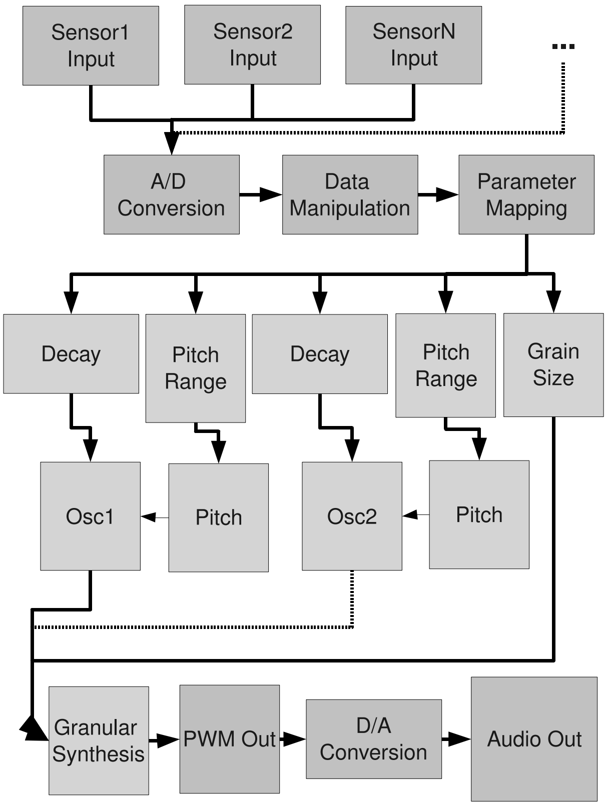


Figure 6: *Synthesizer scheme*

The result is no single tone, but a complex sound, that is subject to manipulation with our sensors and switches and the produced sounds are unlike most other synthesis techniques. By varying the waveform, envelope, duration, spatial position, and density of the grains many different sounds can be produced.

3.2. The Two Basic Sonification Modes

We discern two different sonification types according to the directness of auditory feedback.

1. Continuous Sonification: This method allows the continuous control of a movement or parts of it in real-time. The “shaping of a figure” is translated directly into a sound feedback. Especially the filter-like sound composition sounds appealing and sounds similar to popular musical effects users are used to listen to anyway.
2. Case-Triggered Sonification: This means, the sound only appears, if a certain problem or deviation appears. The

sonification can be changed and turned on and off manually, so the dancers have permanent control. This allows the individual assignment of a specific sound or sound effect to each sensor or condition, or to group useful sensor combinations.

3.3. Wearable Embedded Sonification

Our integrated and wearable devices have at least one built-in loudspeaker. If acoustical feedback occurs, the position in the 3D-space is automatically given through the sound emitted by the device. In result, no complex pointers or 2D- or 3D-sound systems are necessary to point to the relevant position. The spatial hearing of the humans allows exact and fast location of the sound source, without having to turn the head or to change any corporal position. Figuratively, every device is an active moving sound source, meeting the human habit of hearing and reacting to noises and sounds in everyday life and environments. The directional characteristic of the built in loudspeakers allows even the acoustical recognition of the gyration of the wrist, which would hardly be possible to simulate in virtual sound environments.

3.3.1. Pulse-Width Modulation, digital to analog conversion and amplification

For audio out, the Pulse-Width Modulation (PWM) outs are used. Pulse-width modulation uses a rectangular pulse wave whose pulse width is modulated resulting in the variation of the average value of the waveform. A standard digital to analog converter circuit from [14] is used to receive the analog voltage (see fig. 7). This voltage is amplified with a transistor to drive the loudspeaker

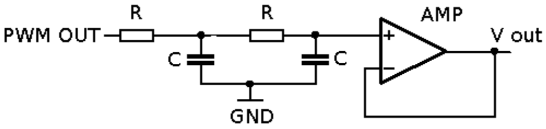


Figure 7: *Digital analog converter with amplifier*

3.3.2. Loudspeakers

One or more small speakers are used for audio out. The frequency range is quite small but the sensitivity of the human ears in the frequency range is high. It means, the sounds are good to hear and easy to locate, but the sound quality, caused by the small housing and form factor (see fig. 5) of the loudspeakers, is low.

3.4. Vibro-Tactile Feedback

3.4.1. The Vibration Motor

Several vibration devices were taken into account, including simple vibration motors, solenoid piezo-electric elements and voice coils. Besides the simple control, weight and form factor, the availability and price have been important criteria for the choice. The left vibration motor in fig. 8 with the dimensions 5x15mm, lightweight and cylindric shape seemed to be the best compromise. Furthermore, this kind of motor is typically used in mobile phones and is easy available for around 1 euro. Suitable vibration frequencies are around 250 Hz, since fingers and skin are most sensitive to these frequencies (see [15]). In this paper we present a new developed active vibrotactile feedback system, easy to use, lightweight and very flexible attachable to manifold objects and body parts. In this case two vibration motors are fixed to our board.

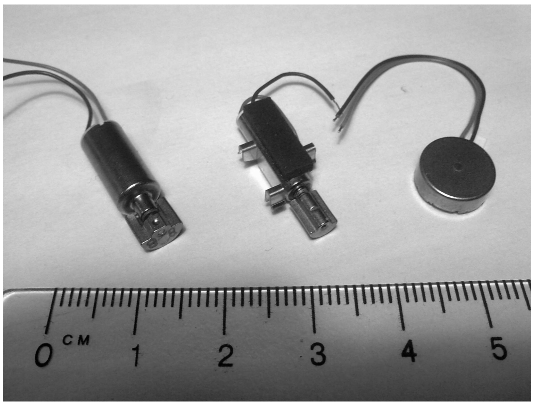


Figure 8: *Several vibration motors*

3.4.2. Listening with the Skin

We call “listening with the Skin” the awareness of local distributed and dynamically triggered vibro-tactile feedback. The vibrations are short rhythmic bursts between 40Hz and 800Hz, which is the sensitive range of the mechanoreceptors in the fingers. The distance between the two motors is big enough for easy identification which one is vibrating. The amplitude and frequency can be varied independently. This allows to evoke more or less attention, increasing and decreasing of the vibration and at least 4 significant combinations between the two motors: (1) both motors on, (2) motor 1 on, motor 2 off, (3) motor 2 on and motor 1 off and (4) both motors off. As described in Bird [16] the touch-sense feedback channel is extended and the awareness of the vibrotactile feedback is increased and trained.

3.5. Multi Channel versus Direct Sound

Compared to existing standard audio setups, especially multi channel systems, our wearable device is very simple, but very easy locateable in the 3D listening space. A simple example is, if you try to locate an alarm clock just by hearing the alarm, you know very simple and exact, where it is and the sound comes from. On the other hand, finding the exact position of a sound source in a stereo or multi channel sound field, is much more difficult and dependent of the position of the listener. If there are more than one persons, trying to describe the same source, it is already nearly impossible. If you perform this tasks with headphones, it is easier, but usually headphones are not applicable in many situations.

More advanced technologies like 3D Audio, Spatial Audio, and WFS systems improve partly the stability of the sound source, but again, the complexity and form factor of the equipment does not fit into the idea of a new, unobtrousive wearable interface.

The developed device can not only be fitted with more loudspeakers for multi channel audio out, even more than one wearable device itself can be fixed on the body or clothes. In this case, more different sounds from more directions can be provided and produce interesting interferences.

3.6. The “Audio Loupe”

A conventional loupe or magnifier glass is a [17] “type of magnification device used to see small details more closely”. In this paper we introduce the “audio loupe”, a acoustical magnifying glass for motion, posture and gesture in an auditory form. This means: Acoustical time stretching of fast motion, like an echo, and, similar to visual zooms, a acoustic zoom in- and out function to magnify or demagnify positions or movements. If for example a constant movement variies in speed, the deviation is not perceived physically, but measured and sonified. similar to an “loupe” are developed. This approach does not exclude visualisation at all, but combinations of several feedback channels in future projects might provide additional help for understanding and learning. The magnifying works precisly, especially with the goniometers. Smallest deviations of a continous movement are detected or smallest movements in stagnant postures. This is important in dance and coordination training as described in sec. 4.1.

4. APPLICATIONS

One basic idea of this new interactive interface was, to receive a 3D audio-haptic feedback in the most easy but realistic, precise and useful way. In the end, the user and performer should be able to set up the device alone, without the support of a technician. This will help to increase the acceptance of this new technologies and methods. In the following, two applications are described for dancing and the data exploration. Further interessting fields would be learning and improvising music, similar or additional to the applications and systems from Beilharz [18] and Bevilacqua [19].

4.1. Dance

As dancers are used to coordinate to music, sound and rhythm, sonification in this case can depict complex dependencies between action and reaction. Accordingly these dependencies can be understood easier through listening. Examples of dance training and learning scenarios for teacher to student or self assessment and analysis for dance motor skill learning are developed.

The most relyable data in fast motion scenarios and jumps are the goniometer data. The calculated data of the accelerometers and the gyroscopes still have a certain drift and an infeasible repeat accuracy. The professional system of XSense [20] is expensive, too large housings, and not flexible enough, especially if additional sensors are needed. The 6-DOF Board is used for tilt detection and acceleration measurements of jumps.

4.1.1. Constancy of Motion

The specific task described in the following section was the constant speed of motion in specific planes and lines and stretching.

The data in fig. 9 show two slow motion arm movements, the first one a correct movement and the second one an incorrect. The upper line is the goniometer data, the two lines below are the acceleration data. Here we have the problem, that it is hard to see the difference between the good and bad example, but the sensor data mapped to sound in realtime creates useful assistant feedback. The sensor data of differing speeds in the same motions are sonified with the following possibilities:

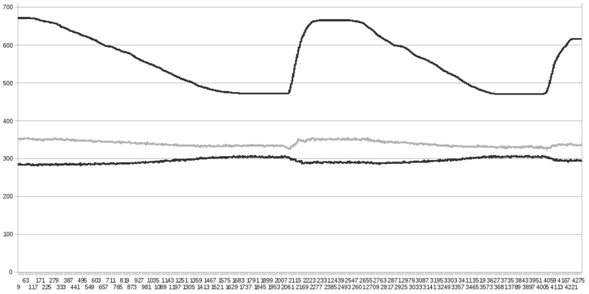


Figure 9: *Sensor data of constant correct and incorrect arm movement*

- Sonififcation position of the upper arm.
- Continous sonification of the angular rate changes of the ellbow, or the deviation of the ideal speed.
- Continous sonification of the angular rate changes of the wrist, or also again the deviation.

In other examples, the device provides feedback, if a certain point or e.g. height of the hand or foot is reached. This means a simple way of controlling the quality of the exercise or right amount of stretching is reached. Here Continous Sonification is used (ref. sec. 3.3).

4.1.2. Group Dancing and Synchronisation

In group dancing situations synchronised motion is an important issue and difficult issue to train. It is simplified by measuring the speed of the angular rate changes. Differences in speed are displayed in acoustic or tactile form. Here Case-Triggered Sonification is used (ref. sec. 3.3).

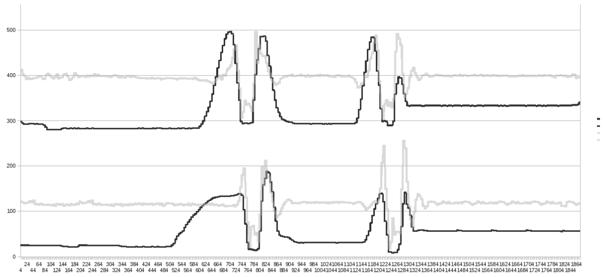


Figure 10: *Sensor data of two synchronised jumps*

The fig. 10 shows a recordings of coordinated movements. Here again, sonification of the turning points indicates the synchronisation with one or more dancers. This data are recorded with two dancers and data transmission via Bluetooth to a standard laptop computer. This allows later examination of the data, but in real life training situations, real-time audio or tactile feedback inreases training efficiency.

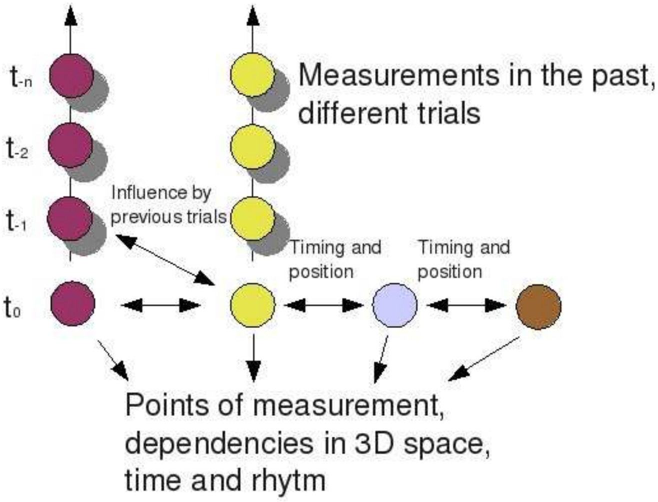
Several aspects of synchronisation and alikeness are clearly displayed by sonification:

- synchronised starting points at the beginning of the jumps
- timing of the landing

- body, foot and leg elasicity of the landing
- posture before and after the jump

4.1.3. Jumps

Another exemplary scenario is a jump in different variations. The jumps are not only of good or bad quality, they include small faults and deviations causing different results. In dance movements and jumps it is usually not only a question of “correct” or ”incorrect“, it is more a ”thinking” about complex dependencies between many parameters to find a final coherent result (see fig. 12).



t_{-n}
t_{-2}
t_{-1}
t_0
Measurements in the past, different trials
Influence by previous trials
Timing and position
Timing and position
Points of measurement, dependencies in 3D space, time and rhytm

Figure 11: *temporal, local and rhythmic dependencies*

This means, certain values are sorted e.g. knee angles or time between lift off and point of return in the air. These previous points are sonified during the next, or the further following jumps. This enables the dancer to compare different trials from the past with your current jump in realt-time. In both mentioned examples, only the maximum values are sonified, to hear the diference between the trials exactly. This acoustical repitition of the past in this case jumps, allows efficient online support and investigation possibilities during the active training phase.

The basic setup consists of 2 Goniometers, 2 foot switches, the accelerometer and gyrometer board. In fig. 12 only two of the 10 data channels, meaning two sensors of the left leg sensors are plotted. The upper line is the vertical acceleration, the lower line is the angle of the knee. It is quite hard to see the differences of the correct example (first jump) and the incorrect eaxmple (second one) and only post exploration of the data is possible. With realtime, onboard embedded sonification, the differences are more easy to investigate during the training. This is supported by recommender systems like in sec. 3.7.

Fig. 12 shows the sensor data of differing speeds and correct (the first jump) and incorrect (the second jump) body springiness distribution in the same motions, with the following sonification possibilities:

- Sonififcation of the knee bending.
- Audio cue, when the food leaves the floor.

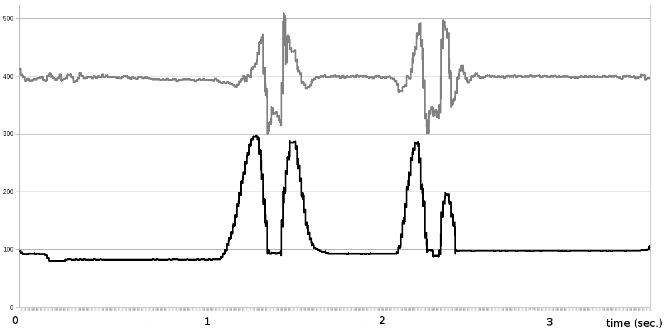


Figure 12: *2 jumps, good and bad version, acceleration versus knee angle*

- Over all acceleration sonififcation.

This measuring method also allows dancers extensive “offline” analysis of their movements, if the sensor data are saved. Sonified variations of body parts and joints, different trails with several changes of certain parameters, positive and negative progress and dependencies between all of them are shown and sonified. Audio feedback of different trials, again similar to a loupe, for professional dancers in an auditory form will provide more possibilities in the future, the longer this setup is evaluated.

Different useful combinations of important parameters in 3D-space and temporal flow are sonified. Also positive or negative skill developments between different trials and rhythmical and temporal synchronization of motion sequences are explored and sonified. Some combinations are:

- Combination of single motion points and sequences (of different trials).
- Combination of different trials with important positive or negative changes.
- Combination of different trials with important changes within longer and shorter sequences.

A further idea is the reduction of a complex movement sequence to small steps and working out of a personal “best case” scenario, to be achieved later again and more and more often after a certain amount of trials.

4.2. Sports/Every-Day Postures and Gestures

Walking with insole e.g. realised by Benocci et al. [21] and Kong [10] with several pressure senors in the sole of the shoe. Our system is simplier, as we don’t need the pressure data for our investigations. Walking, Running, “Rhythmische Gleichmäßigkeit und/oder Abweichung davon. Recommender System etc.

- Rhythmical regularity of the single steps
- Regularity of the step size
- Measurement of the constancy of the upright acceleration

Fig. 13 shows data of a leg movement while running with small rhythmical deviations from a steady running flow. The “acoustic augmentation“ of the running shows the rhythmical regularity and, even more important, the flow of the motion. For examples a typicall symptom of fatigue is irregularity of the leg motion, which can be sonified with continous sonification or case-triggered sonification, if it reacts to abrupt angular rate changes.

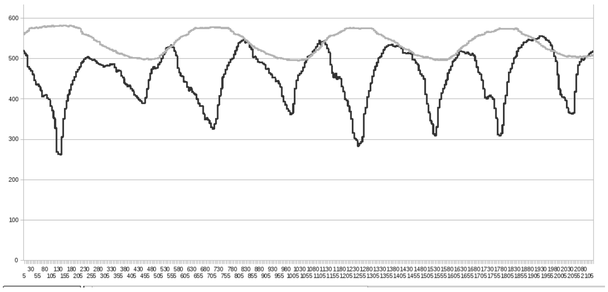


Figure 13: *Footsteps while running, with knee angle versus walking speed measurement*

5. CONCLUSION

Sonification and haptic feedback addresses, besife the visual sense, a wid range of feedback channels available. For that end we presented a multimodal audiohaptic and wearable sensor/actuator system to support human activity for many possible applications. The described way of the integration of many sensors and output possibilities are expected to have a positive effect in many learning scenarios and multi-sensorial perception. The audio-haptic feedback possibilities demonstrate that changes in movement - here in 3D-space - can be signaled unobtrusively and quite intuitively using combined haptics and audio as indexical and information carrying sign. Even real-time correction or an overdone correction can be shown.

Our first impression is that the continuous sensor data based closed-loop audio-haptic feedback described above works well and is quite efficient to direct the attention to improper executions. As promising prospect, the system may for example lead to learning aids for visually impaired people, especially as they are more biased to use their non-visual senses to compensate the lacking visual information.

The feedback helps to understand quite intuitively, how a special and complex movement is executed and trained. Further developments in augmenting both areas, the sensor and the feedback side, will show how learning processes can be improved and adapted to situated demands in everyday life situations. Especially the wrist-mounted device with the multi-modal feedback and multi sensory input is adaptable to different scenarios such as in sports, music, dance, games and many more interaction scenarios. Also interactive music systems for improvisation are considered with this setup.

The ”Audio loupe“ is a promising method in the field of high level dance and motor skill learning especially for examination and monitoring of progress and hard to understand complex movements and dependencies.

6. ACKNOWLEDGEMENT

The work has been supported by CITEC – Center of Excellence in Cognitive Interaction Technology.

7. REFERENCES

[1] G. Kramer, *An Introduction to auditory display*, G. Kramer, Ed. Santa Fe Institute Studies in the Sciences of Complexity, Proceedings vol. XVIII. Addison-Wesley, Reading, MA, 1994, vol. Auditory Display: Sonification, Audification, and Auditory Interfaces.

[2] A. Hunt, T. Hermann, and S. Pauletto, “Interacting with sonification systems: Closing the loop,” *Information Visualisation, International Conference on*, vol. 0, pp. 879–884, 2004.

[3] T. Hermann and A. Hunt, “Guest editors’ introduction: An introduction to interactive sonification,” *IEEE MultiMedia*, vol. 12, no. 2, pp. 20–24, 2005.

[4] K. E. MacLean, “Designing with haptic feedback,” in *IEEE International Conference on Robotics and Automation, 2000. Proceedings. ICRA ’00. IEEE International Conference on*, vol. vol.1, 2000, pp. 783 – 788.

[5] D. Morris, H. Tan, F. Barbagli, T. Chang, and K. Salisbury, “Haptic feedback enhances force skill learning,” in *WHC ’07: Proceedings of the Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 21–26.

[6] T. Grosshauser and T. Hermann, “The sonified music stand – an interactive sonification system for musicians,” in *Proceedings of the 6th Sound and Music Computing Conference*, X. S. Fabien Gouyon, Alvaro Barbosa, Ed. Casa da Musica, Porto, Portugal, 2009, pp. 233–238. [Online]. Available: http://smc2009.smcnetwork.org/programme/pdfs/238.pdf

[7] ——, “Augmented haptics - an interactive feedback system for musicians,” in *Haptic and Audio Interaction Design, 4th International Conference, HAID 2009, Dresden, Germany, September 10-11, 2009, Proceedings*, ser. Lecture Notes in Computer Science, M. E. Altinsoy, U. Jekosch, and S. A. Brewster, Eds., vol. 5763. Springer, September 2009, pp. 100–108. [Online]. Available: http://www.springerlink.com/content/688162j31k304h84/

[8] T. Hermann, C. Nölker, H. Ritter, C. T. Hermann, A. T. Hermann, U. Bielefeld, T. Fakultät, A. Neuroinformatik, T. Hermann, C. Nölker, and H. Ritter, “Hand postures for sonification control (extended abstract),” 2001.

[9] V. Verfaille, O. Quek, and M. M. W, “Sonification of musicians’ ancillary gestures,” 2005.

[10] S.-J. KONG, K. Chul-Seung, and G.-M. EOM, “Portable gait-event detection system with built-in wireless sensor configuration,” in *11th Annual Conference of the International FES Society*, Zao, Japan, 2006.

[11] S. J. M. Bamberg, A. Y. Benbasat, D. M. Scarborough, D. E. Krebs, and J. A. Paradiso, “Gait analysis using a shoe-integrated wireless sensor system,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 4, pp. 413–423, 2008.

[12] X. S. Bufu Huang, Meng Chen and Y. Xu, “Gait event detection with intelligent shoes,” in *Proceedings of the 2007 International Conference on Information Acquisition*, Jeju City, Korea, July 9-11, 2007.

[13] "http://code.google.com/p/tinkerit/," January, 31th 2010.

[14] "http://www.mikrocontroller.net/wikifiles/1/1f/pwm_filter_1 .png," internet, January, 31th 2010.

[15] M. T. Marshall and M. M. Wanderley, "Vibrotactile feedback in digital musical instruments," in *NIME '06: Proceedings of the 2006 conference on New interfaces for musical expression*. Paris, France, France: IRCAM — Centre Pompidou, 2006, pp. 226–229. [Online]. Available: http://recherche.ircam.fr/equipes/temps-reel/nime06/proc/nime2006_226.pdf

[16] J. Bird, S. Holland, P. Marshall, Y. Rogers, and A. Clark, "Feel the force: Using tactile technologies to investigate the extended mind," in *Proceedings of Devices that Alter Perception*, 2008. [Online]. Available: http://www.k2.t.u-tokyo.ac.jp/perception/dap2008/papers/Birddap2008.pdf

[17] [Online]. Available: 02.02.2010,http://en.wikipedia.org/wiki/Loupe

[18] K. Beilharz and S. Ferguson, "Gestural hyper instrument collaboration with generative computation for real time creativity," in *C&C '07: Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition*. New York, NY, USA: ACM, 2007, pp. 213–222.

[19] F. Bevilacqua, F. Guédy, N. Schnell, E. Fléty, and N. Leroy, "Wireless sensor interface and gesture-follower for music pedagogy," in *NIME '07: Proceedings of the 7th international conference on New interfaces for musical expression*. New York, NY, USA: ACM, 2007, pp. 124–129.

[20] [Online]. Available: 02.02.2010,http://www.xsens.com/en/general/mvn

[21] M. Benocci, L. Rocchi, E. Farella, L. Chiari, and L. Benini, "A wireless system for gait and posture analysis based on pressure insoles and inertial measurement units," in *4th International ICST Conference on Pervasive Computing Technologies for Healthcare*. University of Bologna, 2009.

SONIFICATION OF ACE LEVEL 2 SOLAR WIND DATA

Robert Alexander, Thomas H. Zurbuchen, Jason Gilbert, Susan Lepri, Jim Raines

Department of Atmospheric, Oceanic and Space Sciences,
University of Michigan, Ann Arbor, Mi, USA
robertalexandermusic@gmail.com

ABSTRACT

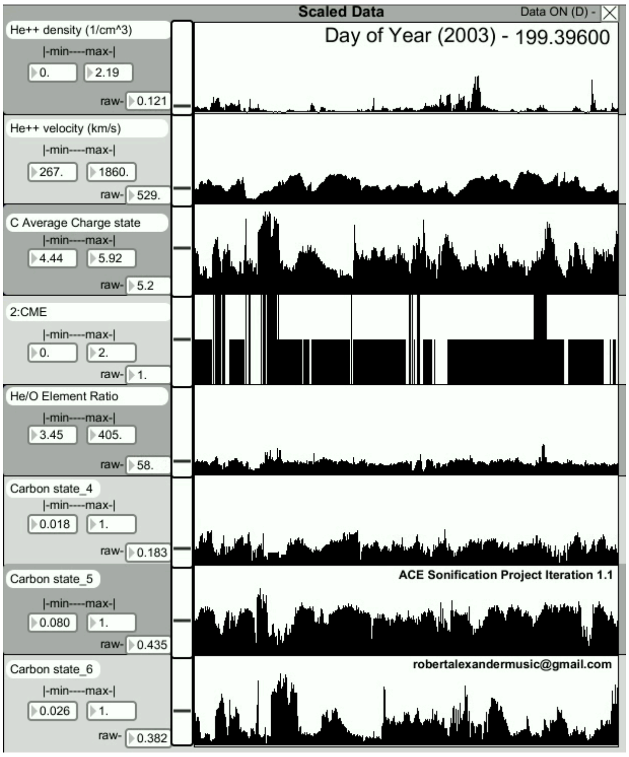
This paper provides a brief overview of the sonification research conducted by the Solar and Heliospheric Research Group at the University of Michigan. The team collaborated with composer and multimedia artist Robert Alexander to gain a new perspective of the underlying patterns behind recurring solar wind phenomena. This sonification effort was one in which a high level of creative freedom was provided to the composer, while scientific accuracy was maintained through adherence to the original data set. An interface was constructed in Max/MSP that allowed ACE-SWICS Level 2 solar wind data to be graphed visually and represented aurally through both acoustic and synthesized timbres. This document will explore the sonification methods behind iteration 1.1, which is a sonification of solar wind activity from 2003.

1. INTRODUCTION

The solar wind, originating from the Sun and carried into interplanetary space, is highly dynamic and punctuated by abrupt explosive events. This dynamic nature provides the ideal medium to experiment with sonification. Iteration 1.1 is well suited for absorbing data over a long period of time. The listener should relax and allow their attention to drift between the various sounds. The algorithm has been refined to create a balance between all parts. This balance is punctuated by Coronal Mass Ejection (CME) events, which are explosions of vast amounts of material and energy from the Sun.

Upon first experiencing the iteration, it is possible to deduce the audio-visual correlation by closely viewing each individual data parameter as seen in Figure 1. These data parameters have been scaled to make full use of the visual space, and the maximum and minimum values used for this scaling are displayed below the data type. The original value is displayed in a number box to the left of the graphic representation. The data points included in this sonification are:

1) Helium (He++) density (1/cm^3)
2) He++ speed (km/s)
3) Carbon average charge state
4) Solar Wind Type (0. Streamer Wind 1. Coronal Hole Wind 2. Coronal Mass Ejection).
5) Helium to Oxygen (He/O) element ratio
6) Carbon charge state 4+
7) Carbon charge state 5+
8) Carbon charge state 6+



Scaled Data

Data ON (D) -

Day of Year (2003) - 199.39600

He++ density (1/cm^3) |-min----max-| 0. 2.19 raw- 0.121

He++ velocity (km/s) |-min----max-| 267. 1860. raw- 529.

C Average Charge state |-min----max-| 4.44 5.92 raw- 5.2

2:CME |-min----max-| 0. 2. raw- 1.

He/O Element Ratio |-min----max-| 3.45 405. raw- 58.

Carbon state_4 |-min----max-| 0.018 1. raw- 0.183

Carbon state_5 |-min----max-| 0.080 1. raw- 0.435

Carbon state_6 |-min----max-| 0.026 1. raw- 0.382

ACE Sonification Project Iteration 1.1

robertalexandermusic@gmail.com

Figure 1: Visual representation of sonified data.

From Figure 1, the reader can see that some quantities vary more smoothly and are less variable, while other quantities vary quite drastically and often in unison with other parameters.

2. IMPLEMENTATION IN MAX/MSP

The data file used in this sonification contains a combination of 2-hour averaged solar wind parameters and 2-hour averaged charge-state distributions from the Solar Wind Ion Composition Spectrometer (SWICS) on the Advanced Composition Explorer (ACE), which orbits at the gravitational saddle point between the Sun and the Earth. These data sets are provided to the public by the ACE Science Center (ASC). The data is loaded into the “text” object in MAX and the “line” message is used to represent successive data entries. The minimum and maximum values of each parameter are used to scale the data, which is then plotted. Any 3 entries can be plotted next to one another for comparison purposes, and selected data can then be sent directly to the sonification section of the patch.

3. SONIFICATION METHODOLOGY

The section of the patch that was devoted to sonification began as a relatively blank-slate, such that ideas could be quickly implemented. This multi-layered sonification was constructed through an iterative process of experimentation with various data mappings. This section will deconstruct each element of the sonification.

The sweeping wind sound is generated by both He++ density and speed. The speed parameter controls the cutoff frequency of a band-pass filter, which causes the "whooshing" noise that sweeps up and down. The density parameter controls the loudness of this wind sound, i.e., the higher the density the louder the wind. The use of filtered noise creates a sound that is reminiscent of terrestrial wind phenomena. During a CME, the wind is further amplified and processed with a form of distortion known as overdrive. This causes the wind to swell in a more violent fashion. One particularly interesting moment occurs at 3:05, during an extended CME.

The different charge states of Carbon provide information on the temperature of the corona; higher charge states originate in a hotter region of the corona, often associated with CMEs. The basic vocal ambience layer is created with 6 distinct vocal layers that each correspond to a charge state of carbon. A recording of a female voice (alto vocalist Amanda Alexander) was conducted in a small room with a condenser microphone. Each note was recorded individually, and each file was subsequently edited to create one long extremely smooth tone. The prevalence of one charge state over the other, as determined by the distribution ratio, is used to modulate the gain of each vocal layer. For example, charge state 6+ (the bottom box) corresponds to the higher voices, which are panned to the left ear (this is easier to distinguish on headphones). As this charge state becomes more predominant, the higher voice will stick out.

Carbon charge states 4+ and 5+ are easily discernible by listening for the absence of charge state 6+. They occur as lower sets of voices that are panned to the right. The pitch cluster is reinforced by a set of sinusoidal tones at the same frequencies; the volume of these tones is linked to the predominance of Carbon charge state 4+. These tones are quite soft, but their presence significantly adds to the texture of the sonification.

The value "C Average Charge State" is represented by another set of voices that sing in a higher octave. The easiest way to pick these voices out is to listen during a Coronal Mass Ejection event; at this time the highest of these voices bends upward in pitch. A chord composed of an extremely high frequency set of triangle waveforms represents the He/O element ratio. This sound can be described as a "glistening," it clearly stands out during the CME at 3:07.

Solar wind type is the most readily discernible feature in this sonification. During a CME the reverb quickly swells to a much higher volume before slowly attenuating back to the original volume, which creates the feeling of a sudden expanse. The difference between Streamer wind and Coronal Hole Wind is subtler. During streamer wind the level of reverb is further attenuated, and the chanting vocal ostinato is cut completely. The soft vocal chanting layer doubles in loudness during a CME. The bass-line is also played two octaves higher during a CME; this sound quite muffled.

A Low-Frequency bass tone was generated with a saw-tooth waveform that traveled algorithmically between a pre-determined group of pitches. The changes in the bass not only mark time, but also provide a sense of forward momentum through harmonic progression (the movement of the bass creates a pseudo-random progression between I, ii, IV, and vi chords in C major). The change in pitch happens once every half sidereal Carrington rotation; two changes in pitch mark one full sidereal Carrington rotation (25.38 days). To further demarcate the rotation process, an automated low-pass filter was applied. The cutoff frequency of this low-pass filter travels up and down with one full rotation. During one half-rotation the bass sound becomes muffled (the cutoff frequency is lowered), and during the other half it is slowly un-muffled (the cutoff frequency is raised).

4. CONCLUSION AND FUTURE DIRECTION

The team was able to hear complex interactions between multiple data entries, but has yet to unearth any new findings from initial experimentations. The sonification work resulting from this project has gained wide attention due to its aesthetic appeal and scientific potential. For future iterations the team is interested in taking on larger time scales. The number of active sunspots could provide a potentially compelling arc in the sonification of a complete solar cycle. This sonification effort is still in its early stages, and the team is hopeful about the potential of future work in this area.

The technique has the potential for impact in two distinct ways. First, making numerical solar wind data into music can make it much more accessible the public at large. This is a long standing difficulty with data of this sort: While movies of solar explosions sometimes make into the evening news, few would consider including a bunch of wiggly lines. This project has already made strides in this direction, including posting online videos that includes solar wind composition data. With continued work, future dramatic sonifications could further inform non-scientists about the complex behavior of the sun.

Second, sonification has the potential to advance this scientific field by helping researchers to find patterns and features in the data that went undetected through other means. Humans have a well-refined ability to appreciate complex sonic environments and pick out individual details. The human brain has powerful pattern-detecting mechanisms; many scientific leaps in the past have sprung directly from human intuition. Custom-designed software tools that enabled researchers to build and vary sonifications in near real time, much like is currently done with visualizations, could potentially improve scientific understanding of the data and lead to new ideas for exploration.

SONIFICATION OF ACE LEVEL 2 SOLAR WIND DATA

Robert Alexander, Thomas H. Zurbuchen, Jason Gilbert, Susan Lepri, Jim Raines

Department of Atmospheric, Oceanic and Space Sciences,
University of Michigan, Ann Arbor, Mi, USA
robertalexandermusic@gmail.com

ABSTRACT

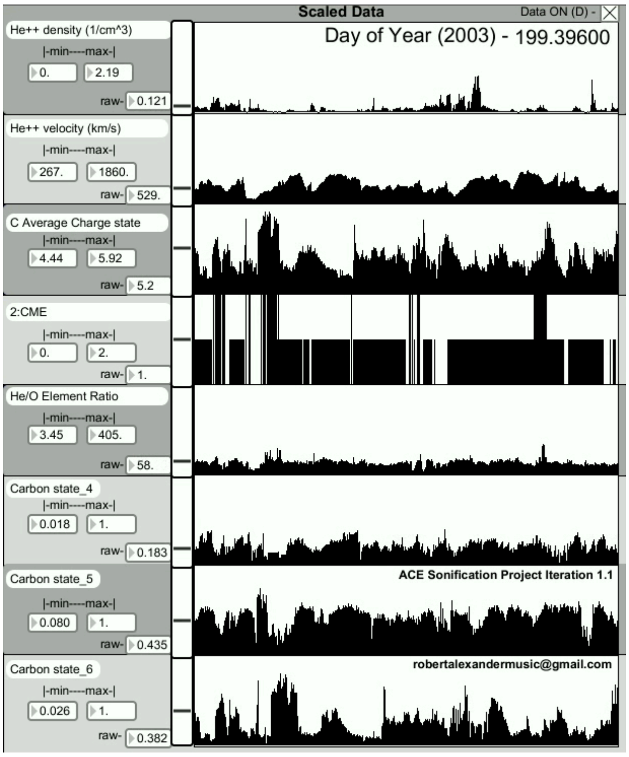
This paper provides a brief overview of the sonification research conducted by the Solar and Heliospheric Research Group at the University of Michigan. The team collaborated with composer and multimedia artist Robert Alexander to gain a new perspective of the underlying patterns behind recurring solar wind phenomena. This sonification effort was one in which a high level of creative freedom was provided to the composer, while scientific accuracy was maintained through adherence to the original data set. An interface was constructed in Max/MSP that allowed ACE-SWICS Level 2 solar wind data to be graphed visually and represented aurally through both acoustic and synthesized timbres. This document will explore the sonification methods behind iteration 1.1, which is a sonification of solar wind activity from 2003.

1. INTRODUCTION

The solar wind, originating from the Sun and carried into interplanetary space, is highly dynamic and punctuated by abrupt explosive events. This dynamic nature provides the ideal medium to experiment with sonification. Iteration 1.1 is well suited for absorbing data over a long period of time. The listener should relax and allow their attention to drift between the various sounds. The algorithm has been refined to create a balance between all parts. This balance is punctuated by Coronal Mass Ejection (CME) events, which are explosions of vast amounts of material and energy from the Sun.

Upon first experiencing the iteration, it is possible to deduce the audio-visual correlation by closely viewing each individual data parameter as seen in Figure 1. These data parameters have been scaled to make full use of the visual space, and the maximum and minimum values used for this scaling are displayed below the data type. The original value is displayed in a number box to the left of the graphic representation. The data points included in this sonification are:

1) Helium (He++) density (1/cm^3)
2) He++ speed (km/s)
3) Carbon average charge state
4) Solar Wind Type (0. Streamer Wind 1. Coronal Hole Wind 2. Coronal Mass Ejection).
5) Helium to Oxygen (He/O) element ratio
6) Carbon charge state 4+
7) Carbon charge state 5+
8) Carbon charge state 6+


Scaled Data
Data ON (D) -
Day of Year (2003) - 199.39600
ACE Sonification Project Iteration 1.1
robertalexandermusic@gmail.com

He++ density (1/cm^3)
|-min----max-|
0. 2.19
raw- 0.121

He++ velocity (km/s)
|-min----max-|
267. 1860.
raw- 529.

C Average Charge state
|-min----max-|
4.44 5.92
raw- 5.2

2:CME
|-min----max-|
0. 2.
raw- 1.

He/O Element Ratio
|-min----max-|
3.45 405.
raw- 58.

Carbon state_4
|-min----max-|
0.018 1.
raw- 0.183

Carbon state_5
|-min----max-|
0.080 1.
raw- 0.435

Carbon state_6
|-min----max-|
0.026 1.
raw- 0.382

Figure 1: Visual representation of sonified data.

From Figure 1, the reader can see that some quantities vary more smoothly and are less variable, while other quantities vary quite drastically and often in unison with other parameters.

2. IMPLEMENTATION IN MAX/MSP

The data file used in this sonification contains a combination of 2-hour averaged solar wind parameters and 2-hour averaged charge-state distributions from the Solar Wind Ion Composition Spectrometer (SWICS) on the Advanced Composition Explorer (ACE), which orbits at the gravitational saddle point between the Sun and the Earth. These data sets are provided to the public by the ACE Science Center (ASC). The data is loaded into the “text” object in MAX and the “line” message is used to represent successive data entries. The minimum and maximum values of each parameter are used to scale the data, which is then plotted. Any 3 entries can be plotted next to one another for comparison purposes, and selected data can then be sent directly to the sonification section of the patch.

3. SONIFICATION METHODOLOGY

The section of the patch that was devoted to sonification began as a relatively blank-slate, such that ideas could be quickly implemented. This multi-layered sonification was constructed through an iterative process of experimentation with various data mappings. This section will deconstruct each element of the sonification.

The sweeping wind sound is generated by both He++ density and speed. The speed parameter controls the cutoff frequency of a band-pass filter, which causes the "whooshing" noise that sweeps up and down. The density parameter controls the loudness of this wind sound, i.e., the higher the density the louder the wind. The use of filtered noise creates a sound that is reminiscent of terrestrial wind phenomena. During a CME, the wind is further amplified and processed with a form of distortion known as overdrive. This causes the wind to swell in a more violent fashion. One particularly interesting moment occurs at 3:05, during an extended CME.

The different charge states of Carbon provide information on the temperature of the corona; higher charge states originate in a hotter region of the corona, often associated with CMEs. The basic vocal ambience layer is created with 6 distinct vocal layers that each correspond to a charge state of carbon. A recording of a female voice (alto vocalist Amanda Alexander) was conducted in a small room with a condenser microphone. Each note was recorded individually, and each file was subsequently edited to create one long extremely smooth tone. The prevalence of one charge state over the other, as determined by the distribution ratio, is used to modulate the gain of each vocal layer. For example, charge state 6+ (the bottom box) corresponds to the higher voices, which are panned to the left ear (this is easier to distinguish on headphones). As this charge state becomes more predominant, the higher voice will stick out.

Carbon charge states 4+ and 5+ are easily discernible by listening for the absence of charge state 6+. They occur as lower sets of voices that are panned to the right. The pitch cluster is reinforced by a set of sinusoidal tones at the same frequencies; the volume of these tones is linked to the predominance of Carbon charge state 4+. These tones are quite soft, but their presence significantly adds to the texture of the sonification.

The value "C Average Charge State" is represented by another set of voices that sing in a higher octave. The easiest way to pick these voices out is to listen during a Coronal Mass Ejection event; at this time the highest of these voices bends upward in pitch. A chord composed of an extremely high frequency set of triangle waveforms represents the He/O element ratio. This sound can be described as a "glistening," it clearly stands out during the CME at 3:07.

Solar wind type is the most readily discernible feature in this sonification. During a CME the reverb quickly swells to a much higher volume before slowly attenuating back to the original volume, which creates the feeling of a sudden expanse. The difference between Streamer wind and Coronal Hole Wind is subtler. During streamer wind the level of reverb is further attenuated, and the chanting vocal ostinato is cut completely. The soft vocal chanting layer doubles in loudness during a CME. The bass-line is also played two octaves higher during a CME; this sound quite muffled.

A Low-Frequency bass tone was generated with a saw-tooth waveform that traveled algorithmically between a pre-determined group of pitches. The changes in the bass not only mark time, but also provide a sense of forward momentum through harmonic progression (the movement of the bass creates a pseudo-random progression between I, ii, IV, and vi chords in C major). The change in pitch happens once every half sidereal Carrington rotation; two changes in pitch mark one full sidereal Carrington rotation (25.38 days). To further demarcate the rotation process, an automated low-pass filter was applied. The cutoff frequency of this low-pass filter travels up and down with one full rotation. During one half-rotation the bass sound becomes muffled (the cutoff frequency is lowered), and during the other half it is slowly un-muffled (the cutoff frequency is raised).

4. CONCLUSION AND FUTURE DIRECTION

The team was able to hear complex interactions between multiple data entries, but has yet to unearth any new findings from initial experimentations. The sonification work resulting from this project has gained wide attention due to its aesthetic appeal and scientific potential. For future iterations the team is interested in taking on larger time scales. The number of active sunspots could provide a potentially compelling arc in the sonification of a complete solar cycle. This sonification effort is still in its early stages, and the team is hopeful about the potential of future work in this area.

The technique has the potential for impact in two distinct ways. First, making numerical solar wind data into music can make it much more accessible the public at large. This is a long standing difficulty with data of this sort: While movies of solar explosions sometimes make into the evening news, few would consider including a bunch of wiggly lines. This project has already made strides in this direction, including posting online videos that includes solar wind composition data. With continued work, future dramatic sonifications could further inform non-scientists about the complex behavior of the sun.

Second, sonification has the potential to advance this scientific field by helping researchers to find patterns and features in the data that went undetected through other means. Humans have a well-refined ability to appreciate complex sonic environments and pick out individual details. The human brain has powerful pattern-detecting mechanisms; many scientific leaps in the past have sprung directly from human intuition. Custom-designed software tools that enabled researchers to build and vary sonifications in near real time, much like is currently done with visualizations, could potentially improve scientific understanding of the data and lead to new ideas for exploration.

EFFECTS OF INTERFACE TYPE ON NAVIGATION IN A VIRTUAL SPATIAL AUDITORY ENVIRONMENT

Agnieszka Roginska[1], *Gregory H. Wakefield*[2], *Thomas S. Santoro*[3], *Kyla McMullen*[2]

[1]Music and Audio Research Lab, New York University, 35 West 4th St, New York, NY 10012

[2]EECS Department, The University of Michigan, Ann Arbor, MI 48109

[3]Naval Submarine Medical Research Lab, SUBASE NLON, Groton, CT 06349

roginska@nyu.edu, ghw@umich.edu, thomas.santoro@med.navy.mil, kyla@umich.edu

ABSTRACT

In the design of spatial auditory displays, listener interactivity can promote greater immersion, better situational awareness, reduced front/back confusion, improved localization, and greater externalization. Interactivity between the listener and their environment has traditionally been achieved using a head tracker interface. However, trackers are expensive, sensitive to calibration, and may not be appropriate for use in all physical environments. Interactivity can be achieved using a number of alternative interfaces. This study compares learning rates and performance in a single-source auditory search task for a head-tracker and a mouse/keyboard interface within a single source and multi-source context.

1. INTRODUCTION

The use of auditory cues to help navigators explore unfamiliar environments is of ancient origin. Horns have led ships through the foggy seas just as more contemporary portable sound devices have led the blind through urban environments (e.g. [4][5][9]). Similarly, spatial auditory displays can be used to communicate spatial information about a virtual environment to the user. In this type of interface, locations are represented as sound sources, and a user may navigate and explore this virtual environment as they would the more familiar natural environment.

An important factor in fully-immersive systems is the degree to which the participant and the virtual environment interact at the participant's sensorimotor level. Interaction supports the participant's active exploration of their environment through which they become better oriented spatially and can, therefore, navigate more accurately.

One of the challenges with virtual spatial audio is the type of interface used to inject the user into the virtual world. Hardware for sensing head orientation and position has been used extensively as a means to track users. However, there are several issues associated with the use of head-tracking systems – they are expensive, susceptible to calibration issues, require more specialized application development, and can't always be used in all physical environments. In addition, due to the fact that many users are unfamiliar with the interface, the head tracker involves training. An alternate interface is desirable, which doesn't compromise the user experience or performance. Although the sensorimotor integration between changes induced in a spatial audio display by other interfaces' motion may be less natural, we hypothesize that similar performance can be achieved with alternate interfaces and correspond to an equally compelling experience as with a head tracker.

In this study, we propose the use of an *avatar* interface as an alternative. The interface involves a mouse and keyboard as a means to navigate through and interact with an environment. The mouse controls the x/y position of the listener, and the keyboard controls their orientation.

We compare the use of the mouse/keyboard interface to the head tracker interface in a search and navigation task. We focus on comparing human performance in a search task of a single source within a single- or multi-source environment with both interfaces. We explore the differences in the use of the two interfaces. Through a subjective experiment, we look at how participants learn to use these interfaces, the effect each interface has on their performance, and search strategies developed and used by the participant during a search task.

2. EXPERIMENT

An experiment was designed to assess the extent to which auditory search in a virtual acoustic environment (VAE) can be mediated through an avatar interface. The VAE was comprised of acoustic sources arranged along a circle in an otherwise anechoic environment. Participants could move and orient through this environment either directly, by walking and turning their head, or indirectly, by moving the location and angular orientation of an avatar on a computer display presented in a top-down perspective. In what follows, the former will be called *natural mediation* and the latter will be called *avatar mediation* of user position in the VAE.

The task required that participants locate a source in the VAE by moving to the location of that source. To acclimate participants to the apparatus, the experiment was conducted in two phases. During the training phase, a single source was presented during a trial and the participant moved from the center of the circle to the location of the source as quickly as possible. During the test phase, four sources were presented during a trial and the participant was to move to the location of each source until all four sources were found.

Because it draws upon the standard means by which we, as listeners, navigate through our environment, we hypothesize that natural-mediated search will require fewer trials than avatar-mediated search to reach asymptote for the training phase. Nevertheless, because both forms of mediation engage a common representation of auditory space, we expect that the asymptotic search strategies of each will be similar.

When multiple sources are present, it is not clear how search times should be affected. An increase in the time it takes to locate the first source would be expected if the presence of multiple sources interferes with the cues used to locate any one source. Alternatively, a participant may choose to minimize total search time by using a portion of their first search to establish a general mapping of all the sources before moving to the first source. In the absence of interference or a global strategy, the time it takes to locate the first source during the test phase should be the same as the asymptote reached during the training phase.

Finally, we are interested in whether some users are generally faster than others when performing an auditory search and in the strategies they use. Accordingly, each participant was tested under both forms of mediation. Half the subjects were trained and tested first under natural mediation, before going on to training and testing under avatar mediation, while the other half underwent initial training and testing under avatar mediation. We hypothesize that experience in either modality (natural or avatar mediation) will transfer to the other as evidenced in fewer trials to reach asymptote when shifting to the alternative modality and that there will be a high degree of correlation between fastest and slowest performers across modality.

2.1. Procedure

For both training and test phases of the experiment, a trial began with a source (or sources) positioned randomly along a fixed circle placed horizontally in the 0-degree elevation plane and the participant positioned in the center of that circle. Participants were notified by a diotic auditory cue when they arrived within a fixed radius of the source. During the training phase, a single source was presented and the participant re-centered him or herself after notification to begin another trial. Training continued until a participant's current and past four search times had a standard deviation of 2.5 seconds or less.

The test phase consisted of four sources. At the beginning of a trial, participants were informed which source they should search for first by a diotically-presented four-second sample of the selected source. Following the cue, the four sources were presented and the participant began their search. Upon successfully locating the first source, the sources were turned off, and the second cue was presented, after which the sources were turned back on again. This sequence continued until all four sources were located.

Once a participant finished both the test and training phases for one modality, they repeated the procedures for the alternate modality.

2.2. Apparatus

Avatar mediation was controlled by a mouse/keyboard interface. A mouse controlled the position of the participant in the acoustic space. The left and right arrow keys controlled the yaw of the participant's head, in steps of two degrees. Natural mediation was controlled by a Polhemus Liberty electromagnetic 6DOF system head tracker, with a 240Hz update rate. The sensor was mounted at the center of the headphone band worn by the participant. The tracker emitter was mounted at the end of an arm placed at least 0.5 m above the participant's head. The system was maximally sensitive to within a 1.5-m radius sphere of the emitter, which is similar in size to a CAVE.

Yaw and position were sampled at a rate of 10 Hz to drive a real-time spatial audio system programmed in Matlab using HRIRs obtained from KEMAR using the NSMRL measurement facility [2]. Audio streaming was implemented as follows:

- A Matlab timer was programmed to generate a new frame of audio based on the participant's current position in the virtual environment. The timer called on routines to convolve audio input read from disk using the appropriate yaw-adjusted interpolated HRIRs and adjustment in position-dependent gain.
- Audio was controlled through the PsychToolbox extension of OpenAL by double buffering. The same Matlab timer was responsible for querying the OpenAL source to determine when one of its (two) buffers had finished playing. The next frame of audio was then loaded into the spent buffer and re-queued.

HRIR interpolation was implemented by constructing the minimum phase impulse response of a system whose magnitude spectrum is determined from a log mixture of the adjacent measured HRTFs (sampled every 10 degrees) and convolving the result with an all-phase system using a fractional-delay method.

Stimuli were presented over Sennheiser open ear HD650 headphones. The listening room was a sound-treated standard 4.5m x 7m acoustic research space in the Music Technology program at New York University. Depending on the condition, participants were either seated before the computer console in which a window with a listener icon was displayed or standing in the middle of the room beneath the Polhemus emitter.

2.3. Sources and VAE

Four sources were selected from a publicly available database of audio recordings [7]. Because the present experiment is the first in a broader study on auditory-guided search through multiple-source virtual environments, the sources were chosen to be (1) sufficiently varying in spectro-temporal features, (2) mutually discriminable, and (3) mutually *inconsistent*, e.g., unlikely to be commonly occurring together in naturally occurring acoustic environment. Among the variety of options, we selected recordings of a **typewriter, street crowd**, **brook**, and **electronic sounds**, as might be heard in a piece of computer music. Each recording was between 23 and 80 seconds in duration and repeated continuously.

Stimulus levels were balanced by one of the authors using method of adjustment to achieve equal sensation level by determining the detection threshold of one source in the presence of the other three when presented diotically. These levels were confirmed by informal listening among all authors.

An inverse square law was used to determine the amplitude of the source as the participant moved through the environment. The dimension of the circle in the VAE was scaled so that there was a 13 dB drop in gain for a source that was one diameter away from the participant. Under natural mediation, this scaling created a non-veridical percept as the attenuation within the VAE is much greater than that associated with a 1.5 m displacement. The radius of the acceptance zone for locating a source was approximately 7.5% of the radius of the circle and was chosen to be roughly within the size of a participant's quarter step under natural mediation.

2.4. Subjects

Eighteen paid volunteers participated in the experiment. Half began with training and testing using the natural mediation while the other half trained and tested on avatar mediation first. Training and testing under both modalities took approximately 75 minutes. Each participant completed the experiment in one session.

3. RESULTS

3.1. Training

Training data was obtained from all subjects before each interface was used for testing. During the training period, subjects were presented with a single source and asked to either physically move to (in the natural mediation), or position their mouse (in the avatar mediation) at the location of the source. The search path and amount of time taken for a subject to "find" the source were measured A minimum of ten trials were presented. When ten trials were completed, results were analyzed. If the standard deviation of the last five contiguous trials was less than 2.5 seconds, it was said that the subject had reached optimal performance. If optimal performance was not reached, training was continued until optimal performance was reached.

Figure 1 and Figure 2 contains an example of search times for a subject who began training under the avatar mediation condition. The white bar represents the trial at which optimal performance was reached. The results for this subject show evidence of substantial learning before reaching optimal performance under avatar mediation, but little improvement in performance over time under natural mediation. For most subjects, the natural mediation (regardless of whether it was the first or second training condition) did not exhibit the type of substantial learning demonstrated when training under avatar mediation.

The scatter plot in Figure 3 represents the mean search times, once optimality is reached, for avatar-mediation first (x-marker) and natural-mediation first (circles). The x-axis shows the avatar search times, the y-axis represents the natural mediation search times. In general, there is considerable scatter in performance across subjects with some trend towards fast and slow subjects being such under both forms of mediation. Order of training does not appear to have an effect: prior exposure to natural mediation (or avatar mediation) does not appear to help nor hinder search times achieved under subsequent training. Finally, when averaged over all subjects, there is no significant difference between avatar-mediated and natural-mediated search times.

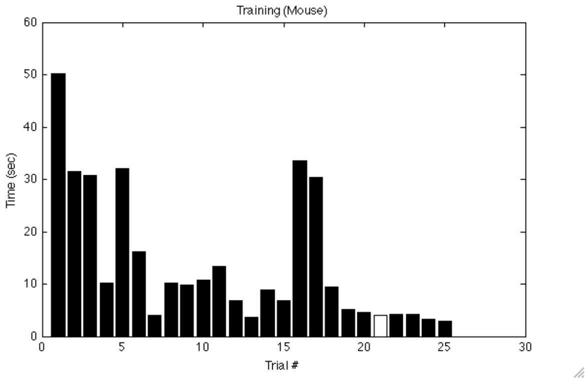


Figure 1 Example of training time results for a subject using the avatar mediation. The subject was presented with the Natural mediation first. The trial marked in white represents when subject has reached optimal performance.

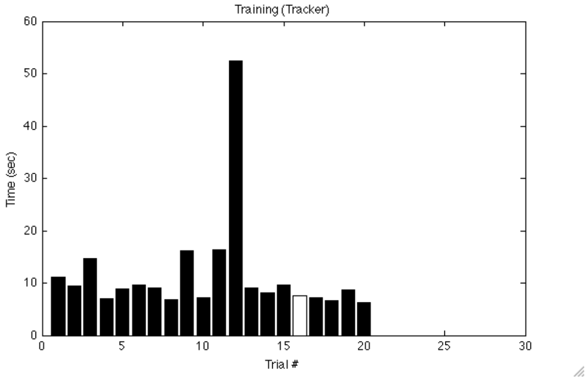


Figure 2 Example of training time results for the same subject as in Figure 1, using the Natural mediation. The trial marked in white represents when subject has reached optimal performance.

Results from the number of trials it took to reach optimality are shown in Figure 4. The scatter plot shows the number of trials for avatar mediation along the x-axis and that for natural mediation along the y-axis. As above, results are shown by the x-markers for those first trained under avatar-mediation, while circles indicated results for those first trained under natural-mediation. Out of the 18 total subjects, 5 showed virtually equal learning times with both interfaces; 7 (4 who used the natural mediation first, 3 avatar first) reached their optimal performance faster using the avatar; and 6 (3 avatar first, 3 natural mediation first) reached optimal performance faster using the natural mediation.

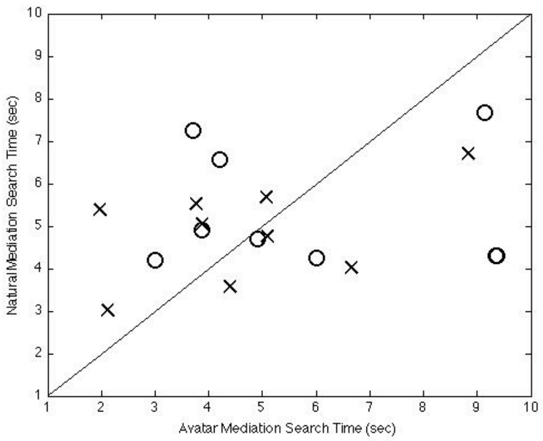


Figure 3 Mean search times are shown for the avatar-mediation first (x-marker) and natural mediation-first (circles) during the training phase.

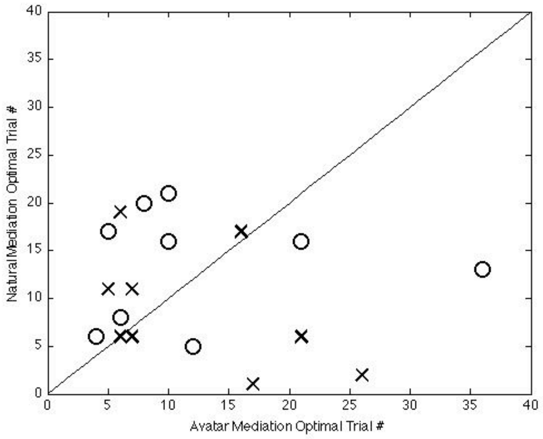


Figure 4 Optimal trial number for the avatar-first (x-marker) and natural mediation-first (circles).

Training results show that the mean search times for all subjects for the natural and avatar mediations are comparable: 5.29 sec for the avatar, and 5.11 sec for the natural mediation. These results suggest that, in the configuration used in this experiment, the type of interface does not play a role on the resulting search time.

The search time does not significantly decrease from the first to the second interface, nor does the number of trials to reach optimal performance decrease from the first to second interface. Based on these two facts, there does not appear to be any transference of performance from one interface to the other.

3.2. Test Results

Results were analyzed separately for (first) target search in the single- and four-source environments. It is beyond the scope of this paper to analyze search times and strategies for all sources in the four-source environment.

When comparing the search time results between the training session and the test trials in the single-source environment, we see similar performance between the two phases of the experiment. Figure 5 compares training search times (x-markers) and the test search times (open circles) for the two interfaces. In most cases, very similar results can be seen during the test trials, as when optimal performance is reached during training. In other words, it appears that once a subject reached a certain level of performance during the training phase, they managed to maintain this level after a break.

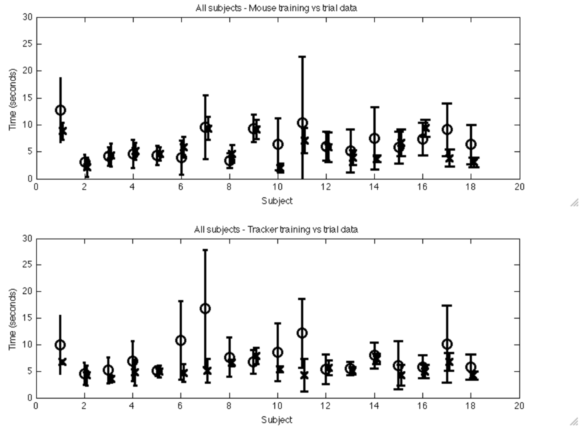


Figure 5 Error bar plot comparing results of training data (x-markers) to test data (open circles) for the avatar (upper) and natural (lower) mediation.

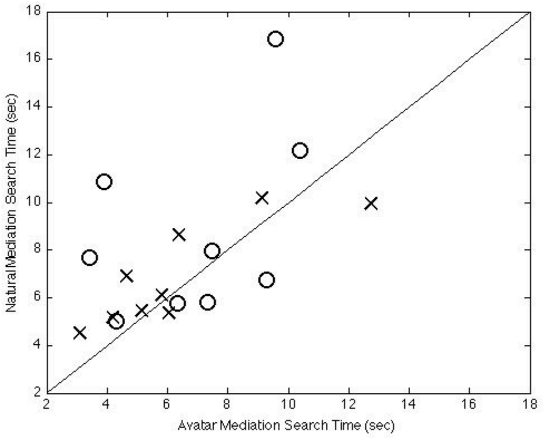


Figure 6 Single-source environment search times for avatar mediation-first (x-marker) and natural mediation-first (circles).

The mean search times during the testing phase for the single-source environment are presented in the scatter plot in Figure 6, for the avatar (x-markers) and natural (open circles) mediation. For many subjects (almost 50%) the search times for both types of mediation for each subject are very similar. A subject tended to spend an equal amount of time finding a single source regardless of whether they used avatar or natural mediation. This is consistent with our results from the training sessions, where we saw a similar search time for both types of mediation. However, when looking at the raw data for all subjects for the 1-source context, we see an overall increase in search time going from the avatar to the natural mediation. When looking at all subjects, the mean search time for the avatar is 6.2 sec, and 7.8 sec for the natural mediation. Subjects who were presented with the avatar first, show a mean response time of 6.4 sec with the avatar mediation, and 6.9 sec with the natural mediation. Subjects who were presented with the natural mediation first have a response time of 6.9 sec using the avatar, and 8.8 sec using the tracker.

Search times for the first source in a 4-source environment are similar to those for the single-source environment. The mean search time for all subjects increases from 6 sec, using the avatar mediation, to 7.6 sec with the natural mediation. This is confirmed with subjects who were presented with the avatar mediation first, where the mean time is 5.9 sec with the avatar mediation, and 7.2 sec with the natural mediation. Subjects who were presented with the natural mediation first also exhibit a similar increase from 6.2 sec with the avatar mediation to 8 sec with the natural mediation. This overall increase in search times from the avatar to the natural mediation can be seen in the scatter plot in Figure 7.

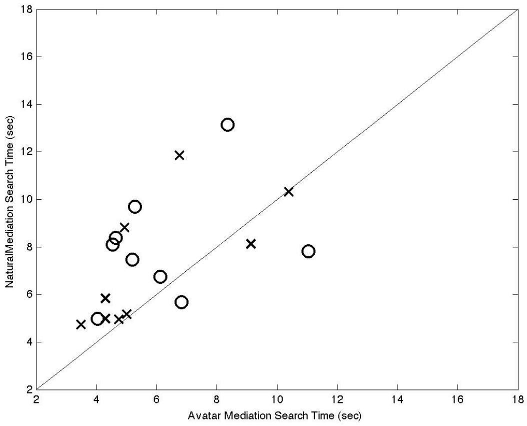


Figure 7 Search times for the first source within the 4-source context for avatar mediation-first (x-marker) and natural mediation-first (circles).

3.3. Search strategies

To further evaluate the subjects' performance using each form of mediation, we analyzed the paths taken by each subject when finding the source. These paths are indicators of the search strategies used by the subject, and give us insight into how well the user's spatial knowledge of the interface is being utilized. In his study of navigation behavior, Tellevik [5] found that a listener's search strategy changes over time as a result of learning. Many virtual environment and spatial cognition researchers (Buechner et. al [1], Hill et. al [3], and Thinus-Blanc & Gaunet [7],) have classified spatial search patterns into those that indicate novice search performance and those which indicate a more experienced search technique. It is by these classification schemes that we have categorized each subject's path data. Figure 8 shows an example from our data of a path that would be classified as a novice (left) strategy and a path that would be classified as an experienced (right) strategy.

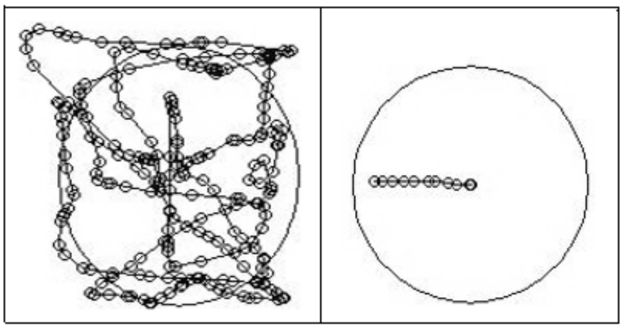


Figure 8 Classification of search strategies. The path on the left is classified as a *directed random* strategy and the path on the right is classified as an *enfilading* strategy.

Figure 9 shows the frequency of usage across subjects of an experienced search strategy during training. Subjects using natural mediation who trained first under avatar mediation exhibited the highest proportion of experienced search strategies. This trend can also be seen in the frequency of usage results during the test phase of the experiment as shown in Figure 10. Subjects using natural mediation, who trained on the natural mediation first also exhibited a high proportion of usage of experienced search strategies, although slightly lower than that of the subjects who trained first under avatar mediation. For the single-source environment, subjects who trained on the natural mediation first, when moving to the avatar mediation, showed a decline in performance. The subjects in this condition began the test, using a high proportion of sophisticated search strategies and later ended the test, using the least proportion of experienced search strategies.

Figure 11 examines the usage of an experienced search strategy to find a single source in the 4-source environment. Here, we can see that performance is very similar under natural and avatar mediation: there is no performance difference in the usage of sophisticated search strategies for either form of mediation.

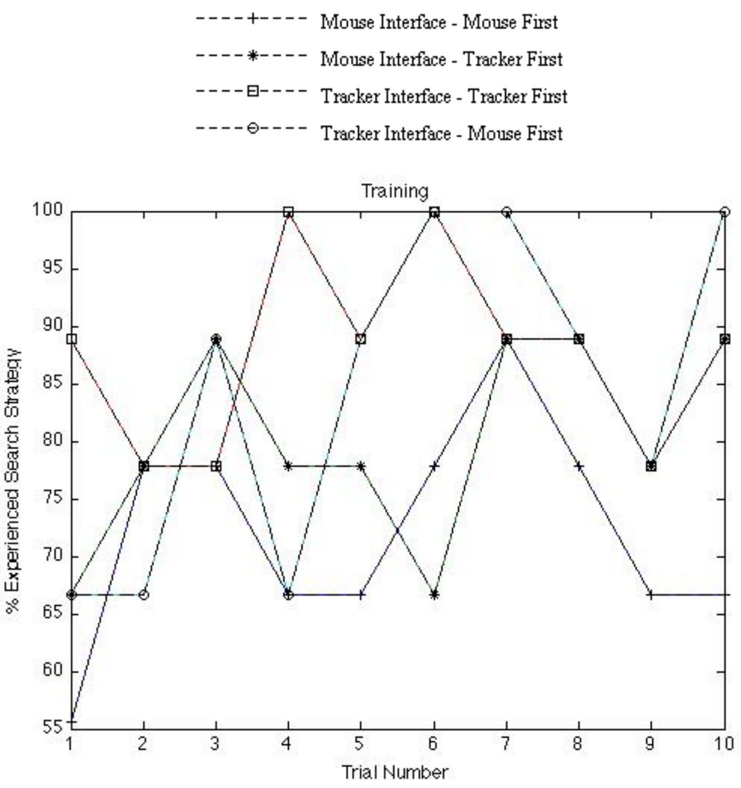


Figure 9 Usage of an experienced search strategy while training

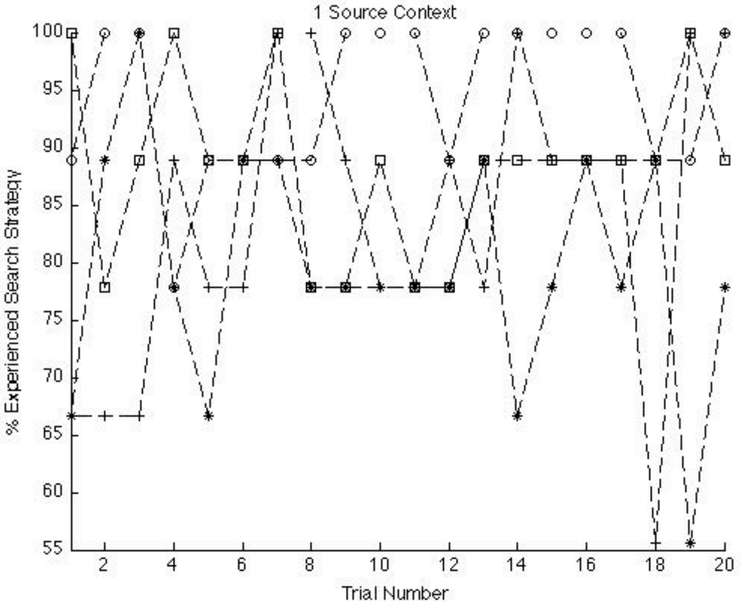


Figure 10 Usage of experienced search strategy in 1 source context environment

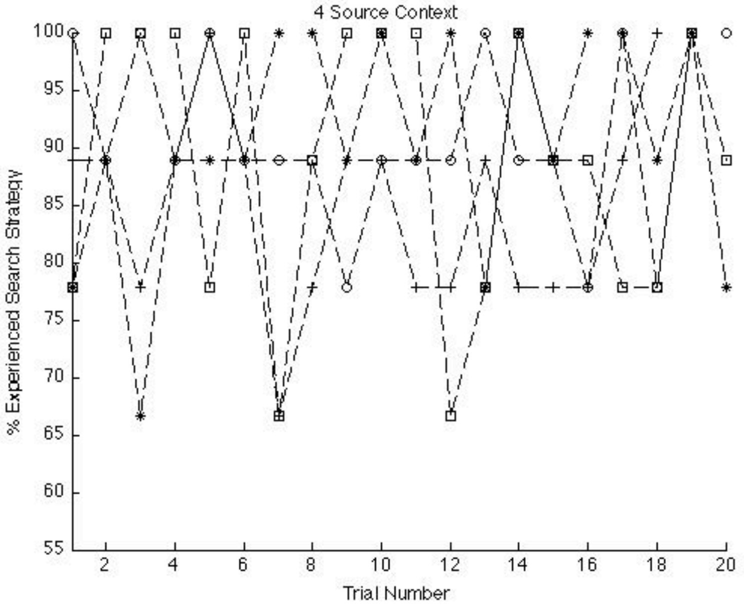


Figure 11 Usage of experienced search strategy in 4-source context environment

4. DISCUSSION AND CONCLUSIONS

This paper presents results of an experiment that compares search times and strategies of an avatar and natural mediation interface when finding one source within the context of a single and four-source auditory environment. Results from the avatar and natural mediation training and test phases of the experiment suggest that, although the training trends for the two interfaces are different, the resulting search times are similar.

Results from the training phase, during which subjects familiarized themselves with each interface and the task, show that in many cases, the number of trials necessary to reach optimal performance with each interface was similar. However, looking closely at the trial data we notice that there is clear evidence of learning to use an avatar to interact with the acoustic environment. . Such steep learning curves were not seen in most subjects under natural mediation. These results are independent of whether the avatar mediation was presented to the subject as the first or the second interface and suggest no transfer of experience across the two interfaces.

The asymptotic search times achieved during training were very similar in both interfaces. During the testing phase we saw an increase in the search times in both the single and multi source context conditions (from 5.29 sec to 6.2 sec with the avatar, and from 5.11 sec to 7.8 sec). Although the training and test search times in most subjects were very similar, in a few subjects we observed an increase in search times in the testing phase of the experiment, which could be due to fatigue. Further testing is needed to validate the cause of the search time increase.

Regardless of the number of sources in the context (one or four sources), results show nearly identical search times. This suggests that the number of sources in the background does not create a distraction for the subject, at least for finding the first source. Further analysis is needed to describe search times for the remaining sources in a multi source context.

Congruent with the search time data, the search strategy data also indicate that there is no clear difference in the quality of a user's search strategy under natural mediation compared to avatar mediation for finding a single source in a four source-environment. Small differences exist in the proportion of experienced search strategies used, while training as well as in the one source environment. Although these differences can be teased out, they are not significant enough to suggest that one form of mediation is significantly superior to another.

The experiment was setup in a room where the physical configuration and the limitations of the sensitivity and range of the tracker used in the natural mediation limited the physical space during testing to a radius of 1.5 meters. Although we have not performed any testing in different sized configurations, we speculate that the size of the effective area during the testing was one of the contributing factors to the similar time scales of the results. Had the area been much larger, the physical constraints of human movement would have most likely produced different results, as it is doubtful, for example, that a virtual acoustic source placed somewhere in a football field would be found by most players in under 10 seconds! The key finding of our experiment is that the only penalty in using an avatar to explore one's acoustic environment is that of learning to use the interface in the first place. Once learned, participants appear to use it as effectively as they would their own bodies in exploring a new acoustic space.

5. ACKNOWLEDGMENT

This work was supported by the Office of Naval Research Award Number N0001409WR20103, 2006-10

6. REFERENCES

[1] Buechner S. J., Hölscher, C. & Wiener, J., (2009) “Search Strategies and their Success in a Virtual Maze”. Proceedings of the 31th Annual Conference of the Cognitive Science Society, 1066-1071

[2] Cheng, C. and Wakefield, G. H. (2001). "Moving Sound Source Synthesis for Binaural Electro-acoustic Music Using Interpolated Head-Related Transfer Functions (HRTF's)," Computer Music Journal, 25(4), 57-80.

[3] Hill E.W, Rieser J.J., Hill M.M., Halpin J., (1993) “How persons with visual impairments explore novel spaces: strategies of good and poor performers.” Journal of Visual Impairment and Blindness*,* 87(8)

[4] Sandberg S., Hakansson C., Elmqvist N., Tsigas P., and Chen F.. (2006) “Using 3D audio guidance to locate indoor static objects”. Human Factors and Ergonomics Society Annual Meeting Proceedings, 50(4), 1581-1584.

[5] Shoval, S., Borenstein, J. and Koren, Y. (1998) “Auditory guidance with the Navbelt - a computerized travel aid for the blind”, IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, August, 28(3), 459-466.

[6] Tellevik, J.M., (1992) “Influence of spatial exploration patterns on cognitive mapping by blindfolded sighted persons”. Journal of Visual Impairment & Blindness, 92, 221-224.

[7] The BBC Sounds Effects Library. Princeton, N.J.: Films for the Humanities & Sciences vol. 1-40 (1991)

[8] Thinus-Blanc, C. & Gaunet, F., (1997) “Representation of space in blind persons: Vision as a spatial sense?”. Psychological Bulletin, 121, 20-42.

[9] vOICe: http://www.seeingwithsound.com/

AURALLY AIDED VISUAL SEARCH WITH MULTIPLE AUDIO CUES

Brian Simpson and Nandini Iyer

Air Force Research Laboratory
Wright-Patterson Air Force Base, OH
brian.simpson@wpafb.af.mil
nandini.iyer@wpafb.af.mil

Douglas S. Brungart

Army Audiology and Speech Center
Walter Reed Army Medical Center
Washington, DC
douglas.brungart@us.army.mil

ABSTRACT

In many applications, the primary goal of a spatialized audio cue is to direct the user's attention to the location of a visual object of interest in the environment. This type of auditory cueing is known to be very effective in environments that contain only a single visual target. However, little is known about the effectiveness of this technique in environments with more than one possible target location. In this experiment, participants were asked to identify the characteristics of a visual target presented in a field of visual distracters. In some conditions, a single auditory cue was provided. In other conditions, the auditory cue was accompanied by one or more audio distracters at different spatial locations. These conditions were compared to a control condition in which no audio cue was provided. The results show that listeners can extract spatial information from up to three simultaneous sound sources, but that their visual search performance is significantly degraded when more than four simultaneous sounds are present in the stimulus.

1. INTRODUCTION

In many practical applications of virtual audio displays, the primary purpose of the spatialized auditory cue is to direct the user's attention to the location of a target object so a positive visual identification can be made. In situations where the visual field is cluttered and the target object is difficult to distinguish from other visual objects in the environment, dramatic reductions in visual search time can be achieved simply by turning on a broadband sound at the location of the target. For example, Bolia et al. [1] examined the benefit of audio cueing as a function of visual scene complexity by manipulating the number of visual objects in the scene (i.e., the set size). In a two-alternative, forced-choice task, the subjects were to detect and identify which of two target light arrays was presented on each trial. They found that, when no audio cue was presented, visual search times increased with increasing set size, consistent with a limited-capacity attentional process in which an observer must serially scrutinize each display element individually. However, when an audio cue was presented from the location of the target, response times were significantly reduced relative to the no-cue condition (by up to 93%), and were essentially independent of set size, suggesting that the benefit of providing an auditory cue that is spatially coincident with a visual target not only reduces target acquisition times dramatically, but in fact changes the nature of the search strategy. Specifically, the salience of the auditory cue leads to searches that are more characteristic of parallel search processes, and thus are essentially independent of set size.

From these results, it is evident that a continuous spatialized audio cue at the location of the target is almost always an advantageous display strategy in cases where the listener's visual attention should be directed to a single known location in space. However, the situation gets more complicated in cases where it is necessary to cue more than one target location at the same time. This might occur because the range of possible target locations has been narrowed down to one of N possible locations, or it might occur because more than one simultaneous target exists and the relative priority of each target cannot be determined without visual inspection by the operator. In either case, care must be taken in determining how to provide spatial auditory cues at the location of more than one simultaneous target. The simplest strategy is to turn on a different independent continuous sound source at each potential target location. However, each additional simultaneous sound source will reduce the localizabilty of the individual sources in the mixture [2], and as a result one would expect the advantage of audio cueing to decrease as the number of cued target locations increases. In the limit, one would expect performance to deteriorate to the point where no measurable advantage in search time is observed from the addition of spatialized audio cues at the locations of the potential targets.

In this experiment, the aurally-aided visual search paradigm employed by Bolia et al. [1] has been adapted to examine how visual search times change as a function of the number of auditorally-cued *potential* target locations within a set of 50 visual distracters. The next section describes the experimental procedures in more detail.

2. METHODS

2.1. Apparatus

The experiments were conducted in the Auditory Localization Facility (ALF) at Wright-Patterson AFB in Dayton, Ohio (Figure 1). The ALF is a geodesic sphere 4.3 m in diameter that is equipped with 277 full-range loudspeakers spaced roughly every 15° along its inside surface. The ALF facility is connected to a high-powered signal switching system that allows up to 16 different sounds to be routed to any or all loudspeakers from a multichannel digital soundcard (RME).

Mounted in front of each loudspeaker in the ALF facility is a small visual display consisting of a cluster of four red LEDs arranged in a square pattern, with each diode subtending a visual angle of approximately 0.5°. Figure 2 shows an illustration of the possible modes of these LEDs.

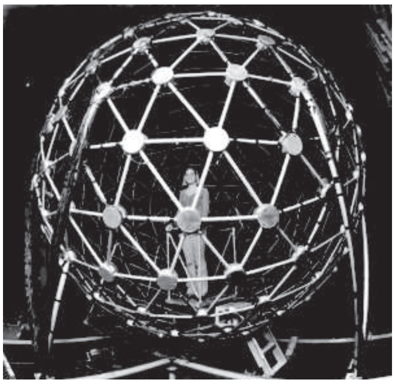


Figure 1: Auditory Localization Facility used for HRTF collection

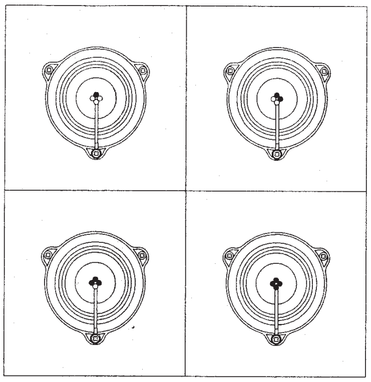


Figure 2: Configurations of LEDs used for visual display in experiment. LED clusters with an odd number of active LEDs (1 or 3) were used as visual distracters (left column). An LED cluster with an even number of active LEDs (2 or 4) was used to designate the target location. Participants were required to search all the loudspeaker locations to find the location with an even number of active LEDs, and then press a button to identify whether the target had 2 or 4 LEDs active.

2.2. Participants

A total of 8 paid volunteer listeners participated in the experiment, including 4 males and 4 females. All had normal audiometric thesholds, and their ages ranged from 19 to 25 (Mean age 23 years). All were screened to have uncorrected 20/20 vision in both eyes.

2.3. Procedure

The experiment was conducted with participants standing on a platform in the center of the ALF facility. The participants wore a headband with a 6-DOF headtracking sensor (IS-900) attached, and, at the start of each trial, they were asked to turn and face the front speaker in the ALF facility until an LED cursor slaved to the participants' head orientation was activated at that loudspeaker. They then pressed a button to indicate their readiness to begin the trial. At that point, two things happened. First, a visual display was generated by randomly selecting one loudspeaker as the target loudspeaker and turning on either two or four LEDs at that loudspeaker location, and then randomly selecting 8, 16, or 50 other loudspeaker locations as "visual distracters" and turning on 1 or 3 LEDs at each of those loudspeaker locations (see Figure 2). Second, a broadband continuous noise signal was switched on at the location of the target, and additional, statistically-independent random noise signals were simultaneously switched on at 0, 1, 2, 3, 5, 7, or 15 other audio distracter locations. These audio distracter locations were chosen randomly from among the locations of the visual distracters. Thus, in the condition with 50 visual distracters and 8 audio distracters, the 277 speakers in the ALF facility included: one *target* speaker with a continuous noise signal and either 2 or 4 active LEDs; seven *audio distracter* loudspeakers with a continuous noise signal and either 1 or 3 active LEDs; and 43 *visual distracter* loudspeakers with no sound but either 1 or 3 active LEDs.

In all cases, the participant's task was the same: search all the loudspeaker locations with active sound sources for the target location with an even number of active LEDs (2 or 4), and press a response button to indicate whether there were 2 or 4 LEDs active at the target location.

Responses were collected in blocks of 20 trials. On each trial, the number of audio distracters and visual distracters was randomly chosen. Most of the data were collected in conditions with 50 visual distracters. Over the course of the experiment, each of the eight participants provided responses in 60 trials in conditions with 50 visual distracters and 0, 1, 2, 3, 4, 5, 7 or 15 auditory distracters. They also participated in three visual-only control conditions with 8, 16, or 50 visual distracters but no auditory signals. In total, a minimum of 700 trials were collected on each of the eight participants in the experiment.

3. RESULTS

Listeners were instructed to conduct the task as quickly as possible while ensuring a very high level of accuracy on the identification of the number of LEDs at the target location. As a result, overall accuracy on the LED identification task was extremely high- listeners correctly distinguished between target configurations containing 2 or 4 LEDs in 99.82% of all trials.

The more meaningful metric of performance in the task is response time, measured from the presentation of the stimulus at the beginning of the trial to the time when the participant pressed the response button identifying the target, which terminated the trial. Figure 3 shows performance as a function of the number of visual distracters in the visual-only control condition, where no audio stimulus was presented, averaged across all participants. As would be expected, the amount of time required to complete the task increased systematically as the number of visual distracters increased, suggesting a serial search process. When 50 visual distracters were present, response time was on average about 8 seconds.

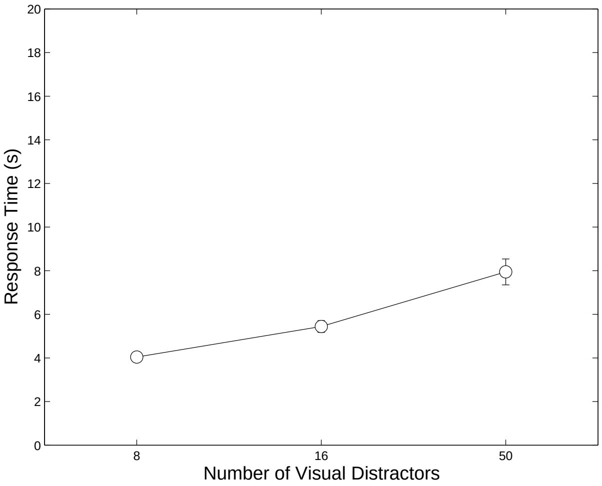


Figure 3: Response times, averaged across all participants, plotted as a function of the number of visual distracters in the visual-only control condition. The error bars show the 95% confidence intervals around each data point.

Figure 4 shows mean response time in the experiment as a function of the number of audio distracters in the conditions with 50 visual distracters. For comparison purposes, the shaded bar in the middle of the figure shows performance in the visual-only condition with 50 visual distracters and no audio stimuli. Again, as expected, the overall visual search time was found to increase with the number of audio distracters. Moreover, as expected, the benefit of having an audio signal at the location of the target disappears after the addition of a certain number of audio distracters. Specifically, there is no longer a difference between the visual-only condition and the audio condition when three audio distracters are added to the stimulus.

What is somewhat surprising about the data, however, is that performance does not merely plateau when enough audio distracters are added to the stimulus to eliminate any useful information the listener might obtain from the audio cue at the location of the target. Rather, it continues to worsen, and when 15 audio distracters were present, the total search time to find the target was almost twice as long as it was when no audio signals were presented at all. Importantly, this result suggests that listeners are not generally able to determine when audio information no longer provides any advantage in this visual search task. In such cases, one might expect that the participants would adjust their strategy and ignore this distracting audio information. Rather, the fact that response times continue to increase with the number of sounds suggests that participants are searching serially through the sounds in order to locate that sound associated with the visual target. This means that audio display designers must use extreme caution when they implement audio displays that have the potential to generate large numbers of spatialized cues at the same time. The results of this experiment suggest that the users of these systems may not be able to accurately determine when the audio information should be relied upon for a visual search task, and when it should be ignored. Consequently, it seems that there may be some cases where the provision of additional audio information might actually significantly degrade the operator's performance in complex visual search tasks.

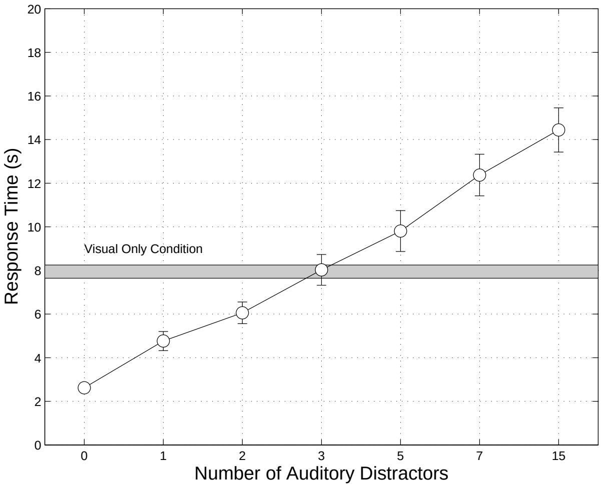


Figure 4: Response times, averaged across all participants, plotted as a function of the number of audio distracters in the conditions with 50 visual distracters. The error bars show the 95% confidence intervals around each data point. The shaded bar shows mean performance in the visual-only condition with 50 visual distracters.

4. CONCLUSIONS

In this experiment, we examined how well participants were able to perform a complex aurally-aided visual search task when one or more distracting sounds were presented concurrently with the audio cue from the location of the visual target. The experiment was intended to replicate the kind of scenario that might occur when an operator is required to investigate more than one simultaneous visual target, or when a visual target or threat is known to be present at one of a small number of possible locations. In cases where there are fewer than four simultaneous targets, these results suggest that some advantage can be gained simply by providing a co-located continuous sound source at all the possible locations in the target set. However, when more than four target locations need to be cued, the presentation of simultaneous co-located audio cues at the target locations actually results in a significant *degradation* in performance relative to the visual-only case where no cueing sounds are provided.

However, it is important to note that these results only apply to the worst-case condition where the exact same audio cue is provided at all the possible locations in the target set. While there is no guarantee that performance would be improved by other types of cueing sounds, it is likely that some alternative audio symbology incorporating sounds that do not overlap either in time or frequency might be able to produce better performance in this task than was obtained with the continuous broadband noises used in this study. We are currently conducting experiments to explore this possibility in more detail.

5. REFERENCES

[1] R. Bolia, W. D'Angelo, and R. McKinley, "Aurally aided visual search in three dimensional space," *Human Factors*, vol. 41, pp. 662–669, 1999.

[2] D. Brungart and B. Simpson, "Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty," *Journal of the Acoustical Society of America*, vol. 115, pp. 301–310, 2004.

INFORMATION-THEORETIC PERFORMANCE ANALYSIS OF AZIMUTHAL LOCALIZATION FOR SPATIAL-AUDITORY DISPLAY OF BEAMFORMED SONAR DATA

Jason E. Summers

Applied Research in Acoustics (ARiA) LLC
Washington, D.C.
jason.e.summers@ariacoustics.com

ABSTRACT

An information-theoretic model of azimuthal localization is presented. The number of distinct source locations that can be encoded by a set of head-related impulse response functions (HRIR) is predicted in terms of information transfer as a function of the properties of the source signal and a quantization interval that is related to the level of internal perceptual noise. The model also predicts how source locations should be distributed in azimuth in order to maximize the information transferred through the set of HRIR for a given set of input conditions. The predictions are related to design considerations for a spatial-auditory display of beamformed sonar data in which time series associated with fixed beams of a one-dimensional array are mapped to virtual sources located at fixed radius from the listener in the horizontal plane.

1. INTRODUCTION

A recent review by Arrabito et al. cites, “the ability to present sonar beams in a three-dimensional auditory display where the spatial position of each sonar beam corresponds to the actual position of the source in the ocean,” as a key research area for enhancing the role of the auditory modality in processing of sonar data [1]. While such an auditory equivalent of a low-level geographic situation (GEOSIT) display is ultimately limited by the beamforming algorithm and the physical receiver array, the number and spacing of the virtual-source locations to which beams are mapped should be governed by the spatial resolution of human hearing. Here, a model is developed that predicts and bounds human azimuthal localization performance based on the amount of spatial information encoded in a set of head-related impulse responses (HRIR). A well-defined signal-processing algorithm is described which determines the information content of the HRIR as a function of several variables, including internal perceptual noise and external source-signal spectrum.

Though direct methods for binaural presentation of sonar signals have been used since the earliest days of sonar [2, 3], spatial auditory display of sonar beam data in which beams are mapped to virtual sources was first mentioned in the open literature much more recently by McFadden and Taylor [4] and has been a topic of ongoing interest [5, 6, 7, 1] .

A simple and obvious approach for presentation of sonar beams from a one-dimensional array via a spatial-auditory display is mapping of the time-series associated with each beam to a virtual source located at a particular azimuth on a circle in the horizontal plane. Such a scheme raises two primary technical concerns. First, it must be known how many independent (virtual) source locations can be identified, which determines the number of beams that can be mapped. Second, it must be known how the virtual sources associated with the beams should be distributed in azimuth in order to realize the desired level of performance. Two theoretical questions underlie these technical concerns: How much azimuthal information about source location can a set of HRIR encode and how much of this information can a listener extract? Further, how is the information density of a set of HRIR allocated in azimuth?

To address these questions, spatial hearing in the horizontal plane is here recast as a communication problem in which scattering from the head and torso, described by the set of HRIR, encodes source location. This representation of the problem provides an information-theoretic framework for the analysis of localization performance. By postulating coding and decoding in terms of the coefficients of a particular orthogonal decomposition of the set of HRIR, the amount of information transferred and, consequently, localization performance is mathematically determined as a function of the particular set of HRIR, the source-signal, and, through a resulting quantization interval, the perceptual signal-to-noise ratio (SNR). For each set of input parameters the model yields a map of information density as a function of azimuth, which indicates the locations of virtual sources required to realize the maximum information transfer possible for a given set of conditions.

2. INFORMATION-THEORETIC MODEL

The task of identifying the direction of a sound source from the received binaural signal is similar to the problem of localizing a radiator in a sound channel from measurements of the sound field in the channel (i.e., matched-field processing, see, e.g., [8]). In both cases spatially dependent variations in the impulse response (of the channel or the scattering from the head and torso) encode information about source position and one estimates the location of the source from measurements of the sound field at the receiver. By recasting this problem as a gridded search in which the task is identifying which cell of the grid contains the source, Buck et al. [9, 10, 11] formulated source localization as an unconventional communication problem and developed an information-theoretic framework for characterizing localization performance. Though Buck et al. investigated the standard matched-field problem of a single-frequency continuous source with the measurement of the sound field being the vector of complex pressures received on a vertical array of hydrophones, Gaumond later used a con-

This work was supported in part by the Office of Naval Research and was performed in part while the author held a National Research Council Research Associateship Award at The U.S. Naval Research Laboratory.

structive approach within this information-theoretic framework to characterize localization performance for a band-limited impulsive source with the measurement of the sound field being the pressure time series received on a single hydrophone [12]. It is this latter approach that is followed here.

In the information-theoretic representation, the identity of the cell containing the source is the message that is encoded into the sound field and one estimates the identity of the cell containing the source based on noisy measurements of the sound field at the receiver. Note that this formulation, though isomorphic to more conventional communication problems, is fundamentally unlike them in interpretation. Whereas the signal transmitted by the source contains the encoded message in conventional communication problems, in this problem the source location itself is the message.

To characterize azimuthal localization performance, assume that there is a source located on a circle in the horizontal plane with unknown azimuth Θ described by the probability density function $p_\Theta(\theta)$. Following Buck, the continuous set of input conditions is discretized according to $\beta(\theta)$ described by the probability mass function $p_\beta(m)$ where $m = 1, 2, \ldots, M-1, M$.

The time series of acoustic pressure at each ear resulting from a source located at $\Theta = \theta$ are given by

$$x_{left}(\theta; t) = s(t) * h_{left}(\theta; t), \tag{1a}$$

$$x_{right}(\theta; t) = s(t) * h_{right}(\theta; t), \tag{1b}$$

where $s(t)$ is the source waveform and $h_{left}(\theta; t)$ and $h_{right}(\theta; t)$ are the HRIR associated with azimuth θ and the left and right ears, respectively. These can be represented as a single time series by concatenating the responses from the left and right ears in a single time series

$$x(\theta; t) = [\ x_{left}(\theta, t)\ \ x_{right}(\theta, t)\ ]. \tag{2}$$

This received signal is corrupted by internal processes that produce perceptual and criterial noise [13, p. 458], $n(t)$, which can be modeled as additive white Gaussian noise (AWGN). The resulting time series is given by

$$y(\theta; t) = x(\theta; t) + n(t). \tag{3}$$

While the prior equations are expressed in terms of the continuous azimuthal variable Θ, they are equivalent for the discretized case when expressed in terms of the discrete azimuthal index β.

In the discrete case, the complete set of HRIR can be represented as an M-by-N matrix $\mathbf{X}$, where each row of $\mathbf{X}$ is the discrete-time signal of length N given by (2). As in [12] the matrix is constructed so as to exclude initial time-delay because it does not encode any information about source location. However, it is constructed to preserve all aspects of $x(\beta; t)$ that do encode information about source location—interaural time difference (ITD), interaural level difference (ILD), spectral variation and other cues such as those described in [14]. This is accomplished by sequentially time shifting each $x(m; t)$ in order to maximize the cross correlation between the high-energy peaks at the ipsilateral ear with those of the adjacent $x(m-1; t)$. Because each $x(\beta; t)$ is shifted as a whole, ITD is preserved.

In this signal-processing approach, the goal is estimation of the source position as a function of the corrupted signal given by (3)

$$\hat{\beta} = g(y(\beta, t)). \tag{4}$$

In contrast to localization-performance metrics that measure the mean-square error between the estimated and true source position, such as the Cramer-Rao lower bound (CRLB) [8, 15], the information-theoretic framework characterizes performance by the probability of error P_e in assigning the source to a discrete cell

$$P_e = Pr\{\hat{\beta} \neq \beta\}. \tag{5}$$

To characterize the information content of the set of HRIR in terms of a set of discrete states, the time series associated with each discrete source location given by β is expanded in a set of empirical orthogonal functions (EOF) $\nu_n(t)$

$$x(\beta = m; t) = x_m(t) = \sum_{n=1}^{N} \alpha_{mn} \nu_n(t), \tag{6}$$

such that the information content of the set of HRIR is described by the set of coefficients $\{\alpha_{mn}\}$. In this representation each $x_m(t)$ describes the mapping from one of the M discrete positions in azimuth described by β to a position in an N-dimensional space given by the vector of coefficients

$$\mathbf{a}_m = [\alpha_{m1} \cdots \alpha_{mN}]. \tag{7}$$

This expansion is realized by singular value decomposition of the matrix $\mathbf{X}$

$$\mathbf{X} = \mathbf{USV}^T, \tag{8}$$

where the coefficients α_{mn} are given by the product of the matrix of singular vectors and the diagonal matrix of singular values

$$\mathbf{US} = \begin{bmatrix} \alpha_{11} & & \\ & \ddots & \\ & & \alpha_{M,N} \end{bmatrix}, \tag{9}$$

and the EOF $\nu_n(t)$ are given by the matrix of singular vectors

$$\mathbf{V} = \begin{bmatrix} \nu_1(t) \\ \vdots \\ \nu_N(t) \end{bmatrix}. \tag{10}$$

As Gaumond observed [12] for underwater sound channels, coefficient vectors $\mathbf{a}_m$ are not necessarily unique; some may be degenerate. The same holds true for the coefficient vectors of the HRIR. For example, a spherical model of the head yields a set of HRIR that do not resolve source positions lying on cones of confusion. However, HRIR from a realistic head-and-torso model are able to encode substantially more information about source position (see, e.g., [16] and [17, p. 274]) such that the primary source of degeneracy is limited resolution of the encoding due to quantization. For human localization this quantization of coefficients serves as model for the effects of internal noise processes. Expanding $n(t)$ in the same set of EOF

$$n(t) = \sum_{n=1}^{N} \eta_n \nu_n(t), \tag{11}$$

yields a set of independent identically distributed (iid) coefficients η_n that are zero mean with variance σ^2. This corresponds to the uncertainty being uniform in each dimension of the N-dimensional space to which the HRIR maps source position β.

Following Gaumond [12], it is assumed that σ^2 defines the quantization interval ρ_0 through a multiplicative constant $\rho_0 = a\sigma^2$. The vector of quantized coefficients is thus given by

$$\mathbf{c}_m = \left[\frac{\alpha_{m1}}{\rho_0} \cdots \frac{\alpha_{mN}}{\rho_0}\right]^T. \tag{12}$$

As in the case of $\mathbf{a}_m$, not all $\mathbf{c}_m$ are unique.

The information input into the channel is given by the entropy of the probability mass function of the random variable describing source position

$$H_{in} = H(\beta) = -\sum_{m=1}^{M} p_\beta(m) log_2 p_\beta(m). \tag{13}$$

Assuming that all source positions are equally probable (maximum entropy) yields

$$p_\beta(m) = 1/M, \tag{14}$$

so that

$$H(\beta) = log_2 M. \tag{15}$$

This corresponds to $2^{H(\beta)}$ discernible source positions. The joint source-channel coding theorem requires that the mutual information $I(\beta;\hat{\beta})$ satisfy

$$H(\beta) \leq I(\beta;\hat{\beta}), \tag{16}$$

in order to estimate β with arbitrary small probability of error ($P_e \rightarrow 0$). Mutual information can be expressed as

$$I(\beta;\hat{\beta}) = H(\hat{\beta}) - H(\hat{\beta}|\beta), \tag{17}$$

where $H(\hat{\beta})$ is a measure of prior uncertainty about which azimuthal cell contains the source and the conditional probability $H(\hat{\beta}|\beta)$ represents that uncertainty remaining about $\hat{\beta}$ once the source transmits from cell $\beta = m$. When all $\mathbf{c}_m$ are unique, $H(\hat{\beta}|\beta) = 0$ and $H(\mathbf{c}) = H(\beta)$. Thus, it is required that $H(\hat{\beta}) = H(\beta)$ for ($P_e \rightarrow 0$). As in [12] the constructive approach postulates that the decoding from HRIR to $\hat{\beta}$ is done in terms of the coefficients of a particular orthogonal decomposition. This is only one possible decoding scheme, motivated more by mathematical convenience than an underlying verisimilitude to the actual human processes. Other decoding schemes such as those described in [17] that are more closely related to psychophysical localization cues are also possible. However, the decoding scheme presented here has the significant analytical advantage of producing a set of orthogonal coefficients. Therefore it is postulated that $\hat{\beta} = g(\mathbf{c})$. The data-processing theorem [18, Thm. 2.8.1, pp. 34–35], then requires that

$$I(\beta;\hat{\beta}) \leq I(\beta;\mathbf{c}), \tag{18}$$

such that for $H(\mathbf{c}) = H(\beta)$ must hold true for ($P_e \rightarrow 0$). Because $\hat{\beta} = g(\mathbf{c})$, the output information is given by the entropy of those L vectors of discrete coefficients that are unique

$$H_{out} = H(\ell) = -\sum_{\ell=1}^{L} p(\ell) log_2 p(\ell) \quad \in [0, H(\beta)], \tag{19}$$

where $p(\ell)$, $\ell = 1, 2, \ldots, L-1, L$, is the sum of the probabilities over all input states $\beta = m$ that result in the ℓth coefficient vector

$$p(\ell) = -\sum_{m|c_\ell = c_m} p_\beta(m). \tag{20}$$

If only noise is present

$$p(\mathbf{c}_m) = \left\{ \begin{array}{ll} 1 & m = 1 \\ 0 & m > 1 \end{array} \right. , \tag{21}$$

such that $H(\mathbf{c}) = 0$, indicating that no information is transferred. In contrast, if all $\mathbf{c}$ are unique, $p(\mathbf{c}) = p_\beta(m)$ such that $H(\mathbf{c}) = H(\beta)$, which is $log_2 M$ for equiprobable source positions. However, if all $\mathbf{c}$ are not unique $H(\mathbf{c}) \neq log_2 L$ because equipartition of probability over source positions does not correspond to equipartition of probability over unique coefficient vectors when some coefficient vectors are degenerate. This is made clear by the following reformulation of the information theoretic error metric given in (5)

$$P_e = \sum_{(m,\hat{m})} p_{\hat{\beta}}(\hat{m}|m) p_\beta(m) d(m,\hat{m}), \tag{22}$$

where $p_{\hat{\beta}}(\hat{m}|m)$ is the conditional probability of the source-location estimate given the distribution of source positions and $d(m,\hat{m})$ is the Hamming distortion measure

$$d(m,\hat{m}) = \left\{ \begin{array}{ll} 0 & m = \hat{m} \\ 1 & m \neq \hat{m} \end{array} \right. . \tag{23}$$

Because $p_{\hat{\beta}}(\hat{m}|m) \neq 0$ for $\hat{m} \neq m$ if all $\mathbf{c}$ are not unique, (22) indicates that $P_e > 0$ if $p(m) = 1/M$.

3. NUMERICAL RESULTS

To illustrate the theory developed in Sec. 2, consider the set of HRIR for the KEMAR dummy-head microphone [19]. The set of HRIR were measured in the horizontal plane at a fixed radius from the center of the head in 5 deg. increments of azimuth ($M = 72$). A small loudspeaker served as the source and transmitted maximum-length pseudorandom binary sequences of length 16383, which were recorded at a sampling rate of 44.1 kHz, resulting in a nominal SNR of 65 dB. The impulse response of the measurement loudspeaker was removed from the HRIR measurement using a inverse filter calculated from its measured impulse response using a Mourjopoulos least-squares technique [20], yielding a response that is approximately flat over the bandwidth of the loudspeaker.

As in [12] the source waveform is a band-limited impulse. Four different source spectra are considered, unfiltered, having the full bandwidth of the measurement system, and three octave-band-filtered impulses with center frequencies of 250Hz, 1kHz, and 4kHz. Signal-to-noise ratio is specified indirectly in terms of the maximum number of quantization levels q for any coefficient in $\mathbf{c}$. While this avoids specifying the explicit relation between SNR and quantization interval ρ_0, it means that results for different source waveforms are not directly comparable.

The analysis for each of the source waveforms is the same: $\mathbf{X}$ is expanded in a set of EOF in order to define α_{mn}. The vectors of discrete coefficients are then calculated according to (12) for each of the 72 source positions for $q = 1, 2, 3, 4$. In Tables 1– 4 the number of unique coefficients, the output information given by the entropy of the coefficients $H(\mathbf{c})$, and the corresponding number of distinct source positions that can be identified with arbitrary small probability of error are specified for each value of q, under the assumption of equiprobable source positions. Figures 1 – 4 display the azimuthal distribution of unique coefficients as a color

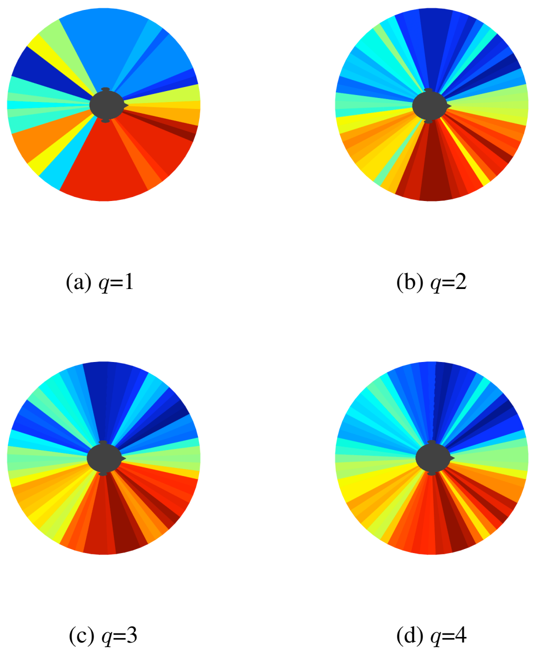

(a) q=1
(b) q=2
(c) q=3
(d) q=4

Figure 1: Azimuthal distribution of unique coefficients for the full-bandwidth impulse.

plot for each of the four q values. In these figures each distinct color corresponds to a unique coefficient vector. Each value of q corresponds to a set of discrete coefficients $\mathbf{c}_{mn}^{(q)}$ and a set of HRIR given by

$$\hat{x}_m^{(q)}(t) = \sum_{n=1}^{N} \mathbf{c}_{mn}^{(q)} \nu_n(t). \tag{24}$$

These data were sonified by sequentially convolving the set of HRIR with a band-limited noise signal in order to generate a series of auditory displays equivalent to Figs. 1 – 4. The stimulus signal was a band-limited, time-windowed noise pulse of 150 ms total duration with 10 ms raised-cosine onset and offset transitions. The noise was band-pass filtered to have the same four source spectra as the band-limited impulses. For each of the four source spectra and each of the four values of q, the stimulus signal was convolved in sequence with the HRIR associated with each of the 72 source positions, beginning with 0 deg. source position and ending at 355 deg. source position. A 150 ms silence was inserted between the signals corresponding to each source position. The resulting binaural sounds are provided for the full-bandwidth case with $q = 1$, 2, 3, and 4, the 250 Hz octave-band-filtered case with $q = 1$, 2, 3, and 4, the 1 kHz octave-band-filtered case with $q = 1$, 2, 3, and 4, and the 4 kHz octave-band-filtered case with $q = 1$, 2, 3, and 4.

For the 250 Hz octave-band-filtered impulse, the only coefficient that carries azimuthal information is essentially a measure of interaural time difference (ITD), as shown in Fig. 5a. Finer quantization simply allows for more values of ITD to be encoded. At higher frequencies, there are multiple coefficients that carry azimuthal information, some of which are symmetric about the median plane and others that are antisymmetric, as shown in Fig. 5b. As observed in [16], 2 kHz is roughly the dividing point between the low-frequency region in which temporal cues dominate and the high-frequency region in which spectral cues dominate. However, in general, the coefficients do not correspond directly to psychophysical cues such as those described in [14].

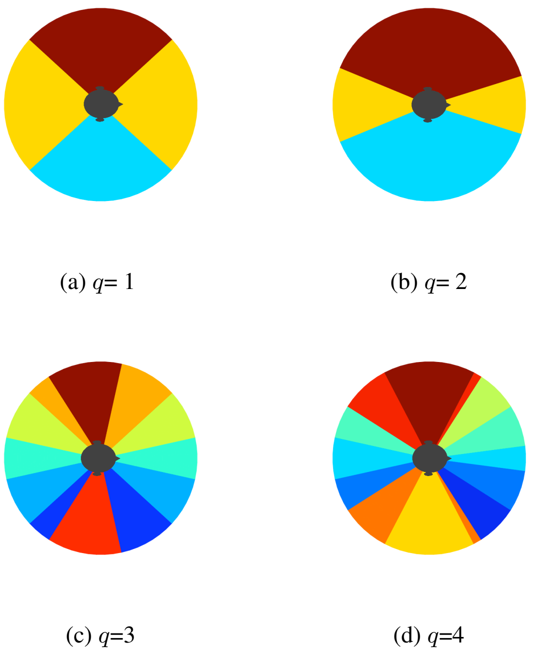

(a) q= 1
(b) q= 2
(c) q=3
(d) q=4

Figure 2: Azimuthal distribution of unique coefficients for the 250 Hz octave-band-filtered impulse.

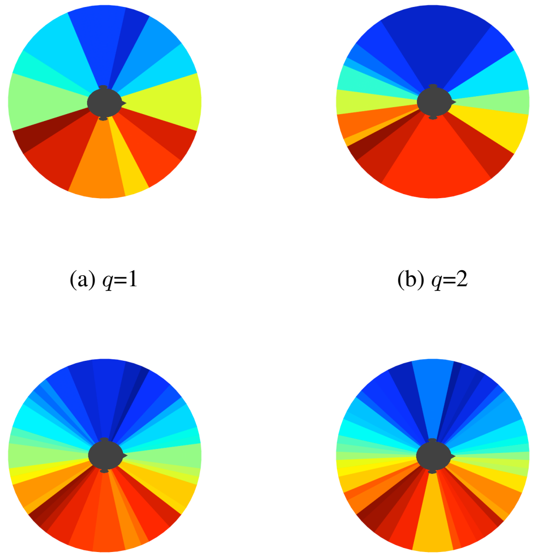
Finally, it is important to note that, because a uniform quantization interval is not optimal for the distribution of coefficient values, the increase in the number of unique coefficients with decreasing quantization interval is not strictly monotone.

4. DISCUSSION

Head-related impulse responses encode a substantial amount of information about azimuthal source location but, in general, the information density is not uniformly distributed in azimuth. Both the amount of information and the azimuthal distribution vary as a function of source signal and quantization interval. At low frequencies HRIR contain primarily ITD cues while, at higher frequencies, HRIR contain additional ILD and spectral cues. However, given a sufficiently small quantization interval, the azimuthal capacity of the HRIR exceeds that of the measured set of discrete HRIR in $\mathbf{X}$, even for the 250Hz octave band considered.

levels	unique coefficients	entropy (bits)	source positions
1	21	3.6767	12.788
2	60	5.7810	54.988
3	66	5.9823	63.219
4	72	6.1699	72

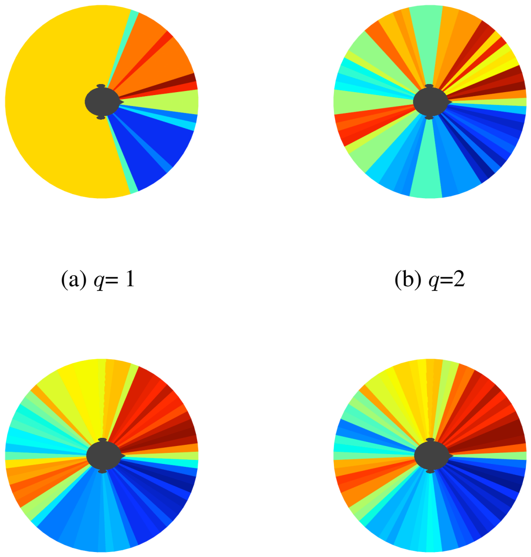
Table 1: Performance metrics for the full-bandwidth impulse



(a) q=1 (b) q=2

(c) q= 3 (d) q=4

Figure 3: Azimuthal distribution of unique coefficients for the 1 kHz octave-band-filtered impulse.



(a) q= 1 (b) q=2

(c) q=3 (d) q=4

Figure 4: Azimuthal distribution of unique coefficients for the 4 kHz octave-band-filtered impulse.

In those cases for which performance is noise (i.e., quantization-interval) limited, the number of unique coefficients is less than the number of source positions. Because the azimuthal distribution of those unique coefficients does not correspond to the distribution of source positions, the spatial information transmitted through the channel is less than its capacity. There are two possible methods for realizing the capacity of the set of HRIR to encode azimuthal information (i.e., maximize $I(\hat{\beta};\beta)$). A new partition of azimuth $\beta'(\theta)$ can be defined such that there are L discrete sources located at θ_ℓ. Alternately, the probability distribution $p_\beta(m)$ can be modified to be nonuniform so that $p(\ell)$ corresponding to the azimuths with unique coefficient vectors $\mathbf{c}_\ell$ are uniform with probability $p(\ell) = 1/L \;\; \forall \ell$. The first of these two options is related to the design of a virtual-source array for spatial-auditory display.

For example, suppose there are eight equiprobable, uniformly distributed sources but, for a given set of conditions, two source positions have coefficient vectors that are degenerate so that there are only seven unique coefficient vectors. Though the output information $H(\mathbf{c})$ is necessarily less than the input information $H(\beta) = log_2 8 = 3$, given a uniform probability distribution for each source, $H(\mathbf{c}) = 0.25 log_2 0.25 + 6(0.125 log_2 0.125) = 2.75$ is less than the maximum possible information $H(\mathbf{c}) = log_2 7 = 2.8074$ and it would not be possible to identify seven source positions with arbitrary small P_e.

As a practical example, consider the case of the full-bandwidth impulse with $q = 1$. For this level of quantization there are 21 unique coefficient vectors, as indicated in Table 1. The distribution of these coefficient vectors in azimuth, as shown in Fig. 1a, indicates that coefficient degeneracy will lead to a number of ambiguities in localization with multiple noncontiguous source locations corresponding to a single coefficient vector. This is well illustrated by the sonification of these data described previously. If, however, the number of source locations is reduced to 21 so that each of the L source positions corresponds to a single unique coefficient, the ambiguity is removed and all source positions can be resolved, as demonstrated by a sonification of the modified configuration.

Because the distribution of unique coefficients is not generally uniform in azimuth, increasing coefficient degeneracy due to decreasing SNR does not lead to a uniform loss of azimuthal resolution. Rather, information density is generally greater for frontal

levels	unique coefficients	entropy (bits)	source positions
1	3	1.5256	2.8790
2	3	1.5420	2.9119
3	7	2.7983	6.9563
4	9	3.1187	8.6860

Table 2: Performance metrics for the 250 Hz octave-band-filtered impulse

levels	unique coefficients	entropy (bits)	source positions
1	12	3.4345	10.811
2	14	3.3804	10.414
3	38	5.0441	32.994
4	40	5.1042	34.398

Table 3: Performance metrics for the 1 kHz octave-band-filtered impulse

source locations than for lateral source locations, as has been observed in psychoacoustic experiments [21]. This also can cause ambiguity in the encoding of azimuth, which leads to such phenomena as cones of confusion or the front-back confusion shown in Fig. 1a. In such cases maximum information transfer is achieved by placing only one source associated with a degenerate coefficient in one of the contiguous azimuthal sectors associated with that coefficient.

The information-theoretic framework on which the localization theory is predicated requires that all predictions be made relative to quantization interval, which is here assumed to be associated with a particular SNR. Noise in this context represents all aspects of audition that prevent perfect recovery of the spatial information encoded in the HRIR. Consequently, it cannot be measured directly and must be inferred from localization performance. To do so requires establishing a link between a model of human perception and a model of human decision. This is traditionally supplied by signal-detection theory (see, e.g., [22]).

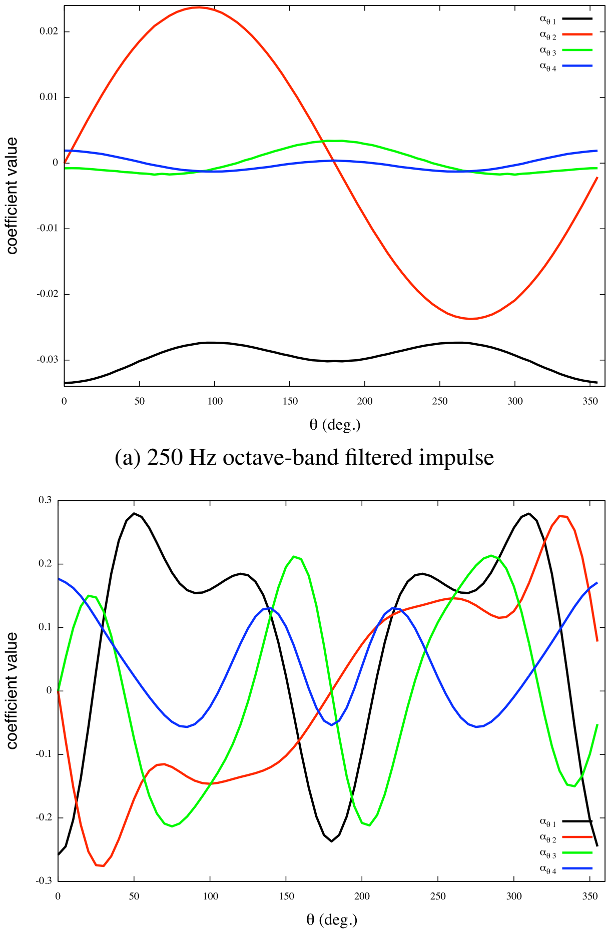
4.1. Relationship to psychoacoustic metrics of localization

While there is some recent work that applies a signal-processing method to compute performance bounds on HRIR-based localization [15], localization performance is most typically characterized through psychoacoustic metrics based on experiments with human subjects. In particular, azimuthal localization performance is characterized by absolute and relative localization thresholds. Absolute-localization experiments measure the accuracy and precision of source-location estimates. In comparison, relative-localization experiments measure acuity: the minimum audible angle of difference that can be perceived between two successive stimuli [22]. These two thresholds are linked in that the width of the distribution of absolute localization judgments (i.e., precision) is related to the relative-localization threshold [23], though in a somewhat more complex manner than suggested by a straightforward application of signal-detection theory [22], as discussed by [24].

The information-theoretic performance analysis described in this paper characterizes absolute-localization. In particular it bounds the performance of the source-identification method [25], which formulates the absolute-localization task as source identification over a grid of equal-azimuth sectors. In contrast, the Cramer-Rao lower bound [15] characterizes the mean-square error of absolute localization and thus is more closely related to relative localization. One of the primary spatial-hearing tasks that an operator could perform using an azimuthal spatial auditory display is aurally detecting and estimating the bearing of a signal associated with one of many virtual sources. For this task the relevant measure of performance is absolute localization, particularly the source-identification method.

levels	unique coefficients	entropy (bits)	source positions
1	9	1.9874	3.9652
2	42	5.1175	34.7152
3	58	5.7391	53.412
4	68	6.0588	66.663

Table 4: Performance metrics for the 4 kHz octave-band-filtered impulse



(a) 250 Hz octave-band filtered impulse

(b) 2 kHz octave-band filtered impulse

Figure 5: Coefficient α_{mn} plotted as a function of azimuth (m) for $n = 1, 2, 3, 4$.

4.2. Relationship to models of localization

Unlike some other models of localization, the objective of the information-theoretic model developed here is providing insight and upper bounds on performance rather than specific predictions of localization performance. In Colburn and Kulkarni's taxonomy [17, p. 272], the model follows a signal-processing approach to localization, as contrasted with psychophysical and physiological approaches. Like other signal-processing approaches such as [22, 25], the model is not congruent with human processes. Though the coefficients α_{mn} correspond to localization cues such as ITD in some instances, the relation between HRIR and location estimate is not explicitly made in terms of psychoacoustic parameters, as in [14] or neurological correlates of localization [26]. Moreover, the model does not account for "biologic constraints" due to the auditory periphery [16] and neural processing (see, e.g., [26]) that limit the spectral and temporal resolution with which localization information can be extracted from HRIR [17]. While the internal noise level and corresponding quantization interval limit the resolution with which information is extracted from a set of HRIR, homoskedastic noise in a multidimensional space of coefficients may not be a good model for the limitations of human perception. For these reasons the model cannot provide verisimilitude to human localization, particularly its more subtle aspects. For example, it does not address performance variations due to spectral integration that arise when the spectrum of the stimulus exceeds a critical band.

5. CONCLUSIONS

Spatial hearing and localization in the azimuthal plane can be interpreted as a communication problem in which scattering from the head and torso, as described by the HRIR, encodes information about source location. Information theory gives bounds on the performance of this communication channel as a function of source signal and places an upper limit on the number of sources that can be identified for a given set of conditions comprising source signal and internal noise level. Further, it indicates how to maximize performance under those conditions. In particular, an array of virtual sources distributed uniformly in azimuth does not maximize the amount of spatial information that can be encoded from noisy observations of the set of HRIR. Instead, maximum information transfer is realized when virtual source positions correspond to those azimuthal locations which are uniquely encoded by the HRIR given the quantization interval associated with a particular set of conditions.

In future work it would be possible to extend this model to elevation-angle localization in the medial plane or to more general localization of source varying in range and over the full 4π steradian of solid angle. Because the model extracts spectral features without the need for explicitly defining them, it may be of particular use to the study of localization in the medial plane for which there is not consensus regarding the spectral cues used by human listeners [16]. Similarly, the model may offer some insight into localization in the presence of multiple reflections and reverberation, as discussed in [27].

6. ACKNOWLEDGMENT

Portions of this work have been presented at the 151st meeting of the Acoustical Society of America [J. Acoust. Soc. Am. 119, 3396 (A) (2006)]. J.E.S. gratefully acknowledges seminal contributions to this work by and ongoing discussion with C. F. Gaumond.

7. REFERENCES

[1] G. R. Arrabito, B. E. Cooke, and S. M. McFadden, “Recommendations for enhancing the role of the auditory modality for processing sonar data,” *Appl. Acoust.*, vol. 66, pp. 986–1005, 2005.

[2] H. C. Hayes, “World War I—submarine detection,” *Sound*, vol. 1, no. 5, pp. 47–48, 1962.

[3] R. D. Fay, “Underwater-sound reminiscences: *Mostly* binaural,” *Sound*, vol. 2, no. 6, pp. 37–42, 1963.

[4] S. M. McFadden and M. M. Taylor, “Human limitations in towed arrray sonar recognition,” in *Proc. 24th Defense Research Group Seminar, The Human as a Limiting Element in Military Systems*, ser. DS/A/DR(83) 170, vol. 1, May 1983, pp. 431–451.

[5] S. M. McFadden and R. Arrabito, “Proposals for enhancing the auditory presentation of sonar information,” Defense and Civil Institute of Environmental Medicine, Report 94-52, November 1994.

[6] S. Richardson and C. Loeffler, “Three-dimensional auditory display of passive sonar data,” in *Collected Papers, 137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustical Association*, Berlin, 14–19 March 1999.

[7] B. E. Cooke, “Uses of spatial audio in sonar,” Defense and Civil Institute of Environmental Medicine, Contract Report CR 2002-054, Feb 2002.

[8] A. B. Baggeroer, W. A. Kuperman, and H. Schmidt, “Matched field processing: Source localization in correlated noise as an optimum parameter estimation problem,” *J. Acoust. Soc. Am.*, vol. 83, no. 2, pp. 571–587, 1988.

[9] J. R. Buck, “Information theoretic bounds on source localization performance,” in *Proc. 2nd IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM-2002)*, Washington, D.C., 2002, pp. 184–188.

[10] ——, “Fading channel capacity and passive sonar performance bounds,” in *Proc. 4th IEEE Workshop on Sensor Array and Multichannel Processing (SAM-2006)*, Waltham, MA, July 2006, pp. 294–298.

[11] T. Meng and J. R. Buck, “Rate distortion bounds on passive sonar performance,” in *Proc. 4th IEEE Workshop on Sensor Array and Multichannel Processing (SAM-2006)*, Waltham, MA, July 2006, pp. 636–640.

[12] C. F. Gaumond, “Broadband information transfer from oceanic sound transmission,” *Acoust. Res. Lett. Online (ARLO)*, vol. 5, pp. 44–49, 2004. [Online]. Available: link.aip.org/link/?ARLOFJ/5/44/1

[13] F. G. Ashby, “Multidimensional models of categorization,” in *Multidimensional models of perception and cognition*, F. G. Ashby, Ed. Hillsdale, N.J.: Lawrence Erlbaum, 1992, pp. 449–483.

[14] C. L. Searle, L. D. Braida, M. F. Davis, and H. S. Colburn, “Model for auditory localization,” *J. Acoust. Soc. Am.*, vol. 60, pp. 1164–1175, 1976.

[15] S. Sen and A. Nehorai, “Perfromance analysis of 3-D direction estimation based on head-related transfer functions,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 4, pp. 607–613, May 2009.

[16] C. Jin, M. Schenkel, and S. Carlile, “Neural system identification model of human sound localization,” *J. Acoust. Soc. Am.*, vol. 108, pp. 1215–1235, 2000.

[17] H. S. Colburn and A. Kulkarni, *Sound Source Localization*, ser. Springer Handbook of Auditory Research. New York: Springer, 2005, vol. 25, ch. Models of sound localization, pp. 272–316.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New york: Wiley, 1991.

[19] B. Gardner and K. Martin, “HRTF measurements of a KEMAR dummy-head microphone,” MIT Media Lab, MIT Media Lab Perceptual Computing Technical Report 280, May 1994. [Online]. Available: http://sound.media.mit.edu/KEMAR.html

[20] A. Farina. Inverse filter of the MIT Medialab (sic.) measurement setup. [Online]. Available: http://pcfarina.eng.unipr.it/Public/Sursound/

[21] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *J. Acoust. Soc. Am.*, vol. 87, pp. 2188–2200, 1990.

[22] W. M. Hartmann and B. Rakerd, "On the minimum audible angle—a decision theory approach," *J. Acoust. Soc. Am.*, vol. 85, pp. 2031–2041, 1989.

[23] G. H. Recanzone, S. D. D. R. Makhamra, and D. C. Guard, "Comparison of relative and absolute sound localization ability in humans," *J. Acoust. Soc. Am.*, vol. 103, pp. 1085–1097, 1998.

[24] J. M. Moore, D. J. Tollin, and T. C. T. Yin, "Can measures of sound localization be related to the precision of absolute localization estimates?" *Hear. Res.*, vol. 238, pp. 94–109, 2008.

[25] W. M. Hartmann, B. Rakerd, and J. B. Gaalaas, "On the source-identification method," *J. Acoust. Soc. Am.*, vol. 104, pp. 3546–3557, 1998.

[26] N. H. Salminen, H. Tiitinen, S. Yrttiaho, and P. J. C. May, "The neural code for interaural time difference in human auditory cortex," *J. Acoust. Soc. Am. Express Lett.*, vol. 127, pp. EL60–EL65, 2010.

[27] J. E. Summers, "Information transfer in auditoria," in *Proc. 19th Intl. Congress Acoust. (ICA)*, Madrid, 2–7 September 2007.

AUDIO AUGMENTED REALITY IN TELECOMMUNICATION THROUGH VIRTUAL AUDITORY DISPLAY

Hannes Gamper and Tapio Lokki

Aalto University School of Science and Technology
Department of Media Technology
P.O.Box 15400, FIN-00076 Aalto, Finland
[Hannes.Gamper,ktlokki]@tml.hut.fi

ABSTRACT

Audio communication in its most natural form, the face-to-face conversation, is binaural. Current telecommunication systems often provide only monaural audio, stripping it of spatial cues and thus deteriorating listening comfort and speech intelligibility. In this work, the application of binaural audio in telecommunication through audio augmented reality (AAR) is presented. AAR aims at augmenting auditory perception by embedding spatialised virtual audio content. Used in a telecommunication system, AAR enhances intelligibility and the sense of presence of the user. As a sample use case of AAR, a teleconference scenario is devised. The conference is recorded through a headset with integrated microphones, worn by one of the conference participants. Algorithms are presented to compensate for head movements and restore the spatial cues that encode the perceived directions of the conferees. To analyse the performance of the AAR system, a user study was conducted. Processing the binaural recording with the proposed algorithms places the virtual speakers at fixed directions. This improved the ability of test subjects to segregate the speakers significantly compared to an unprocessed recording. The proposed AAR system outperforms conventional telecommunication systems in terms of the speaker segregation by supporting spatial separation of binaurally recorded speakers.

1. INTRODUCTION

Audio augmented reality (AAR) aims at enhancing auditory perception through virtual audio content. Virtual auditory display (VAD) is used to present the content to the AAR user as an overlay of the acoustic environment. This principle is applicable to telecommunication systems as a new interface paradigm. Conventional telecommunication systems often provide the user only with a monaural audio stream, played back via a headset or a hand-held device. The term "monaural" refers to the fact that only one ear is necessary to interpret the auditory cues contained in the audio stream. However, face-to-face communication, which is considered the "gold standard" of communication [1, 2], is inherently binaural. In a face-to-face conversation, a listener is able to segregate multiple talkers based on their position, a phenomenon referred to as the "cocktail party effect" [3]. Monaural audio employed in conventional telecommunication systems, such as mobile phones and voice-over-IP (VoIP) softwares, does not support interaural cues and hence deteriorates the communication performance compared to face-to-face communication [4, 5].

AAR helps overcoming these limitations by embedding spatialised virtual audio into the auditory perception through VAD. The "cocktail party" principle holds also for a multi-party telecommunication scenario. Using VAD to separate the speech signals of participants spatially improves the listening comfort and intelligibility [6, 7]. In contrast to the sense of vision, auditory perception is not limited to a "field of view". The participants of a teleconference can thus be distributed all around the user, regardless of the orientation of the user. By registering the virtual speakers with the environment, the user can turn towards a conferee the same way as in a face-to-face conversation.

A major challenge in telecommunication lies in the physical distance itself, which puts limits to the naturalness of interaction with a remote end. Communication over distance suffers from a lack of "social presence", compared to face-to-face communication [8]. Through spatial audio, an AAR telecommunication system improves the sense of "presence" [9, 10] and "immersion" [6]. In this work, algorithms are presented to process binaural recordings and embed them into the auditory perception of a user. This serves as a proof-of-concept for employing AAR in a telecommunication scenario. A user study is conducted to analyse the ability of users to localise and segregate remote speakers with the proposed system.

2. EXPERIMENTAL SETUP

The basic principle of AAR is to augment, rather than replace, reality. Therefore, the transducer setup for AAR needs to be acoustically transparent to ensure unaltered perception of the real environment. If a headset is used for the reproduction of virtual audio content, acoustical transparency is achieved by capturing the real-world sounds at the ears of the user and playing them back through the earphones. Mixing these captured real-world sounds with virtual sounds is the basic working principle of "mic-through augmented reality" [11], which refers to the fact that the real world is perceived through microphones.

In this work, the MARA headset, introduced by Härmä et al. [12], is used. It consists of a pair of insert-earphones with integrated miniature microphones. Insert-earphones provide the advantage of leaving the pinnae of the listener uncovered, thus preserving the pinna cues, which are important for the localisation of real-world sounds [13]. Inserting the earphones into the ear canal minimises effects of the transmission paths from the earphone to the ear drum.

The MARA microphone signals provide a realistic representation of the acoustic environment [12]. In the proposed AAR telecommunication system, these signals are transmitted to the other end of the communication chain, where they are embedded

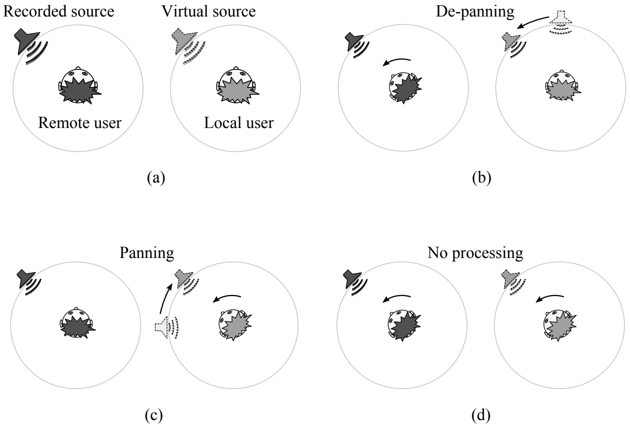


Figure 1: *De-panning and panning of binaural recordings.* (a) The source recorded at the remote end is perceived at the local end as a virtual source at the same direction. (b) De-panning is applied to compensate for head movements of the remote user. (c) Panning compensates for head movements of the local user, to register the virtual sources with the environment. (d) No processing is necessary if the head orientation of both users is the same, e.g. if both are facing the source.

into the auditory perception of a listener. The listener thus perceives the remote acoustic environment through the ears of a remote user as an overlay of the own acoustic environment. As a test case for the proposed system, a teleconference between the two ends is devised. The conference is held at the remote end, captured through a MARA headset worn by one of the participants (hereafter referred to as the "remote user"). The recording is transmitted to the user at the other end, i.e. the "local user" (cf. fig. 1a).

If both the remote and the local user keep their heads still, the local user perceives each remote conferee at a distinct direction. If the remote user rotates the head, however, the spatial cues of the virtual speakers change, which affects the perceived directions. The resulting lack of distinct spatial cues deteriorates the listening comfort and the speaker segregation ability of the local user.

To restore the spatial cues contained in the binaural recording, the head rotation at the remote end has to be compensated for. After compensation, the virtual speakers have a fixed direction relative to the local user, regardless of the head orientation of the remote user recording the conference. In a telecommunication scenario it might be desirable to employ virtual auditory display (VAD) registered with the environment for each virtual speakers. This allows the user for example to turn the head to look at a remote talker, which is a natural behaviour in face-to-face communication. In the following section, algorithms are presented to compensate for head movements of the remote user and register binaurally recorded speakers with the environment of the local user.

2.1. Compensation for head movements

To compensate for head movements during the recording, the binaural recording has to be processed such as to reposition the virtual speakers. Two measures need to be known for this "de-panning" process: The head orientation of the remote user and the position of the sources. For simplicity, the positions of the conference participants are assumed to be fixed. The head orientation of the remote user is tracked.

The aim of the de-panning process is to remove the alterations of the spatial cues introduced by head movement. These alterations occur both in the time domain and in the spectral domain. The following sections propose methods to remove or minimise these alterations.

2.1.1. Restoring interaural differences

The most important alteration of spatial cues caused by head movement during a binaural recording is a change in the time of arrival of the signal at both ears. This results in an altered interaural time difference (ITD). If, for simplicity, the ITD is assumed to be frequency-independent (cf. Wightman and Kistler [14]), it can be represented by a delay of one ear input signal with respect to the other. The head movement affects this delay. Thus, by delaying the binaural signals appropriately in the de-panning process, the ITD of a virtual speaker can be restored. $TD(\alpha)$ is the frequency-independent delay as a function of the angle of incidence (after Rocchesso [15]):

$$TD(\alpha) = \begin{cases} \frac{f_s}{\omega_0} \cdot [1 - \cos(\alpha)] & \text{if } |\alpha| < \frac{\pi}{2}, \\ \frac{f_s}{\omega_0} \cdot \left[|\alpha| - \frac{\pi}{2} + 1\right] & \text{else.} \end{cases} \tag{1}$$

with

$$\omega_0 = \frac{c}{r}, \tag{2}$$

where α denotes the angle of incidence, r the head radius (i.e. half the distance between the two ear entrances) and c the speed of sound. By delaying each signal with an appropriate $TD_{correction}$ factor, the influence of head rotation on the ITD is eliminated (cf. fig. 2, top graph).

In the frequency domain, head rotation affects the head-related transfer function (HRTF). Pinna and shoulder reflections introduce azimuth-dependent peaks and notches in the HRTF. In addition to direct sound, room reflections and diffuse sound make a compensation of HRTF alterations caused by head movements rather complex and impractical. As a simple approximation, the impact of head rotation on the spectrum of the ear input signals can be described in terms of variations of the interaural level difference (ILD). Rocchesso proposes a simple model for the head shadow effect as a one-pole/one-zero shelving filter [15]. From this model, an azimuth-dependent gain correction $LD_{correction}$ is derived to compensate for the ILD alterations caused by head movement. It is given by

$$LD_{correction}(\alpha) = 1.05 + 0.95\cos(\frac{6}{5}\alpha). \tag{3}$$

To achieve the gain correction, a high shelving filter is applied to each channel, with transfer function

$$H_{LD}(z) = k \cdot \frac{1 - qz^{-1}}{1 - pz^{-1}}, \tag{4}$$

where k is the filter gain, p is the pole and q is the zero of the filter. The pole of the filter is fixed [15]:

$$p_{hs} = \frac{1 - \frac{\omega_0}{f_s}}{1 + \frac{\omega_0}{f_s}}. \tag{5}$$

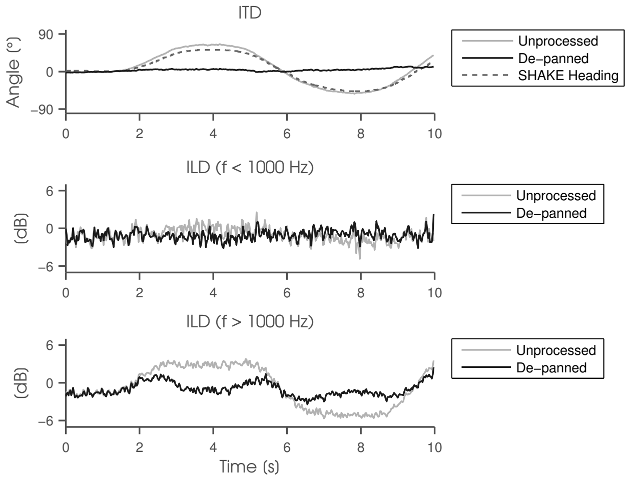


Figure 2: *Restoring interaural differences of a binaural recording.* Head movement (dashed line) causes ITD and ILD variation. The ITD is calculated as the maximum of the interaural cross correlation (IACC). Above 1000 Hz, head rotation causes ILD variations.

with f_s denoting the sampling frequency.
The gain k and zero q of the filter are chosen to meet the following two criteria: At low frequencies, the impact of head shadowing is negligible, thus the filter has a DC gain of unity. At high frequencies, the impact of head rotation on the head shadowing needs to be compensated for. At the Nyquist limit, the filter gain equals the value given by $LD_{correction}$:

$$H_{LD}(z)|_{z=1} = k \cdot \frac{1-q}{1-p} \overset{!}{=} 1, \qquad (6)$$

$$H_{LD}(z)|_{z=-1} = k \cdot \frac{1+q}{1+p} \overset{!}{=} LD_{correction}. \qquad (7)$$

Solving eq. (6) and eq. (7) for q and k yields

$$q = \frac{\phi - 1}{\phi + 1} \qquad (8)$$

for the filter zero q with

$$\phi = LD_{correction} \frac{1+p}{1-p} \qquad (9)$$

and

$$k = \frac{1-p}{1-q} \qquad (10)$$

for the filter gain k. Applying a gain factor $LD_{correction}$ to each channel via a separate shelving filter reduces the impact of head rotation on the ILD of the recorded binaural signals.

The effect of the de-panning algorithm on a binaural recording is shown in Fig. 2. The input signal is white noise, played back from a loudspeaker in a small office environment and recorded via a MARA headset. During the recording, the head orientation changes by ±60°. The resulting ITD variations are compensated for through appropriate delays, defined by $TD_{correction}$, applied to both channels. For frequencies below 1000 Hz, the ILD change due to head shadowing is negligible (cf. Fig. 2, middle graph). Above 1000 Hz, the de-panning algorithm compensates for the head shadowing effect (cf. Fig. 2, bottom graph).

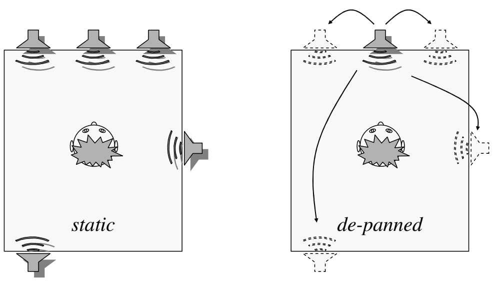


Figure 3: *Recording conditions for speaker localisation task.* For the *static* recording, the speech sample is played back from one of five different loudspeakers. For the *de-panned* recording, only one loudspeaker is used. The spatial separation of the speakers is obtained through de-panning.

2.2. Panning of binaural audio

To register the virtual speakers with the local environment, the head rotation of the local user has to be taken into account. The remote speakers are played back through VADs at fixed positions in the local environment by processing the binaural recording according to head movements of the local user. This “panning” process is analogous to the de-panning process described in the previous section. The head of the local participant is tracked and the spatial cues contained in the binaural recording are adjusted by tuning ITD and ILD. By combining the head orientations of the remote and the local user, the de-panning and the panning process are merged to a single processing stage. Instead of de-panning the recording to the original position (to compensate for head rotation of the remote user) and then panning it to the desired position (determined from the head orientation of the local user), the recording is directly panned to the desired position.

Merging de-panning and panning to a single process provides the advantage of eliminating redundant computations. Low latency is vital in an interactive telecommunication scenario. Processing the binaural audio in a single step has another major benefit: In a communication scenario it is natural for participants to turn towards the speaker. Therefore, the head orientations of both the remote and the local user are assumed to be similar, if the speaker is registered with the local user’s environment. In this case, little or no processing is applied to the binaural recording (cf. Fig. 1d), as the actual source position, relative to the remote user, and the desired source position, determined from the head orientation of the local user, are similar or identical.

3. USER STUDY

To evaluate the performance of the proposed AAR telecommunication system under controlled conditions, a formal user study was conducted. 13 test subjects with normal hearing were used in a within-subjects design. 5 of the test subjects were students of the Department of Media Technology of the Helsinki University of Technology. Having vast experience in using and assessing spatial audio, they were classified as “professional listeners”. The other 8 subjects had little or no experience with spatial audio, and were thus classified as “naïve listeners”. The test subjects were presented with a binaural recording simulating a teleconference. The conference was recorded via a MARA headset in a room with a

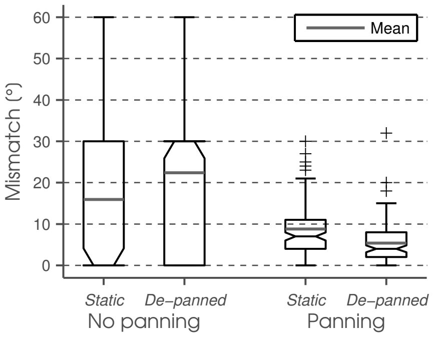


Figure 4: *Absolute angle mismatch.* The mean and median absolute angle mismatch is significantly higher without panning than with panning enabled.

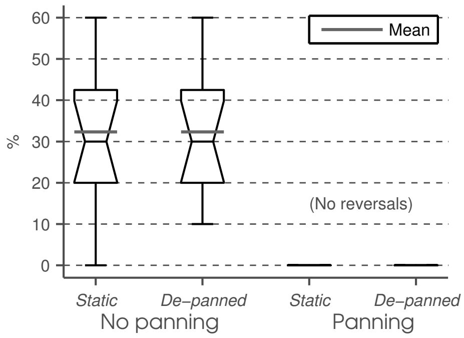


Figure 5: *Front–back reversals.* The *static* and *de-panned* condition yield the same mean reversal rate. No front–back reversal is observed with panning enabled, in either condition.

reverberation time of 0.3–0.5 s.

To analyse the performance in each test condition, mean and median values are analysed and compared using parametric ANOVA and non-parametric Friedman analysis. To compare two matched conditions, a paired t-test is performed. Judgements are based on 0.05 significance level. Results are given as p-values for rejecting the null hypothesis. For multiple comparisons, a post test with Tukey-Kramer correction is applied.

3.1. Localisation of virtual speakers

To test the ability of test subjects to localise a speaker recorded with the MARA headset, a recording was used consisting of ten repetitions of a male speech sample from the "Music for Archimedes" CD [16]. The sample duration is about 11 s, with 1 s of silence between each repetition. Two different conditions were tested: *static* and *de-panned*.

For the *static* condition, the binaural recording was made using five loudspeakers (cf. Fig. 3): three in front (at 30°, 0° and $-30°$ azimuth), one to the right (at $-90°$), and one in the back (at 150°). The anechoic speech sample was played from each loudspeaker, in random order, with each direction occurring twice.

For the *de-panned* condition, a situation was assumed where the remote participant recording the conference is turning towards the currently active speaker. To simulate this scenario, one loudspeaker in front of the MARA headset user at 0° azimuth was used for the recording. The speech sample was the same as in the *static* condition. The recorded sample was then de-panned to encode the interaural cues of the same azimuth angles as used in the *static* condition (i.e. 150°, 30°, 0°, $-30°$ and $-90°$). The listener should thus perceive the speakers as emanating from these directions, even though they were recorded with the remote user facing them. Again, the order of the directions was randomised, with each direction occurring twice.

Presented with a binaural recording from the MARA headset, the test subjects were asked to identify the direction of the speakers from twelve possible directions in the horizontal plane, spaced 30°. The test was conducted using an unprocessed *static* and a *de-panned* recording. To minimise learning effects, the order of the recordings was randomised among subjects.

In the second task, the head of the test subject was tracked, to register the virtual speakers with the environment. The test subjects were asked to turn towards the speakers to specify their directions. Again, this was tested in the *static* and *de-panned* condition, in random order.

3.1.1. Angle mismatch

An objective measure to determine the performance of test subjects to localise speakers from the MARA recording is the angle mismatch between the guess β of the test subject and the actual recording angle α. In the second part, where test subjects are asked to turn towards the speaker, the mismatch is calculated as the offset between the playback angle α and the head orientation β of the test subject.

The mismatch is compensated for front–back reversals and they are analysed separately. A front–back reversal occurs, when the test subject perceives the source as being in front when in fact it is in the back, and vice versa. The error due to the reversal is removed from the angle mismatch, as it would severely distort the measurement results [17].

Boxplots of the mean and median absolute angle mismatches in both subtasks are shown in Fig. 4. To compare performance under the two conditions in each subtask, a paired two-way analysis is performed on the absolute values of the angle mismatches. Applying a two-way ANOVA to the data of subtask I reveals that the mean absolute angle mismatch without panning is significantly smaller with the *static* recording (15.9°) than with the *de-panned* recording (22.4°), $F(1, 12) = 6.57, p_{Cond} = 0.0110$. The Friedman analysis yields an analogous result: The median of the absolute angle mismatch is significantly smaller with the *static* recording (0°) than with the *de-panned* recording (30°), $\chi^2(1, n = 13) = 6.13, p_{Cond} = 0.0133$. No significant difference between subjects is found (ANOVA: $F(1, 12) = 1.02, p_{Subj} = 0.4315$, Friedman: $\chi^2(12, n = 2) = 12.97, p_{Subj} = 0.3714$).

With panning enabled the order is reversed: The mean absolute angle mismatch is significantly lower in the *de-panned* condition (5.4°) than in the *static* condition (8.8°), $F(1, 12) = 14.42$, $p_{Cond} = 0.0002$. The Friedman analysis indicates a significantly smaller median with the *de-panned* recording (4.0°) than with the *static* recording (7.0°), $\chi^2(1, n = 13) = 16.83$, $p_{Cond} = 0.0000$. Again, no significant difference between subjects is found (ANOVA: $F(1, 12) = 1.59, p_{Subj} = 0.0952$, Friedman: $\chi^2(12, n = 2) = 15.87$, $p_{Subj} = 0.1972$). A Tukey-Kramer post test indicates a significantly higher mean absolute angle mismatch in the *de-panned* condition without panning than in

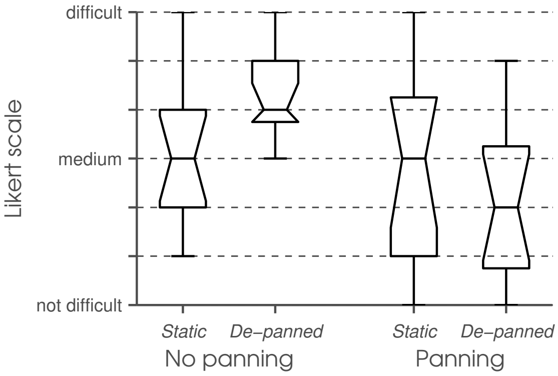


Figure 6: *Perceived difficulty of speaker localisation task.* Localising speakers in the *de-panned* condition without panning is perceived to be significantly more difficult than in the *de-panned* condition with panning enabled.

any of the other conditions.

3.1.2. *Front–back reversals*

The mean front–back reversal rate is equal in both tested recording conditions without panning: 32 percent (cf. Fig. 5). This is close to chance level, as 2 out of the 10 tested directions were at the extreme right ($-90°$), where no reversal can occur. Most of the reversals (83 percent in the *static* and 85 percent in the *de-panned* case) occurred when a source was mistakenly perceived to be in the back. The chance of this kind of error was increased by the fact that frontal source directions prevailed in the test. A Lilliefors normality test indicates that the error rates follow a normal distribution.

With panning enabled, no front–back reversal was observed. All test subjects managed to correctly identify whether a source was in front or in the back when asked to turn towards the virtual source.

3.1.3. *Perceived difficulty*

As a subjective measure, test subjects were asked to judge the perceived difficulty of each subtask. The difficulty was marked on a balanced seven-step Likert scale [18], ranging from *not difficult* to *difficult*, with *medium* marking the centre point. To compare the perceived difficulty of each subtask, a Friedman analysis is performed on the medians (cf. fig. 6). The null hypothesis is rejected for the first subtask, indicating that localisation in the *static* condition is perceived to be significantly less difficult than in the *de-panned* condition, $\chi^2(1, n = 13) = 7.36$, $p = 0.0067$. No significant difference between conditions is found in subtask II, $\chi^2(1, n = 13) = 1.6$, $p = 0.2059$. When comparing both subtasks, the Friedman analysis indicates a significant difference between all tests, $\chi^2(3, n = 13) = 14.5221$, $p = 0.0023$. A post test with Tukey-Kramer correction reveals the speaker localisation in the *de-panned* case without panning to be perceived significantly more difficult than speaker localisation in the *de-panned* case with panning.

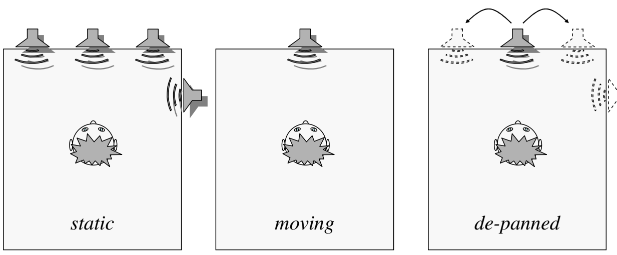


Figure 7: *Recording conditions for speaker segregation task.* For the *static* recording, a separate loudspeaker is used for each speaker. The *moving* and *de-panned* recordings are obtained using just one loudspeaker. De-panning is applied to separate the speakers on the *de-panned* recording spatially, thus simulating the speaker positions used in the *static* recording.

3.2. Segregation of virtual speakers

Speech samples from the TIMIT database [19] were recorded via the MARA headset. Two groups of four male speakers were chosen from the database. Twenty speech samples were recorded per tested condition, five from each speaker. The speakers talk in turns, in random order. The segregation performance is tested in three different conditions: *static*, *moving* and *de-panned* (cf. fig. 7). In the *static* condition, each speaker is assigned one of four different loudspeakers in the recording hall, at $60°$, $30°$, $0°$, and at $-30°$. This simulates a situation where the conference participants are seated around a table with the MARA headset user. In the *moving* condition, just one loudspeaker in front of the MARA user is used. All four speakers are recorded at $0°$ azimuth. This simulates a situation where the MARA headset user turns towards the active speaker during the simulated conversation. For the *de-panned* condition, the same recording setup is used as in the *moving* condition, but de-panning is applied to each recorded speaker, to yield the same perceived speaker directions as in the *static* condition, in random order.

In each condition, the ability of test subjects to segregate the four speakers on the binaural recording is tested. The test subjects are presented with the binaural recordings and asked to mark the words of one of the four speakers in each condition. The *static* and *de-panned* condition are also tested with panning enabled, by tracking the head of the test subject. This way, the recorded speakers are registered with the environment, allowing the test subjects to turn towards them during the test.

The segregation task is repeated three times, to analyse the impact of learning effects on the performance. To counterbalance the order in which the conditions are presented, the order is governed by a *Latin square* [20], and randomised among subjects.

3.2.1. *Error rates*

The segregation performance is measured in terms of the number of correctly identified speaker turns. The most striking result is the speaker segregation performance in the *moving* condition, producing the highest error rates in all three rounds (cf. fig. 8). The differences between mean and median error rates are found to be significant (ANOVA: $F(4, 2) = 26.34$, $p_{Cond} = 0.0000$, Friedman: $\chi^2(4, n = 3) = 71.34$, $p_{Cond} = 0.0000$). Applying a Tukey-Kramer post test to the ANOVA results indicates that

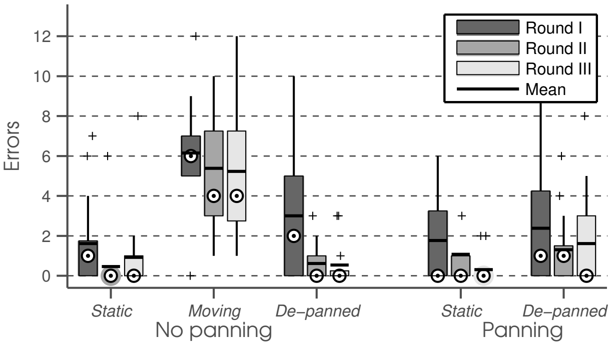


Figure 8: *Error rates of speaker segregation task* for identifying 5 turns of the speaker in question. The mean and median error rates in the *moving* condition are significantly higher than in all other conditions in round II and III. The performance of test subjects improved significantly from round I to round II.

the *moving* condition leads to significantly higher mean error rates compared to all other conditions, in all three rounds. No significant difference is found between the other four conditions, i.e. *static* and *de-panned* with and without panning. Similar conclusions can be drawn from a Tukey-Kramer post test of the Friedman analysis results. In rounds II and III, test subjects performed significantly worse in the *moving* condition than in all other conditions. No significant difference is found between the mean ranks of the *static* and *de-panned* conditions in any of the three rounds. Whether or not panning is used to register the virtual speakers with the environment has no significant impact on the performance, as indicated by a two-way ANOVA, $F(1, 1) = 2.33, p_{Pan} = 0.1287$.

To identify learning effects, the segregation performance of all three rounds is compared. A two-way ANOVA indicates significant differences between the mean error rates in the three rounds, $F(2, 4) = 5.96$, $p_{Rnd} = 0.0031$. A Friedman analysis yields analogous results regarding the median error rates, $\chi^2(2, n = 5) = 14.98$, $p_{Rnd} = 0.0006$. A Tukey-Kramer post test reveals a significant improvement of the segregation performance from round I to round II. No significant improvement from round II to round III is found. A two-way ANOVA indicates that no significant interaction effects exist between the test round and the test condition, $F(2, 8) = 0.48$, $p_{Int} = 0.8699$. The improvement after round I is thus independent of the test condition.

3.2.2. *Perceived difficulty*

A Friedman analysis indicates a significant difference between the perceived difficulty of the five test conditions, $\chi^2(4, n = 11) = 74.44$, $p = 0.0000$. Two test subjects were removed from this analysis due to missing entries. A Tukey-Kramer post test reveals the *moving* condition to be perceived significantly more difficult than all other conditions, as depicted in Fig. 9. No significant difference is found between the other conditions.

3.3. Comments of test subjects

One of the most stated problems in the speaker localisation task was inside-the-head locatedness [13]. Test subjects reported difficulties to localise sound sources that were straight ahead, as they often lacked externalisation. This was said to be confusing. Some test subjects pointed out a lack of depth in the *de-panned* recording. Whereas the sound sources appeared to be positioned on a "clear circle" in the *static* recording, in the *de-panned* recording they seemed to be positioned on a "straight line", ranging from the far left to the far right of the listener. This made it more difficult to map sources to a virtual circle than in the *static* case.

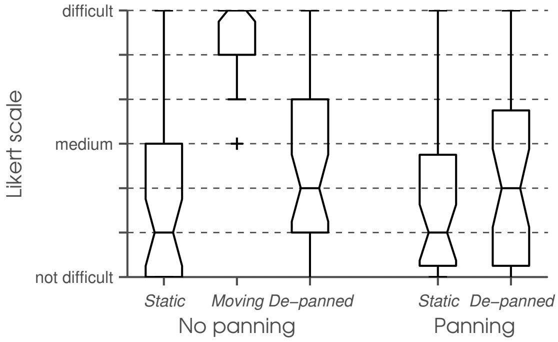


Figure 9: *Perceived difficulty of speaker segregation task.* The *moving* condition is perceived to be significantly more difficult by test subjects than all other conditions.

Most test subjects pointed out difficulties to distinguish speakers in the *moving* recording. Some test subjects said they became more acquainted with the voice of the speaker in question towards the end of the test, and managed to segregate the speakers based on their accents or articulations. In the other test conditions test subjects reported to rely mainly on the direction when segregating different speakers.

Only one test subject named the head tracking as a helpful factor in the speaker segregation task. Another subject stated that turning towards the speaker in question made the segregation task indeed more difficult, as it was easier to localise and identify a speaker a bit off the centre. Yet another test subject named inside-the-head locatedness as a cue for segregating the speakers: After turning towards the speaker in question, that speaker was not externalised anymore, which clearly separated him from the other speakers in the recording.

4. DISCUSSION

The *static* case, made with several loudspeakers at fixed positions, and recorded without head movement, represents the "ideal" case of a binaural recording, preserving the spatial cues of all speakers. In the *de-panned* recording, simulating a situation where the MARA headset user moves the head during the recording, interaural cues are restored by compensating for the head movements through the de-panning algorithm. If no panning is applied during playback to register the recorded speakers with the environment, the *de-panned* recording yields a significantly larger mean and median absolute angle mismatch between the perceived and the actual direction of the recorded speakers than the *static* recording. This indicates that the de-panning algorithm cannot fully restore the spatial cues contained in the recording. Test subjects perceived localisation with the *de-panned* recording to be significantly more difficult than with the *static* recording. This may be related to the fact that some test subjects perceived the speakers in the *de-panned* recording to be positioned on a line, whilst in the *static* recording they appeared to reside on a circle around the listener, with distinct directions.

With head tracking and panning enabled, the mean and median absolute angle mismatch decreased significantly. Test subjects localised speakers significantly more accurately by turning towards them than by indicating their directions. The reduced localisation blur, defined as the minimum audible displacement [13], achieved by facing the virtual speakers implies that registering virtual sources with the environment through panning may lead to better spatial separability of the sources. This is seen as a major benefit in a telecommunication scenario. When comparing the two test conditions with panning enabled, the *de-panned* condition leads to a significantly better localisation performance. As test subjects turn towards the de-panned speaker, their head orientation approximately matches the head orientation during the recording, therefore nearly unprocessed audio is delivered to the test subjects (c.f. fig. 1d). Turning towards a virtual source recorded off the centre, as in the *static* case, increases the localisation blur significantly, as the panning algorithm fails to fully restore the spatial cues.

No effect of the recording condition on the number of front–back reversals is found. The *de-panned* recording does not yield a higher rate of reversals than the *static* recording. We assume front–back reversals to be mainly a result of the ambiguity of interaural cues in general, not of the processing involved in generating them. A more striking finding, however, is the fact that with head tracking and panning enabled, no front–back reversal occurred in any of the 260 observations. This is a strong argument for the hypothesis that panning improves the localisation performance. When a test subject turns the head to search for the virtual sound source, the interaural cues change accordingly, indicating unambiguously whether the source is in front or in the back. Even test subjects without any prior experience with spatial audio and head tracking instinctively interpreted these motional cues correctly.

Results of the speaker segregation task prove the importance of interaural cues to segregate multiple speakers. The *moving* condition, which contains little or no interaural cues to separate speakers, leads to significantly higher mean and median error rates than the *static* and *de-panned* cases, which contain natural or algorithmically restored interaural cues. Even after being presented with the same recording for the third time in round III, the median error rate of test subjects when trying to identify the 5 turns of the speaker in question is 4. Some subjects stated their choices in the *moving* case to be based on pure guessing, others marked no turn at all. In all other conditions the median error rate in round III drops to 0, indicating that more than 50 percent of the test subjects managed to identify all speaker turns correctly. The result is supported by the perceived difficulty, with the *moving* condition rated significantly more difficult than all other conditions. This underlines the importance of spatial cues to segregate multiple speakers in a telecommunication scenario.

No significant differences are found between the *static* and *de-panned* case regarding the speaker segregation. Whilst the de-panning has a negative effect on the speaker localisation, it does not deteriorate the speaker segregation performance. Compared to an unprocessed binaural recording with no or misleading interaural cues, such as the *moving* recording, de-panning significantly improves speaker segregation, and theoretically yields the same performance as the ideal case of a *static* recording devoid of head movements.

The segregation performance improved significantly from round I to round II. This is attributed to the fact that test subjects became acquainted with the test procedure and the a priori unfamiliar voices of the speakers used in the test. No significant improvement from round II to round III is found, indicating that learning effects vanish after round I.

5. CONCLUSIONS

An audio augmented reality (AAR) telecommunication system based on the transmission of binaural recordings from a MARA headset is presented. The binaural recordings preserve the spatial cues of recorded sound sources, yielding a listening experience similar to the natural auditory perception of an environment. Head movements distort the spatial cues and thus the perceived directions of the recorded sound sources. A de-panning algorithm is presented to restore the perceived directions. The localisation accuracy of virtual sources contained on a de-panned recording was analysed in a formal user study with 13 test subjects. After de-panning, test subjects were able to localise speakers in a binaural recording, though with a significant increase of the mean absolute angle mismatch compared to an unprocessed recording not distorted by head movements.

A panning algorithm adjusts the binaural playback according to head movements of the listener, to register the binaurally recorded sound sources with the environment. With panning enabled, the localisation performance of test subjects improved significantly. The test subjects interacted with the system intuitively, using head rotations to “search” for the virtual sources. No significant performance difference was found between subjects, even though about half of the test subjects had no previous experience with spatial audio or head tracking. These results imply that the proposed system is suitable also for “naïve” users. By registering the virtual sources with the environment, no front–back reversal occurred, i.e. all test subjects correctly determined whether a source was in front or in the back.

To analyse their ability to segregate multiple recorded speakers, test subjects were asked to identify speaker turns on a binaural recording. Interaural cues are shown to improve the segregation performance of test subjects significantly, compared to a recording with no interaural cues. In case of misleading spatial cues, i.e. arbitrary changes in the perceived directions of the sources due to head movements during the recording, the performance is expected to be even worse. No significant difference is found between the recordings containing interaural cues. The *de-panned* recording, in which the spatial cues are algorithmically restored, does not lead to a significantly worse performance than the ideal case, an unprocessed binaural recording of sound sources separated in space, devoid of head movements. In a telecommunication scenario, the de-panning algorithm restores the perceived directions of speakers and enhances the ability of a listener to segregate the participants of a meeting. This is assumed to improve the listening comfort and the ability to follow a remote conversation, which is a major argument for the use of AAR in a telecommunication scenario.

Transmitting a binaural recording of one’s environment via a MARA headset is a simple yet effective way to share auditory perception over distance. Tackling issues related to head movements with the algorithms proposed in this work allowed both experienced an inexperienced users to localise virtual sources on a binaural recording. This significantly improved the ability of test subjects to segregate multiple sources on the recording. Due to their simplicity, the proposed de-panning and panning algorithms run on a standard PC, with a responsiveness that was found to be sufficient for the test scenario. System lag was an issue only in the case of fast head movements, due to the limited update rate of the head tracking device. The processing is based on simple ITD (interaural time difference) and head shadowing models, hence the system does not require a dataset of head-related transfer functions (HRTFs). This makes it transferable and relatively robust against individual HRTF variations.

In terms of future research, parametrisation of the algorithms could account for individual differences among users and various recording environments and yield more accurate spatial cues. This might improve the localisation accuracy of virtual sources. A central aspect of AAR is the combination of real and virtual auditory content. An issue further to be investigated upon is the mixing of binaural recordings from a remote end with the pseudo-acoustic environment, perceived through the MARA headset. The biggest limitation of the proposed system is that it currently supports only one virtual source at a time, i.e. speakers talking in turns. To allow for multiple simultaneous speakers, a time-frequency decomposition approach as employed in Directional Audio Coding (DirAC) [21] could be integrated to the system, to segregate and process each speaker individually.

The proposed implementation of an AAR telecommunication system using VAD might serve as a valuable tool to enhance existing telecommunication systems and help overcome the gap to face-to-face communication.

6. ACKNOWLEDGMENT

The research leading to these results has received funding from Nokia Research Center [kamara2009], the Academy of Finland, project no. [119092] and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636].

7. REFERENCES

[1] B. A. Nardi and S. Whittaker, "The place of face-to-face communication in distributed work," in *IN P. HINDS AND S. KIESLER (EDS.), DISTRIBUTED WORK*. MIT Press, 2002, pp. 83–112.

[2] P. Rohde, P. M. Lewinsohn, and J. R. Seeley, "Comparability of Telephone and Face-to-Face Interviews in Assessing Axis I and II Disorders," *Am J Psychiatry*, vol. 154, no. 11, pp. 1593–1598, 1997.

[3] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[4] R. W. Lindeman, D. Reiners, and A. Steed, "Practicing what we preach: IEEE VR 2009 virtual program committee meeting," *Computer Graphics and Applications, IEEE*, vol. 29, no. 2, pp. 80–83, March-April 2009.

[5] M. Billinghurst, H. Kato, K. Kiyokawa, D. Belcher, and I. Poupyrev, "Experiments with face-to-face collaborative ar interfaces," *Virtual Reality*, vol. 6, no. 3, pp. 107–121, 2002.

[6] B. Kapralos, M. R. Jenkin, and E. Milios, "Virtual audio systems," *Presence: Teleoper. Virtual Environ.*, vol. 17, no. 6, pp. 527–549, 2008.

[7] R. Drullman and A. W. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *The Journal of the Acoustical Society of America*, vol. 107, no. 4, pp. 2224–2235, 2000.

[8] M. H. Bazerman, J. R. Curhan, D. A. Moore, and K. L. Valley, "Negotiation," *Annual Review of Psychology*, vol. 51, no. 1, pp. 279–314, 2000.

[9] R. Shilling and S. B. Cunningham, *Virtual auditory displays*, ser. Handbook of Virtual Environments. Mahwah NJ: Lawrence Erlbaum Associates, 2002.

[10] H. Lehnert and J. Blauert, "Virtual auditory environment," in *Advanced Robotics, 1991. 'Robots in Unstructured Environments', 91 ICAR., Fifth International Conference on*, June 1991, pp. 211–216 vol.1.

[11] R. Lindeman, H. Noma, and P. Goncalves de Barros, "An empirical study of hear-through augmented reality: Using bone conduction to deliver spatialized audio," in *Virtual Reality Conference, 2008. VR '08. IEEE*, March 2008, pp. 35–42.

[12] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, 2004.

[13] J. Blauert, *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press, October 1996.

[14] F. L. Wightman and D. J. Kistler, "Factors affecting the relative salience of sound localization cues," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1997, pp. 1–23.

[15] U. Zölzer, Ed., *DAFX:Digital Audio Effects*. John Wiley & Sons, May 2002, ch. Spatial Effects by D. Rocchesso, pp. 137–200.

[16] Bang & Olufsen, "Music for Archimedes," CD B&O 101, 1992.

[17] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[18] H. J. Gardner and M. A. Martin, "Analyzing ordinal scales in studies of virtual environments: Likert or lump it!" *Presence: Teleoper. Virtual Environ.*, vol. 16, no. 4, pp. 439–446, 2007.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.

[20] N. Rapanos, "Latin squares and their partial transversals," in *Harvard College Mathematics Review*, S. D. Kominers, Ed. Harvard College, 2008, vol. 2, pp. 4–12.

[21] V. Pulkki and C. Faller, "Directional Audio Coding: Filterbank and STFT-Based Design," in *Preprint 120th Conv. Aud. Eng. Soc.*, 2006.

VIRTUAL AUDITORY CUEING REVISITED

Derek Brock, Brian McClimens, and Malcolm McCurry

U.S. Naval Research Laboratory,
4555 Overlook Ave., S.W.,
Washington, DC 20375 USA
derek.brock@nrl.navy.mil

ABSTRACT

In a 2002 dual-task experiment involving opposing screens, virtual auditory cueing significantly improved measures of performance and reduced the effort needed to pursue both tasks. An effort to model this result revealed that supplementary empirical information was needed and a new study, reported here, was subsequently carried out. In addition to focusing on modeling issues, the new study also investigated the contribution of both an augmented auditory reality (AAR) style of display and aurally based event identity information. The previously observed benefits of auditory cueing were replicated, but more importantly, neither AAR-based cueing nor the removal of aural identity information meaningfully impacted performance. These findings suggest that simpler auditory information designs for visual attention may, in fact, be preferable to richer designs, and that aural overlays on visual information are unnecessary, but not disadvantageous, in single-use auditory displays.

1. INTRODUCTION

The outcome of four manipulations in a 2002 human performance study involving a dual task displayed on opposing screens demonstrated that virtual (3D) auditory cues can both significantly improve performance and reduce the effort otherwise needed to carry out both tasks [1]. Critical measures in this study included decision response times for secondary task events and switches of attention between tasks. A subsequent effort to comparatively model the baseline and one of these manipulations in the EPIC cognitive architecture [2], using these and derived measures as criteria, revealed that further empirical knowledge is needed for a comprehensive account of how auditory cues assist performance in this operational paradigm. As a consequence, a new dual-task performance study was carried out in 2009 and is reported here.

In addition to addressing modeling issues, the new study specifically investigated the contribution of both an augmented auditory reality (AAR) style of display [3] and aurally conveyed event identity information (c.f., [4], [5], and [6]) through the incorporation of additional manipulations. Interest in both of these latter questions is motivated by the expectation that virtual auditory displays will eventually be routinely used for, or will supplement, more than one visual information and/or decision task at a time. Planned reductions in crew sizes on Navy platforms currently under development, for example, mean that future watchstanders and decision makers will oversee a broader array of concurrent tasks and, in many instances, manage more demanding workflows. Updated workstations designed for this purpose, now referred to as the "Common Display System" [7] (see Figure 1), will in fact feature more visual display space—and thus will likely present more simultaneous information and/or decision tasks—than an individual can monitor at once, which in turn will call for the design of effective strategies for guiding visual attention.

Virtual auditory cueing offers a recommended and reliable technique for this purpose [8], [1], but nevertheless remains subject to a range of human factors and design principles that in many cases have yet to be systematized [9]. The ideal, in what is already an aural information environment in Navy command settings because of voice communications and other uses of sound, is to indicate where visual attention is needed, but to do so in a way that is succinct and unambiguous and avoids overloading or confusing the operator.

2. BACKGROUND

2.1. Test bed

The dual task employed in the 2002 study [1] and in the work reported here (as well as in earlier studies [10], [11]), combines a challenging, continuous tracking activity with a series of rule-

Figure 1: Three-screen console configuration of the Common Display System, the new information workstation being acquired for the U.S. Navy's modernization program and next-generation surface ships.

based decision events. The first of these—the "tracking task"—is performed with a joystick and is presented to participants as their primary activity. The latter, which involves a procession of blips on a simulated radar screen—and is thus referred to in this paper as the "radar task," and also as the "secondary task"—requires participants to evaluate each item's behavior as it moves down the screen and to record their decisions on a numeric keypad after the blips change color. Task scenarios involve three types of blips that are numbered from 1 to 6 and move down the display in respectively distinct fashions that are easy to visually assess as hostile or neutral according to a predefined set of rules. Decision entries require participants to make two key presses, the first indicating their assessment and the second designating the assessed blip by its onscreen number.

Both tasks are visually demanding, and when they are placed on opposing right and left screens, in a manner that corresponds to the "outer" two of the three screens in the Common Display System design, a notable amount of mental and physical effort is needed to prosecute them at the same time [10], [1]. Much of this effort is a result of the distance between the outer screens as well as having to compensate for the loss of peripheral visual access to the opposing task that occurs when one looks to the left or right. In principle, however, a meaningful degree of these performance demands can be reduced by simply notifying the operator when the secondary task requires a response. In the previous dual-task studies cited above, 3D auditory cues have been shown to be a robust and effective technique for bringing this type of information to the user's attention.

2.2. Auditory cueing

The radar task in these earlier studies was augmented with a set of three easily differentiated sounds—one for each type of blip—that signaled the onset of color changes and, thus, when blip decision responses were required. To give the auditory cues a deictic (or indexical) component [5], spatial information that could be intuitively indexed to the visual display was dynamically added, and the sounds were rendered binaurally in stereo headphones with a non-individualized head-related transfer function (HRTF). In the earliest auditory study with the dual task, each sound was spatially correlated with the visual location of its corresponding blip [11]. A simpler scheme indicating only the location of the radar task itself—rather than the location of each blip—was subsequently found to be equally effective in the dual-task study carried out in 2002 [1]. It is important to note, however, that the auditory cues in both of these spatialization schemes were not perceptually co-located with the visual display information they were designed to index.

2.3. Performance questions from the 2002 experiment

The 2002 dual-task experiment was conceived in part as an initial study of the notion of a "mixed-use" virtual auditory display— specifically, an auditory display in which information designated for more than one activity is sounded. A two factor design was employed that crossed two levels of auditory cueing for the tracking task with three levels of auditory cueing for the radar task. The first level in both factors was silence (no-sound), and the remaining levels were, respectively, an auditory cue for poor tracking performance, spatially indexed to the location of the tracking task, and the two spatialization schemes for the radar task that were described in the previous section (the blip location scheme and the simpler task location scheme), both employing the same set of three sounds. This resulted in a performance baseline, three manipulations with auditory cues for only one of the tasks, and two manipulations in which auditory cues were used for both tasks.

Three measures of performance were collected: tracking error and radar task response times (both recorded by the dual-task software), and counts of the number of attention shifts participants made during each exercise, which the experimenter recorded manually on a hand-held computer. With only minor qualifications, a similar pattern of significance emerged for each measure, which supported an encouraging overall result. Specifically, while almost no performance improvement was associated with the auditory cue for the tracking task (suggesting it may have been poorly conceived), the addition of this separately designated alert did not meaningfully impact the significant performance improvements that resulted across the board when auditory cueing was used for the radar task. Put another way, both of the virtual auditory cueing schemes for the radar task had significantly positive impacts on overall dual-task performance and, perhaps more importantly, these improvements persisted in the mixed-use manipulations.

The study, however, also left a number of underlying performance questions unanswered, which became readily apparent when an effort to model and explain the pattern of results with a cognitive architecture was undertaken [12]. Cognitive architectures are essentially theoretical computational frameworks for building explanatory models of human performance based on the constraints of human perceptual, cognitive, and motor processing. Several such architectures exist, but the EPIC architecture ("Executive Process-Interactive Control") was chosen for this endeavor in part because its framework for auditory processing is somewhat more complete than that of other architectures [2].

Cognitive modeling typically begins with a performance study, reduces the performance requirements to a theoretical sequence of goal-directed actions, and then evaluates the resulting model in terms of its correspondence with the observed data. Models of complex activities are often forced to make a number of conjectural assumptions due to gaps in underlying knowledge, and this proved to be the case in modeling the 2002 study. Key unknowns faced in the modeling work include:

- *The basis for switching attention between tasks in the absence of perceptual cues*. Specific questions include how decisions to switch attention are made and how time on task (dwell time) is allocated such that patterns in the hand-collected attention shift and (inferred) task dwell time data can be explained.
- *The radar screen inspection strategy*. The radar screen varies from being empty to showing several blips at once that may or may not have changed color (note that blips that have changed color require a response). How many and which blips are assessed before returning to the tracking task? Are blips assessed before they change color?
- *The blip assessment process*. Questions include how the relevant visual information needed to assess an individual blip is gathered, how long this takes, and whether this is done with a single "look," over multiple looks, or both.

- *Performance associated with auditory cueing*. Do auditory cues prompt immediate switches of attention or is some latency involved? Does the correspondence between aural identity and blip type speed the blip assessment process?

Carefully reasoned answers for these (and other) questions were explored and settled on, but it was also recognized that additional empirical measurements were needed. Accordingly, plans were made for a new dual-task study and the scope of the modeling effort was narrowed to providing an account of performance differences between the baseline (no-sound) condition and the manipulation involving only the spatially simpler of the two auditory cueing schemes for the radar task.

The resulting comparative cognitive model of dual-task performance in these two conditions incorporated a mix of parametric and theoretically plausible solutions, which in some cases (though not others) amounted to predictions that could be empirically tested. Switches of attention to the radar task were deemed to be prompted by knowledge of its status, characterized as the number of blips present, which, in turn, dictates time spent on the tracking task. Strategies for inspecting the radar screen and assessing blips were taken to be both subject to numerous individual differences and too opaque to characterize without eye-tracking studies. (A single-screen variant of the dual-task has subsequently been used with an eye tracker to examine these two issues [13].) As a consequence, solutions for these aspects of performance were parsimoniously modeled in algorithmic terms, and parameterized to balance the demands of both tasks; additionally, it was conjectured that, when possible, blips are assessed before they change color (that is, before a response is required). Finally, related empirical work at that time [14] suggested that responses to auditory cues entail a latency period of approximately 850 ms, and it was conjectured that auditory identity did not measurably facilitate blip assessments.

3. OBJECTIVES OF THE CURRENT STUDY

The study reported below was developed to gather new dual-task performance measures and test several of the model-based predictions outlined above, and also to investigate additional design issues that are thought to be relevant to the successful implementation of mixed-use auditory displays in future Navy decision environments. In particular, the utility of auditory cueing in such settings will largely depend on the ability of deictic sounds to reliably facilitate the performance of concurrent information tasks when these sounds are used in conjunction with the virtual presentation of multiple channels of voice communications (see [15]). This context for auditory design ultimately involves balancing aural attention at more than one level: balancing competing auditory functions that are intended for operators individually, such as auditory cues and voice communications, and balancing competition between the individual's auditory display and sounds in the public setting, such as face-to-face conversation among team members, intercoms, shipboard alerts, ambient noise, and so on.

To minimize the potential for confusion among auditory sources and their informational meanings, it can be argued that virtual auditory displays in this context should be simple (i.e., no more elaborate than necessary) and should function as a fixed aural overlay on the individual operator's visual environment. Simulating the manner in which sounds in the real world are ordinarily perceived as co-located with their apparent sources, regardless of the orientation of the listener's head, is the function of an augmented auditory reality (AAR) display. An important virtue of this type of rendering is that "attaching" or "fixing" a sound to or at a meaningful visual location effectively makes any deictic function the sound is intended to have unambiguous because no perceptual mapping is involved—the sound appears to arise and persist for its duration at the place the listener is intended to look.

Using AAR to simplify auditory deixis in this way is consistent with the broader contention made in the previous paragraph, that auditory displays should be, in principle, no more elaborate than the performance context of any corresponding task calls for. Sounds can be designed to support a multiplicity of information functions—deixis, onset, identity, and disposition, to name a few—but it may well be the case that operators only make use of the information functions present in a particular instance of sound that are the most effective for the purpose at hand. If so, they can be said to adhere to a principle of "least aural effort," implying that any additional task-related auditory information that is superfluous or more readily acted upon from another cognitive or perceptual source will be ignored, if possible. A corollary to this conjecture is that excessive elaboration may be counterproductive.

An immediate test of this notion of least aural effort in the present dual task is the question of whether the correspondence between aural identity and blip type appreciably facilitates the blip assessment process. Another test is whether the unambiguous deixis AAR provides is measurably better than the spatialially relative deixis that is provided by a non-augmented (auditory) reality (NAAR) style of virtual auditory display. Positive differences, if seen in both tests in the same context, could be taken as evidence in support of this proposal, as could a lack of differences, if there is evidence that other, more readily exploited, task information is also available.

Consequently, the new experiment was designed in part to be a replication of the two the manipulations from the 2002 study that were modeled in EPIC—the baseline condition and the condition in which only spatially simplified auditory cues for radar task were used—and in part to investigate the two comparative design questions posed above—the use of an AAR vs. an NAAR display and the relative importance of auditory identity information—in preparation for follow-on studies with a new test bed that will explore other issues for mixed-used displays such as overlapping use of listening space and temporal competition.

The conduct of the experiment also presented a related opportunity to measure the total time required for radar blip assessments, which was not adequately known at the start of the modeling work and had to be partially inferred [12]. The time course of this process in the EPIC model assumes that blips are acquired by the eyes, assessed in some way, and then responded to. Since the time required for this sequence of actions can be measured directly with an appropriate variation of the radar task that displays blips one at a time, scenarios with and without auditory cues were developed and added to the study.

Finally, state information that bears on a number of the performance questions that were confronted in the modeling work was captured in the study. Among the issues this data will eventually help to empirically evaluate are the radar screen inspection strategy, blip assessments, and the relationship between time given to the current task and the status of the unattended task.

4. METHOD AND APPARATUS

4.1. Setup

During the period in which the baseline condition and the simpler radar cueing manipulation from the 2002 experiment were being modeled, the dual-task software was revised to run natively under the current Macintosh operating system. The software was then further modified to communicate with a new virtual audio server and to record state information that can be used to reconstruct scenarios in future analyses of performance data. A separate software package, run under the Windows operating system, was developed to present the auditory cues and utilize an inertial head tracker. As before, the audio component of the study was rendered binaurally in stereo headphones with the same non-individualized HRTF employed in the earlier study. Two flat panel monitors facing the operator on opposite sides, respectively, at 45° angles, were used to display the visual components of the experiment. The radar task was shown on the left, and blip decision responses were entered on a numeric keypad positioned below the monitor. The tracking task, which shows a rapidly moving aircraft silhouette as seen from behind, was presented on the right, and participants controlled the movement of its circular cursor with a Hall effect joystick.

4.2. Recording Switches of Attention

The critical augmentation in the setup for the new study was the addition of a head tracking system, which is necessary for implementing an AAR display but also allows head orientation data to be logged automatically, in contrast to the manual technique that was used before to track shifts of attention between the two tasks. The hand-held computer used for this purpose in the previous study enabled the experimenter's observations to be time stamped, and this, in addition to providing a both a record of attentional transitions and a measure of task switching effort, allowed cumulative distributions of time-on-task between attention shifts (dwell times) in each condition to be developed for the modeling work (allowing for experimenter errors and a one sec. resolution for manual input).

The right-skewed patterns exhibited in these distributions for both tasks yielded a number of important explanatory insights and were among the key criteria the modeling work aspired to account for. For example, differences between the baseline distribution of tracking task dwell times and the corresponding distribution in the (modeled) auditory display condition revealed that most of the significantly greater number of attention switches participants made to the radar task in the absence of auditory cues were associated with very short episodes of tracking. Since all attention to the radar task in the baseline condition was unprompted, the dominance of short tracking dwells indicates that participants were forced to look at the secondary task early and often to maintain sufficient awareness of its status. As noted above, the model's account for this data predicts that short periods of attention to the tracking task correspond to phases in which relatively high numbers of blips are present on the radar screen. In contrast, the smaller numbers of short tracking task dwells in the sound condition demonstrates that the correspondence of auditory cues with blip color changes affords longer periods of attention to the tracking task by reducing the need to see when blip responses are required.

Gathering attention shift data in the new study by automated means is not expected to refute the insights gained from the previous study's manually collected data on the basis of greater accuracy, but, instead, is expected to provide the means for evaluating the analysis of this earlier data, realized as modeling predictions, and, somewhat less importantly, to provide more objective counts of attention switches and a better temporal resolution of dwell times.

4.3. Experimental Design

Twenty NRL staff members volunteered to participate in the experiment. Of these, two individuals had to be dropped due to anomalous attention switching performance, resulting in a group of 6 women and 12 men, ranging in age from 19 to 49 with a mean of 30. Over the course of two days, participants trained to perform the two tasks separately and together, were familiarized with the sounds used in the study, and then carried out the main experiment, which was composed of four dual-task exercises under different treatments in a single-factor, repeated measures design. Treatments were given to participants in counter-balanced order, and independently of this, each exercise was successively scripted by a different radar task scenario involving 65 blip decision events. After completing the main experiment, participants were given two further exercises involving only an altered version of the radar task. A summary of all of the exercises participants were assigned is given in Table 1.

The four treatments in the main experiment entailed a baseline exercise with no sound, designated as NS below, and three manipulations, respectively designated as NAAR3, AAR3, and AAR1, in which the radar task was augmented by progressively different virtual auditory cueing designs. The first of these, used in NAAR3, was an auditory display with three easily differentiated auditory cues (one for each type of radar blip) that were localized in the same manner as the simpler of the two spatialization schemes used in the 2002 study. As its designation implies, this display was an NAAR listening space, meaning that the correspondence between the radar task and the virtual source of the auditory cues—nominally located forward and 45° to the left in the listener's auditory field—was relative to the direction the listener was facing. NAAR3 replicated the aurally-cued manipulation that was modeled in EPIC. The next auditory manipulation, AAR3, was like NAAR3 in all respects except that it used an AAR listening space. Thus, in this second auditory cueing design, the virtual source of all three sounds appeared to be co-located with the radar task regardless of the orientation of the listener's head. The final manipulation, AAR1, used the study's third auditory cueing design, which, like AAR3, was also an AAR display that used a single virtual sound source co-located with the radar task. However, in this final auditory cueing design, only one sound was used instead of three, and its aural identity was different from the three sounds used in the NAAR3 and AAR3 treatments.

The auditory materials used to augment the radar task are short audio files of warning sounds that are played as sound

a) Main Experiment	
Condition	Description
NS	Baseline dual task exercise with **no sound** (i.e., no auditory cueing was used)
NAAR3	Dual-task exercise with a **non-augmented auditory reality** display using **3** auditory cues to signal radar blip color changes - each blip type signaled by an *identifying* sound - one virtual source for all three sounds - correspondence of radar task to perceived location of sounds is relative to orientation of listener's head
AAR3	Dual-task exercise with an **augmented auditory reality** display using **3** auditory cues to signal radar blip color changes - each blip type signaled by an *identifying* sound - one virtual source for all three sounds - radar task and perceived location of sounds are co-located
AAR1	Dual-task exercise with an **augmented auditory reality** display using **1** auditory cue to signal blip color changes - all three blip types signaled by the *same* sound - one virtual source for all three sounds - radar task and perceived location of sounds are co-located
b) Blip-Assessment-Time Study	
Condition	Description
BA-NS	**Blip assessment** exercise—**no sound**
BA-S	**Blip assessment** exercise with **sound**

Table 1: A summary of a) the four experimental conditions in the main experiment, showing their coded designations, and b) the two additional exercises conducted to measure blip assessment times. All exercises were assigned to participants in counter-balanced order.

loops. Loops start when each event's color assignment is made and end when decisions are entered, but are only sounded one at a time and always correspond to the oldest unacknowledged event whenever overlaps occur. The sounds used in the NAAR3 and AAR3 manipulations are a police siren, an air-raid siren, and a diesel truck horn, and the sound used in the AAR1 manipulation is a low frequency pulse alert. Unspatialized examples of each of the auditory cues are given in the audio files accompanying this paper, which are listed below (these files are also available by email from the first author as .wav or .mp3 files).

[SIREN.WAV]
[AIRRAID.WAV]
[HORN.WAV]
[PULSE.WAV]

The two radar-task-only exercises that followed the main experiment were designed to explicitly measure how much time radar blip assessments take. These exercises were conducted as an ancillary study to develop parameters for future modeling work and are analyzed here as a single factor, repeated measures study with two levels, designated as BA-NS and BA-S. In each exercise, a scripted sequence of 72 individual blips was displayed by a version of the radar task that was altered to present a black screen with a red dot corresponding to the center of the radar display before each moving blip was shown. Participants were asked to focus on the red dot and then look at the displayed blip, assess it as they would in the dual task, and enter their decision on the numeric keypad. In the exercise designated BA-NS, participants assessed blips without auditory cues; the BA-S exercise was augmented by the auditory display used in the AAR3 manipulation in the main experiment. A different script was consistently used for each exercise and the manipulations were assigned to participants in alternating order.

4.4. Data and Planned Analyses

As in the 2002 study, three primary measures of performance were collected in the main experiment: tracking error, radar task response times, and counts of the number of attention shifts participants made during each exercise. Based on the previous findings, a correlated pattern of significant differences among the treatment means for each of these measures was expected to be found. Also, because the auditory design questions the main experiment addresses are progressive, the manipulations were specifically ordered to allow planned orthogonal contrasts to be made. A significance level of .05 is used for all analyses.

Only preliminary progress has been made on the more detailed, secondary analyses that are expected to shed light on model-related questions. These results will be reported at a later date. However, implications of the present analyses for the modeling work are covered below, as well as the measures resulting from the two blip assessment exercises.

5. RESULTS

The treatment means for the primary measures in the main experiment are shown in the plots in Figures 2, 3, and 4 (error bars in all of the plots show the standard error of the mean). A consistent pattern of performance differences is present, and a one-way, repeated measures ANOVA for each measure was significant (see Table 2). As in the 2002 study, tracking error data was normalized to compensate for individual differences by subtracting each participant's mean tracking error in their final tracking-only training exercise from their mean tracking error in each manipulation and dividing these differences by the standard deviation of the tracking-only mean. Radar task response times were measured in ms from the point at which blips first change color to the point at which participants made the second of the two key presses required for decision responses (see Section 2.1). The means for these two measures in the NS and NAAR3 conditions are relatively close to the respective values in the corresponding manipulations in the 2002 study: tracking errors are slightly lower and blip response times are a little over 200 ms higher than their earlier counterparts. The mean number of attention switches in the NS and NAAR3 manipulations, though, at 299.5 and 224.4, are

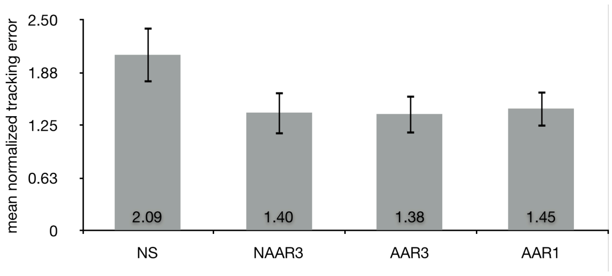


Figure 2: Mean normalized tracking error in the main experiment. The method of normalization is given in the text (Section 5).

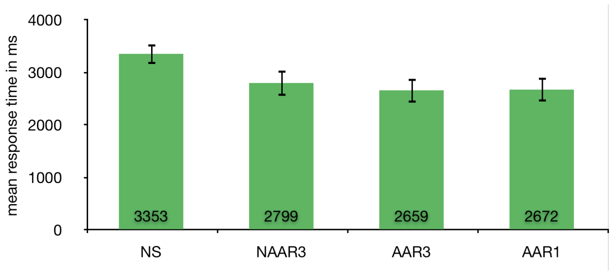


Figure 3: Mean blip response time for the radar task in the main experiment. Measures shown are for the second of the two key presses participants were required to make to record each decision (see Section 2.1).

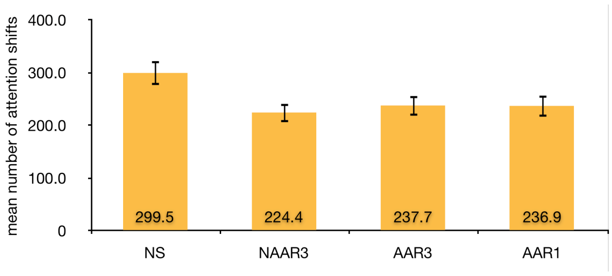


Figure 4: Mean number of attention switches between tasks in the main experiment derived from head tracking data. See the text (Section 5) for additional information about the calculation of these counts.

notably lower than the respective counts of 411.2 and 295.2 that were obtained by hand in the previous experiment, and may, in fact, underreport the number of attention switches participants actually made. The counts published here are a function of the underlying head orientation data collected in each exercise. This measure, which was logged at rate of 20 Hz, proved to be much noisier and subject to individual differences than expected. Although unambiguous shifts from right to left and back again are present in much of the data, many instances where it is unclear whether a genuine change in orientation occurred are also present. To smooth this directional jitter, lower sample rates and a series of distance thresholds were methodically explored. A sample rate of 4 Hz in combination with five thresholds ranging in even steps from 0.02 to 0.1 radians (1.15 to 5.73 degrees) resulted in a stable series of progressively decreasing counts in each of the four treatments. The numbers reported here correspond to the largest threshold and are the most conservative set of the group. However, any of the other thresholds could have been reported with no impact on the significance of the main effect for this measure. Although the empirical counts in Figure 4 potentially challenge the targets for this measure in the modeling work, the ratio of NS to NAAR3, at 1.33, (as well as this ratio for the lower thresholds described above) is quite close to the corresponding ratio of 1.39 in the earlier study.

a) Normalized Tracking Error	
Comparison	Test
main effect	$F(3, 51) = 6.9, p < .001$*
NS with (NAAR3+AAR3+AAR1)/3	$F(1, 17) = 21.1, p < .001$*
NAAR3 with (AAR3+AAR1)/2	$F(1, 17) = 0.006, p > .05$
AAR3 with AAR1	$F(1, 17) = 0.43, p > .05$
b) Blip Response Time	
Comparison	Test
main effect	$F(3, 51) = 10.14, p < .001$*
NS with (NAAR3+AAR3+AAR1)/3	$F(1, 17) = 17.62, p < .001$*
NAAR3 with (AAR3+AAR1)/2	$F(1, 17) = 1.07, p > .05$
AAR3 with AAR1	$F(1, 17) = 0.029, p > .05$
c) Attention Shifts	
Comparison	Test
main effect	$F(3, 51) = 12.36, p < .001$*
NS with (NAAR3+AAR3+AAR1)/3	$F(1, 17) = 20.38, p < .001$*
NAAR3 with (AAR3+AAR1)/2	$F(1, 17) = 1.81, p > .05$
AAR3 with AAR1	$F(1, 17) = 0.006, p > .05$

Table 2: Summary of statistical analyses of the primary performance measures in the main experiment: a) normalized tracking error, b) blip response time, and c) number of attention shifts between tasks. Tests marked with an asterisk are significant.

Planned comparisons among the means for each of the primary measures are also shown in Table 2. These linear contrasts progressively compare a) performance in the baseline condition to the mean of the three auditory display conditions, b) performance in the NAAR design to the mean of the two AAR designs, and last, c) performance with three auditory cues to performance with just one. The first of these is significant for all three measures, and thus provides clear evidence that the auditory treatments meaningfully helped participants carry out the competing tasks at the same time. None of the contrasts comparing the three auditory display designs amongst themselves reached significance, though, and this is an important result that will be considered in greater detail below.

The two blip-assessment exercises with a modified version of the radar task that followed the main experiment yielded a substantial amount of information that will be useful for additional modeling refinements. The scripts for these exercises required participants to decide whether blips from all three of the type categories were hostile or neutral, both before and after they changed color; instances of each of the color assignments (which have not been covered in this paper) were also included, thus giving a balanced set of measures for the different configurations of visual information participants dealt with in the main study. Although comparisons of these breakdowns are not presented here, the means for both treatments, BA-NS and BA-S, are shown in Figure 5. The times shown are for the first of the two key presses participants made for each decision. This measure affords the most straightforward way to use performance constraints to infer the amount of time an operator spends in the overall assessment procedure gazing at a blip to encode its criterial information (see Section 3). Specifically, the time required to acquire each blip visually and the time required to execute the appropriate first key press can be calculated on the basis of standard results in the human performance literature. These intermediate values, which "frame" the core measure of interest, can then be deducted from the gross measure to extrapolate the time spent studying the blip.

Finally, it is interesting to note that while the small, 2.6 percent difference between these means is not significant, $F(1, 17) = 1.09$, $p > .05$, it is nevertheless in the direction that is typically seen when auditory cues accompany visual information. The difference is slightly larger in the same direction, at 3.8 percent, for the mean of the second key presses in these exercises, which are 2476 and 2382, respectively. These latter numbers are essentially measures of distraction-free responses, so it is useful to compare them with the mean blip response times shown in Figure 3 as a way of understanding the impact of the operational paradigm on decision making. In the absence of auditory cues, the presence of an additional task (i.e., tracking) and the distance between the task displays adds 877 ms (nearly a second) to decision responses. And even with auditory cueing, a difference of 277 ms with the measure in the AAR3 condition (the type of display used for the BA-S treatment and the lowest in the study) is still present.

6. DISCUSSION AND CONCLUSION

Two important purposes were met in the design and implementation of this study. The first was to revisit the manipulations from the 2002 dual task experiment that were modeled within the framework of the EPIC cognitive architecture, with the intent of examining specific attentional performance predictions and inferred parameters that came to light in this work. The outcome of this goal was a replication of the main finding of the earlier study, namely, that virtual auditory cueing can meaningfully improve the performance of widely separated concurrent tasks, in large part by significantly reducing the degree of attention switching (taken to be a measure of effort) that is needed to maintain adequate awareness of both tasks. The logging of state information, not gathered in the 2002 experiment, which can be used to reconstruct the status of the dual task at key points, allowing questions about the relationship between courses of action and specific situational patterns to be studied, is expected to be a useful asset for further explanatory and predictive modeling of concurrent tasks involving visual and auditory information.

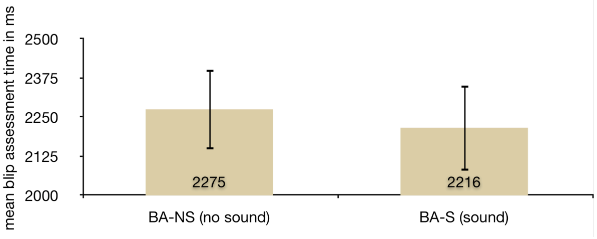


Figure 5: Mean response times from the two blip assessment exercises with a modified version of the radar task that followed the main experiment. Measures shown are for the first of the two key presses participants were required to make to record each decision (see Section 2.1 for the radar task response procedure).

The other major purpose served by the experiment was the methodical investigation of progressively different auditory display designs involving elements that are thought to be important for the composition and use of much richer auditory information displays than the relatively straightforward, single-purpose application that was evaluated here. While the lack of meaningful differences among these treatments may seem puzzling, it is nevertheless a valuable and encouraging result.

None of the prior series of dual-task studies involving auditory cues have utilized an AAR listening space. Yet it seems unlikely that any NAAR design could have reliable utility in real-world settings in which operators must regularly interact with multiple team members, turn to face large, team-oriented displays, and maintain a general awareness of a complex information environment that is likely to include public uses of sound. Disparate uses of virtual audio that had to be mapped to more than one task would potentially invite confusion unless operators were required to remain perpetually oriented toward their workstations. In principle, however, AAR would directly address this concern, particularly in the context of a mixed-use auditory display, by allowing auditory cues and other sound information to be virtually co-located with, and so inherently draw attention to, the different tasks they correspond to, regardless of where the listener might be looking.

On the basis of this reasoning, it is unlikely that the AAR3 treatment would have been in some way inferior to the NAAR3 treatment, and the fact that performance in both AAR treatments was effectively no different than in the NAAR3 manipulation can be taken as persuasive evidence that this is indeed the case. But this finding does suggest that virtual aural overlays on visual information are probably unnecessary—though certainly not disadvantageous—in relatively simple, single-use auditory display applications (e.g., the radar task in the present study), especially when the pace of the environment requires the operator to maintain a high degree of situation awareness and remain oriented toward the performance context. More to the point, it is entirely likely that adding any form spatial information to auditory cues is unnecessary in visually circumscribed, single-purpose applications because operators can readily intuit the import of the sounds.

In an indirect, but principled way, support for this last assertion is arguably provided by the contrasts between the AAR3 and AAR1 treatments, which show, for the purpose of executing the dual task, that the removal of aural blip identity information had no meaningful impact on performance, that is, one auditory cue was as good as three. This outcome implies that simpler auditory information designs for visual attention can be in some cases as good as, or even preferable to, more information-laden designs, which, in turn, may be a particularly useful finding for the design of mixed use auditory displays.

The principle of use that unifies these two outcomes is the notion of least aural effort that was proposed in Section 3, which asserts that, on the whole, listeners only make use of the information functions present in a particular instance of sound that are the most effective for the purpose at hand. (c.f. the "principle of least effort" in [16]). The evidence from the contrasts of auditory treatments in the study is that, beyond the onset function of the auditory cues, listeners were indifferent to the manipulation of two kinds of additional task-relevant information: identity and locational deixis. The most plausible explanation for this indifference is that participants were able to more efficiently gather and process these essential pieces of information for performing the radar task from other sources, one being cognition (where is the task?) and the other being the visual display (what must be decided?). This is not to say that the augmentary aural information could not have been used, only that it appears to have been superfluous in the specific context of the dual task as employed here.

With only a secondary task requiring intermittent attention and all of the criterial information for blip assessments readily available to the eyes, the dual task presents little or no opportunity for listeners to make timely use of the two categories of auditory information that were manipulated in this study. But this circumstance is unlikely to hold where mixed-uses of auditory cues are required. In Navy operations, watchstanders already attend to opposing chat and tactical situation displays and are subject to documented lapses of attention [17]. Virtual auditory cueing is being studied as a strategy for ameliorating this concern, and it is difficult to argue that performance in the absence of aural identities and deixis for these and other tasks in this type of setting will serve the operator well, precisely because these functions index a specific task among several. Additional aural elaboration, though, may be unneeded or counterproductive unless it can be exploited more readily than other sources of task-relevant information.

7. ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research under work request N0001410WX20448.

8. REFERENCES

[1] D. Brock, J. A. Ballas, J. L. Stroup, and B. McClimens, "The design of mixed-use, virtual auditory displays: Recent findings with a dual-task paradigm," in *Proc. of the 10th Int. Conf. on Auditory Display (ICAD)*, Sydney, Australia, July 6-9, 2004.

[2] D. Kieras and D. Meyer, "An overview of the EPIC architecture for cognition and performance with application to human-computer interaction," in *Human Computer Interaction*, 12, 391-438, 1997.

[3] H. Fouad, J. A. Ballas, and D. Brock, "An extensible toolkit for creating virtual sonic environments," in *Proc. of the 6th Int. Conf. on Auditory Display (ICAD)*, Atlanta, USA, April 2-5, 2004.

[4] W. W. Gaver, "Using and creating auditory icons," in *Proc. of the 1st Int. Conf. on Auditory Display (ICAD)*, Santa Fe, USA, October, 1992.

[5] J. A. Ballas, "Delivery of information through sound," in *Proc. of the 1st Int. Conf. on Auditory Display (ICAD)*, Santa Fe, USA, October, 1992.

[6] N. A. Stanton and J. Edworthy, "Auditory warning affordances," in N. A. Stanton and J. Edworthy (Eds.), *Human Factors in Auditory Warnings*, Ashgate, Aldershot, UK, 1999.

[7] General Dynamics Advanced Information Systems, "Common display system," at http://www.gd-ais.com/index.cfm?acronym=cds

[8] G. A. Osga, "Human-centered shipboard systems and operations," in H. R. Booher (Ed.), *Handbook of Human Systems Integration*, Wiley, Hoboken, USA, 2003.

[9] S.C. Peres, V. Best, D. Brock, C. Frauenberger, T. Hermann, J. Neuhoff, L.V. Nickerson, B. Shinn-Cunningham, and A. Stockman, "Auditory interfaces," in P. Kortum (Ed.), *HCI Beyond the GUI.* Morgan Kaufman, San Francisco, USA, 2008.

[10] D. Brock, J.L. Stroup, and J.A. Ballas, "Effects of 3D auditory cueing on dual task performance in a simulated multiscreen watchstation environment," in *Proc. of the Human Factors and Ergonomics Soc. 46th Ann. Meeting*, Baltimore, MD, 2002.

[11] J.A. Ballas, D. Kieras, D. Meyer, D. Brock, and J.L. Stroup, "Cueing of display objects by 3-D audio to reduce automation deficit," in *Proc. of the 4th Ann. Symp. and Exhibition on Situational Awareness in the Tactical Air Environment.*, Patuxent River, MD: Warfare Center Aircraft Division, 1999.

[12] D. Brock, B. McClimens, A. Hornof, and T. Halvorson, "Cognitive models of the effect of audio cuing on attentional shifts in a complex multimodal dual-display dual-task," in *Proc. 28th Ann. Meeting of the Cognitive Science Soc.*, 2006.

[13] A.J. Hornof, Y. Zhang, and T. Halverson, "Knowing where and when to look in a time-critical multimodal dual task, to appear in *ACM CHI 2010: Conference on Human Factors in Computing Systems,* New York: ACM, 2010.

[14] A.J. Hornof, T. Halverson, A. Issacson, and E. Brown, "Transforming object locations on a 2D visual display into cued locations in 3D auditory space," in *Proc. of the 52nd Ann. Meeting of the Human Factors and Ergonomics Soc.,* 2008, pp. 1170-1174.

[15] D. Brock, B. McClimens, J.G. Trafton, M. McCurry, and D. Perzanowski, "Evaluating listeners' attention to and comprehension of spatialized concurrent and serial talkers at normal and a synthetically faster rate of speech," in *Proc. of the 14th Int. Conf. on Auditory Display (ICAD).* Paris, France, June 24-27, 2008.

[16] H.H. Clark, *Using language*, Cambridge University Press, New York, 1996.

[17] J. Cantanzaro, M. Risser, J. Gwynne and D. Manes, *Facilitating critical event detection in chat communications*, (Technical Report). San Diego, CA: Pacific Science & Engineering Group, Inc., 2006.

Sonification of NRL Dual-Task Data

Brian McClimens

Naval Research Laboratory
4555 Overlook Ave, Washington DC
brian.mcclimens@nrl.navy.mil

Derek Brock

Naval Research Laboratory
4555 Overlook Ave, Washington DC
derek.brock@nrl.navy.mil

1. INTRODUCTION

The addition of a head-tracker to the dual-task experiment reported in Brock et al., 2010 led to two distinct sonifications of the collected data [1]. Initial analysis of the head-tracking data was attempted in Microsoft Excel, and the size of the data sets made the visual comparison of multiple log files intractable. The head-tracker logged data approximately every 70 ms, and each condition was thirteen minutes long, so a graph of a single condition contained over eleven thousand points. This proved too large to be managed from within Excel. In exploring possible programs to view the data in we realized that our log files could easily be modified to conform to the .wav file format.

Initially, the realization that the head-tracking data was essentially a digitally sampled analogue signal was used only as a means transform the data into a form that could be visually examined from within sound editing software. Sound Forge^{TM} allowed us to simultaneously view the data from multiple subjects. Although the data had been changed to a .wav format, it was not specifically intended as a sonification in the sense that no thought had been put into how the data would sound. Nevertheless, the format used happened to produce meaningful sound and with very little practice, several colleagues could easily differentiate between sonifications of conditions which had been shown to have statistically significant differences in the head movements of subjects.

This spurred the construction of a more detailed sonification of our data which included not only the data about the motion of subjects' heads, but also data collected about their tracking performance and a detailed event log from the radar task. The goal was to sonify as much of the collected data as possible, and then to see whether or not people could differentiate between data from each of the different conditions used in the experiment.

2. Data and Sonification Methods

The NRL Dual-Task consists of a tracking task and a radar task. The tracking task is a continuous task in which subjects are required to follow the movements of a target on the screen with a joystick. The radar task involves a series of classification events in which the subject responds to the behavior of blips on the screen by classifying them as either hostile or neutral. Each classification event requires two key presses, one to denote the determination of hostility, and one to identify the blip. Each condition had a duration of thirteen minutes, with a total of 65 classification events. A detailed description of the task can be found in Brock et al., 2010[1]. Two sonifications have been created from the data collected in this experiment. The first sonification only made use of the head-tracking data, but a second sonification used performance data from the tracking and radar tasks.

2.1. First Sonification

Subjects wore a pair of headphones fitted with a head-tracking device that logged the direction the subject was facing approximately every 70 milliseconds. This logging was slightly irregular, with some samples separated by as much as 100 milliseconds. In order to most effectively display this data as a .wav file, evenly spaced samples are needed. Because the magnitude of head movements between any two points in our dataset was small, we were able to use a simple linear interpolation between points to create a dataset with points spaced exactly 50 milliseconds apart. These data points were measured in radians and ranged from approximately -0.3 to .3 radians.

The initial sonification used the interpolated, evenly spaced data, and converted it to an audio format in the most direct possible way. Each data point was scaled and converted to a 16-bit integer between -32677 and 32678. These were then written as sample points in a 16-bit, 8kHz PCM format. Because each sample represented 50 milliseconds of time in our experiment, and 8kHz audio stream represented a playback speed four hundred times faster than real-time, and each thirteen minute condition was translated into a 1.95 second sound file.

As examples, eight auditory files are included. The first four are sonifications from a single subject of the four experimental conditions described in Brock et al., 2010 [1]. The conditions are referred to as the “no sound” condition (NS), the “non-augmented auditory reality” with three sounds (NAAR-3), the “augmented auditory reality” with three sounds (AAR-3), and the “augmented auditory reality” with one sound (AAR-1). For details about the difference of the conditions, please refer to the referenced paper. The key issue for the purpose of this example is that in the NS condition, subjects performed significantly worse than in others, and had a greater number of head turns.

The first sound clip (Son1_1-1.wav), is a sonification of subject 1's head movement in the NS condition. When compared to the other three files (Son1_1-2.wav, Son1_1-3.wav, and Son1_1-4.wav), subtle differences are apparent. Although each condition sounds like low-frequency static, the 1-1 file is noticeably smoother. In the other conditions (NAAR-3, AAR-3, and AAR-1), subjects performed better, with fewer unnecessary head turns. In these sonifications, the relatively large gaps of time with no head movement producs short silences followed by a noise sounding like a pop or a click. With a small amount of practice, we found ourselves able to correctly identify sonfications of the NS condition. In order to test this for yourself, four additional auditory clips are provided (Son1_2-a.wav, Son1_2-b.wav, Son1_2-c.wav, Son1_2-d.wav). These four files are the Sonifications of another subject's head movement in the NS, NAAR-3, AAR-3, and AAR-1conditions. There are distinct differences between subjects, but you should be able to identify the sonification of the NS condition. The conditions represented by the four files are revealed in a text file (Son1_key.txt).

3. Second Sonification

A second sonification was created with the goal of including all data that had been collected during the experiment. In addition to the head-tracking data, this sonification would include data about the subjects' performance on both the tracking task and the radar task, as well as data describing the state of the radar task.

3.1. Head-tracker

The second sonification made use of the interpolated data points used in the first sonification described above. Instead of directly mapping those data points to samples, this sonification used simple properties of the data set to modulate a square wave. The value of the data point was used to position the square wave such that the left-right panning corresponded to the direction that the subjects' head was facing. In addition, the rate of head movement as measured by the magnitude of change between neighboring data points was used to modulate the frequency and amplitude of the square wave. In order to make movement seem continuous, all values are interpolated from the two nearest data points. The goal was to have a relatively subtle sound that would indicate the subject's current focus to a listener, and modify it so that head movements would be noticeable with larger head movements more salient than small ones.

3.2. Joystick

In the tracking task, a subject is required to keep a reticle over a target by controlling the reticle with a joystick. The log file for this task records the position of the target, position of the reticle, and the position of the physical joystick on both the x-axis and y-axis. Each of these values is recorded every 83 milliseconds (12 Hz). From this log, interpolated points are calculated every 50 milliseconds in a method equivalent to what was used for the head-tracking data. From those data points, we calculated the distance between the target and reticle at 50 millisecond intervals. White noise was then synthesized, with the distance between target and reticle used to modulate amplitude so that the noise is quiet for periods where the subject performs well, and increases in volume as a subject's performance decreases.

4. Radar data

In the radar task, subjects see blips of three different types traveling from the top of the screen to the bottom. They are required to classify these incoming blips as hostile or neutral based upon rules that differ for each blip type, and must enter their designation via the keyboard. Theses responses are not allowed until the blips are about halfway down the screen. Prior to this point the blips are grey, and when a response is allowed, they change color. In most conditions of the experiment, an alarm is played at the same time as the color change, but in one condition, no alarms are played. The primary metric used to evaluate subject performance on the radar task is the reaction time measured from when a blip changes color to when the subject completes a designation entry.

The log files for the radar task contain several types of information. Each time a blip appears on the radar screen, becomes active or is classified either incorrectly or correctly by the subject, an event is logged. In addition, every keystroke performed by the subject is recorded, and categorized. Each of the events is encoded and represented within the sonification as short sound clips. These are not synthesized or modulated in any way, but are simply a mapping of the possible event types used to trigger the playback of preset audio clips. Finally, a 440 Hz sine wave was used to indicate periods of time in which the task was waiting for a response from the subject. For all time periods in which a colored blip was present on the screen, a sine wave is generated. In this way, a sonification in which the sine wave was more prevalent reflected poorer performance by the subject.

5. Auditory Examples

Two 16-bit, 44.1kHz .wav file examples of this sonification are included (1-1_20-mix.wav, 1-2_20-mix.wav). The algorithm used to sonify the NRL Dual Task data has a parameter for temporal compression. If the temporal compression is set to 1, then the resulting sonification will be equal in length to the data provided. A temporal compression of 400 would result in a file with the same duration as the first sonification. The examples provided are presented with a temporal compression rate of 20.

Though we hoped to create sonifications that would comfortably include all of the different types of data from the experiment, I find that the sonifications are easier to understand when they are split into data from the head-tracker, joystick data and radar data.

5.1.1. Head-tracker audio

The first data file (HT_1.wav) is a sonification in which the data is being presented in real time. This sonification is a short segment of data from subject 1's head motion during the NS condition. When presented with no temporal compression in the sonification algorithm, it is relatively easy to keep track of the head's position, and to get a sense of the speed with which the head was moved for any individual head turn. Unfortunately, it is very difficult to compare two of these sonifications and notice differences in a subject's performance or behavior based upon them. Increasing the compression improves the ability to compare between sonifications, while giving a less clear picture of head positioning. As an example, two sound clips of sonifications with a temporal compression rate of 20 are included (HT_20_1-1.wav, HT_20_1-2.wav), along with two examples at a compression rate of 400 (HT_400_1-1.wav, HT_400_1-2.wav). These files represent the head movements of subject 1 in the NS and AAR-3 conditions at two different speeds. The files with a compression rate of 400 contain the entire data set for their conditions. At a compression rate of 20, individual head turns are difficult, but possible to perceive. However, it is also difficult to distinguish between the two conditions. The files with a compression rate of 400 make it impossible to identify individual motions, but a different property emerges which makes it easy to distinguish between conditions. At the higher compression rate, the average position of a subject's head can be heard more clearly, with the subject's head position falling far right in conditions other than NS due to the smaller number of head turns.

5.1.2. Joystick audio

Of all the portions of the sonification, the joystick data changed least with different temporal compression rates. The joystick sonification is a direct measure of a subject's performance on the joystick task but in most cases, the differences between conditions in the joystick files are not apparent. Included are the four sonifications of subject 9's performance in each condition (JS_9-1.wav, JS_9-2.wav, JS_9-3.wav, JS_9-4.wav). These files have a compression rate of 400, and represent the entire data set for subject 9.

5.1.3. Radar audio

The main quality that stands out as a difference between conditions in the sonification of the radar data is the presence or absence of the sine wave indicating a colored blip on the screen. This closely related to the subject's reaction time since the completion of a response (whether correct or incorrect) will remove a blip from the screen and results in less time with a colored blip onscreen. Again, the differences between conditions are more pronounced when the compression rate is higher. Another indication of condition type is the presence of the (Miss.wav) sound. Most misses occurred in the NS condition. The number of occurrences is small though, and some subjects were able to complete the NS condition without missing any blips, while others made a few errors of omission in other conditions as well. Two radar files at compression rates of 400 are included for comparison (R_400_1-1.wav, R_400_1-2.wav)

6. Discussion

In the process of sonifying the NRL Dual-Task data, we found it possible to highlight all of the statistically significant features reported on in Brock et al., 2010. We had hoped that these sonifications would provide some insight that might compliment our data analysis. Unfortunately, no distinguishing qualities between conditions have yet been noted in the sonifications that was not present in our statistical analysis. One difficulty with exploring the data was that parameters' settings had to be chosen in advance, and each time we wanted to change a parameter, the program needed to be recompiled, and the sonifications regenerated. This process was quite time consuming, especially for the sonifications with lower temporal compression rates. In order to explore this data more effectively, a dynamic interaction between the listener and the sonification may help. A sonification that could be generated in real-time, with controls to change parameters might allow for a more thorough exploration, and may enable a listener to discover features of these data sets that have not yet been identified.

7. REFERENCES

[1] Brock, D., McClimens, B., and McCurry, M. “Virtual auditory cueing revisited,” In *Proceedings of the 16th International Conference on Auditory Display*. Washington, DC, June 9-15, 2010.

POSTER

PRELIMINARY STEPS IN SONIFYING WEB LOG DATA

Mark Ballora

School of Music/Department of Integrative Arts
The Pennsylvania State University
30 Borland Building, University Park, PA 16803
ballora@psu.edu

Brian Panulla

The Pennsylvania State University
College of Information Sciences and Technology
Center for Network Centric Cognition and
Information Fusion
brian@panulla.com

Matthew Gourley

The Pennsylvania State University
Administrative Information Services
mmg207@psu.edu

David L. Hall

The Pennsylvania State University
College of Information Sciences and Technology
Center for Network Centric Cognition and
Information Fusion
dhall@ist.psu.edu

ABSTRACT

Detection of intrusions is a continuing problem in network security. Due to the large volumes of data recorded in Web server logs, analysis is typically forensic, taking place only after a problem has occurred. We are exploring the detection of intrusion signatures and patterns via an auditory display. Web log data is parsed and formatted using Python, then read as a data array by the synthesis language SuperCollider [1], which renders it as a sonification. This can be done either for the study of pre-existing data sets or in monitoring Web traffic in real time. Components rendered aurally include IP address, geographical information, and server Return Codes. Users can interact with the data, speeding or slowing the speed of representation (for pre-existing data sets) or "mixing" sound components to optimize intelligibility for tracking suspicious activity.

1. INTRODUCTION

While the primary task of the sciences may be exploration and the discovery of new knowledge, a critical issue currently facing scientists and researchers is in the area of presentation -- the ability to introduce their discoveries effectively, both to laypeople and to fellow researchers. There is an emerging area of interest in representation of scientific information, and how the use of multi-media technologies, an essential component in disseminating new information, can in turn shape and influence scientific thought [2].

The problem is not only that of dealing with new forms of information, but also with unprecedented quantities of it. In our Information Age, new forms of gathering information are constantly being created. However, this does not necessarily lead to increased understanding. In particular, managing crisis situations or monitoring infrastructures requires the ability to interpret incoming information from multiple sources. With new sources of information constantly becoming available, the challenge becomes how to process it effectively, avoiding the condition described by informatics researchers as *cogmenutia fragmentosa* [3].

Penn State's Center for Network Centric Cognition and Information Fusion (NC^{2}IF) [4] housed in Penn State's College of Information Sciences and Technology, explores the information chain from energy detection via sensors and human observation to modeling, signal and image processing, pattern recognition, knowledge creation, information infrastructure, and human decision-making [5].

The Center's Extreme Events Lab (EEL) is intended to allow researchers to run end-to-end experiments that improve situational awareness and enhance their ability to optimally leverage all available sensors, human observers, and technology in order to escape "information overload" and extract the true meaning hidden within the vast mountains of available data [6].

There is a body of work in the field of cyber security that examines intrusion detection in terms of information theory, noting that the complexity of network activity drops during intrusion attempts [7, 8].

Here we describe initial steps in an experiment created for the EEL in which Web log data is rendered as a sonification. Our goal is to determine whether intrusion attempts produce recognizable patterns that can be detected aurally, either in real time, or as an after-the-fact analysis. Results of this work will become part of a collective pool of methodologies used in ongoing data rendering experiments carried out by the center.

2. PREDECESSORS

This project is related to earlier work [9-11] involving network activity sonifications. Most directly related was the Peep Network Auralizer System [9], which played back various recordings to reflect network conditions such as incoming and outgoing email, load average, number of users logged in, etc. Through various sound libraries, akin to SoundFonts, listeners could be placed in a variety of listening environments – rainforest, desert, and so on. The amount of rain might represent load average, the flow of a waterfall might represent email traffic, and so on. The creation of nature-inspired soundscapes was an effective design choice, as it resulted in a pleasant and unobtrusive listening environment. However, the limitation of this approach is the same as found in sampling synthesizer instruments: simple *triggering* of audio files lacks "control intimacy" (as described by Moore in [12]) whereby variations in the data create variations (often subtle) in the creation of sound, in a manner akin to a musician's many microscopic gestures that affect the sound and character of an instrument during performance.

A closer level of control intimacy is achieved with *parameter-based* sonifications [13], which link the data to the sound at a deeper acoustic level than is possible with simple triggering. The data values are mapped to synthesized sound characteristics such as oscillator frequency, filter cutoff frequency, volume, stereo panning, etc. This methodology runs the risk of turning into "bleep bloop" music that can be a trying listening experience. The key to success lies in effective orchestration strategies. A multi-dimensional data set is sonified as a multi-instrumental synthesizer ensemble. The design challenge is to create timbres that complement each other well when they are combined to represent data dimensions. Parameter-based sonifications also have the potential limitation of being arbitrarily contrived, so that users may have difficulty learning which auditory characteristics reflect which data dimensions, which may have no apparent intrinsic connection.

A further level of control intimacy is gained with another approach, termed *model-based* sonification [14]. This involves mapping data values to resonances and/or mechanics of a *physical model*, typically in a form that can be explored interactively by the analyst, creating inextricable relationship between the data and the resulting sound event. A physical model is a computer synthesis technique based on wave equations that describe vibrating objects. An example model-based sonification might be a multi-dimensional data set mapped to a theoretical grid of masses and springs, simulating a virtual instrument. As the data iterates, the character of the instrument's vibrations changes. This allows the possibility of subtle patterns to emerge within the quality of the sound field that would be lost with realization based on simple triggering, and are more inherently integrated than parameter-based sonification methodologies.

The modeling idea is explored in a simplified form in our sonification, in that we use a simple physical model, although without the high-dimensionality and user navigability of many model-based sonifications. We create additional renderings that are parameter based for a more qualitative realization, as well as triggered sounds that are used for certain alerts. The synthesis parameters were chosen for aesthetic reasons, in some cases to create a pleasing sounding musical instrument, in others to create a pleasant nature-like soundscape.

3. WEB LOG DATA

When someone tries to load a Web page by typing a URL, clicking a link, or by submitting data via a Web-based form, that person's browser sends an HTTP Request to the Web server, which in turn sends back an HTTP Response in the form of a Return Code. The Return Codes consist of a numeric ID, often followed by a text explanation. The ranges of the numeric IDs indicate various levels of Success (2xx), Redirection (3xx), Client Error (4xx), or Server Error (5xx). This exchange is typically invisible to users, although one commonly seen Return Code is the familiar *404: Page Not Found*.

The data set used for our sonification consists of 11,350 entries, which originate from a filtered set of HTTP Requests made to the Web server at Penn State's College of Information Sciences and Technology. The requests span a 24-hour time period. Each point in our data set is an array consisting of information taken from an HTTP Request as well as the corresponding Return. Each array entry includes:

- a timestamp,
- the Request's source IP address,
- the latitude and longitude of the Request,
- the Return Code sent by the server.

4. SONIFICATION STRATEGIES

4.1. Iteration

The data set is loaded into SuperCollider as an Array. A Task process is run, with each iteration loading values from the next array member into variables. Synth objects are then instantiated that use the variables as controls of its various aspects (frequency, modulation rate, and so on).

4.2. Time Stamps

Each data array triggers a sound event in the sonification. The rendering for sound events is based on the relative times between timestamps, multiplied by a scalar. Thus, periods of higher or lower relative activity can be easily recognized, depending on whether one hears sparse, occasional events or a flurry of sound. A pre-existing data set consisting of many hours of activity can be heard over a timescale on the order of minutes or seconds, depending on the iteration rate the listener chooses.

4.3. IP Addresses

The principal focus of this work involves translating the IP addresses of HTTP Requests into sound. Our question is whether a coordinated set of Requests from a single address or a set of related addresses would create an auditory signature of some kind.

IP addresses are typically written in "dot-decimal notation," which consists of four octet values derived from a 32-bit network identification number, as shown in Table 1.

a 32-bit binary identification number

01100100100101101100100011111010

is divided into four 8-bit octets

01100100.10010110.11001000.11111010

commonly written with each octet as a
decimal number in the range of 0-255

100.150.200.250
most significant octet least significant octet

Table 1. IP Address Structuring

We treat the octets separately, thus treating each source IP address as a four-dimensional data point, with the most significant values represented by the leftmost octet, and the least significant values represented by the rightmost.

The binary nature of IP addresses suggests a compatibility with the numbering system used in MIDI (Musical Instrument Digital Interface), a common software protocol understood and transmitted by synthesizer instruments. A note number of 60 is assigned to middle C, with successive values above or below 60 corresponding to half steps above or below middle C.

As a preliminary step, the IP octet values are mapped to MIDI note values. Each octet (0-255) is remapped to a value within a span of 23 half steps. Since the initial range of 0-255 is much larger than the mapped range of 0-23, the resulting values are floating point, meaning that most of the mapped frequencies are microtones, falling between the half-steps represented MIDI integer values.

4.4. Vibraphone

A set of four resonances can describe the timbre of a vibraphone, as shown by the spectrogram in Figure 1. It can be observed that the timbre consists of four resonances that closely correspond to the fundamental, fourth, 10th and 17th harmonics.

We start with the quartet of MIDI note values derived from the source IP address, described in the last section, and transpose them further, such that they function as four harmonics falling roughly within the ranges of a vibraphone's partials. These values are then used in an instantiation of an vibraphone-like instrument that is created with SuperCollider's Klank unit generator, which is a simple and general physical model consisting of an arbitrary number of resonant frequencies with relative amplitudes and ring times:

```
SynthDef("ipVib", {arg fund=293,
        formant1=1173, formant2=2930, formant3=4986,
        v1=1, v2=1, v3=0.3, v4=0.3, pos=0.0;
e=Env.new([0, 1, 1, 0], [0.01, 1.5, 0.01], 'linear');
k=DynKlank.ar(`[[fund, formant1, formant2, formant3],
                [v1, v2, v3, v4],
                [1.5, 1.0, 0.25, 0.1]],
                Impulse.ar(0, 0, 0.1));
p=PanAz.ar(~numChans, k, pos, 1, 3);
Out.ar(0, p*EnvGen.ar(e, doneAction:2))
         }).send(s);
```

Each time the Task iterates, a single impulse (the digital audio equivalent of a percussive strike) is sent to an instantiation of the vibraphone instrument, with four resonances mapped from the values of data's IP address. Thus, each HTTP Request in the data set triggers an instance of the vibraphone-like instrument in the sonification.

The result is a quick, active vibraphone melody. The four octet/frequency values are not heard as a discrete chord, but rather they fuse into a unified timbre with shifting resonances. A short two-channel excerpt can be heard online at http://dl.dropbox.com/u/4128606/vibraphone-like.wav.

4.5. Babbling Brook

The Peep system, mentioned earlier, created a pleasant and informative nature-scape. In an effort to appropriate this idea, an alternate rendering is created of each IP address that is meant to sound like a brook or creek.

This sound model is derived from a synthesis example titled

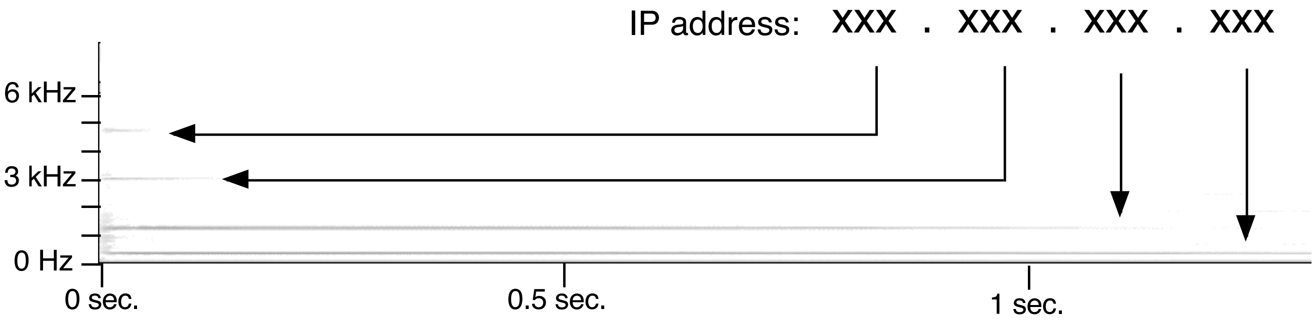


Figure 1. Spectrogram of vibraphone Middle D (293 Hz), with the assignment of IP octets to the most significant partials.

"Babbling Brook," which was created by SuperCollider's inventor, James McCartney, and is included as an example that comes with the SuperCollider program's documentation. The patch consists of filtered noise, with a modulating value controlling the cutoff frequency of the filter.

In our adaptation, each source IP address provides parameters for four instances of the water-like instrument, each with its average cutoff frequency and modulation rate based on the value of its source octet.

```
SynthDef("ipnoisedroplet",
    { arg f1=800, f2=17, vol=1.0, dur=1.0, pos=0.0;
      e=Env.new([0, 1, 1, 0], [0.01, dur, 0.01],
       'linear', nil, nil);
      r=PanAz.ar(~numChans,
       RHPF.ar(OnePole.ar(BrownNoise.ar, 0.99),
       LPF.ar(BrownNoise.ar, f2+5)*f1+f1, 0.03, 0.003),
       pos, vol, 3);
      Out.ar(0, EnvGen.ar(e, doneAction:2)*r);
   }).send(s);
```

The result is a diffuse water-like soundscape that is abstractly related to the IP addresses in the dataset. Since there is no inherently musical relationship between the octet values in IP addresses, the option of creating the babbling brook soundscape allows the possibility of a more qualitative rendering that may be preferable to the musical relationship somewhat imposed by the vibraphone rendering. A short two-channel excerpt can be heard online at http://dl.dropbox.com/u/4128606/water-like.wav.

4.6. Vocal Synthesis

Since a set of 3-5 formants (consistent spectral peaks applied to a signal containing many harmonics) is sufficient for basic vowel synthesis, a third rendering is created that models vowels created from each source IP address. This is somewhat similar to the idea of the vibraphone-based timbre, with the distinction that the four values are mapped to different frequency regions so that they fall within the ranges of vocal formants instead of vibraphone harmonics. The four frequency values are also not given relative ring times, but rather they all sound continuously. A continuous sound that modulates as the data set iterates creates a vocal-like drone that has a shifting vowel quality.

4.7. Server Return Codes

In addition to creating sounds mapped from source IP addresses, an additional annotation is given to each data point by creating a sound based on the Web server's Return Code. Different percussive sounds are assigned to the various Return Codes, such as a model of two river stones clicking together, or noise bursts to suggest splashing, or a pitched ringing sound. These are meant to highlight occurrences of different Return messages, so that a repeated error message, for example, can be made salient.

4.8. Request Location

Since SuperCollider allows sounds to be panned over an arbitrary number of channels, we also sonify longitude of each Request as stereo localization, panning the sound event within an octaphonic ring of loudspeakers, placing the listener in "the center of the world." At this writing, a rendering of latitude is being explored. A likely design will be a model resembling a gong, so that different "elevations" can be represented by differing strike pressures, so that a low tapping can represent lower latitudes, with a louder, ringing strike indicating higher latitudes.

5. INTERFACE

The overall soundscape of the sonification is controllable by a mixing-board like interface. Its functionality follows the design of our earlier work in cardiac data sonification [15]. Users can start, pause, and reset playback via buttons; one slider can speed up or slow the rate at which the dataset is traversed; another slider allows the dataset to be scrubbed so that playback may start from any arbitrary point, with the timestamp of the current position displayed visually. Each sound component has separate volume slider, allowing the overall mix to be controlled.

The interface is displayed on a Lemur LCD controller [16]. This customizable device can send messages to the IP addresses of musical devices via Open Sound Control (OSC). A Lemur can potentially control a synthesizer device placed anywhere on the World Wide Web. Lemur interfaces are created in software on a computer, with objects taken from a palette of knobs, sliders, and other interface elements, and each named individually.

SuperCollider has a class called Lemur, by which an interface can be recognized by its IP address. It can read information from the named objects in the Lemur interface (for example, the position of an object called Slider1), and assign it to a variable within the synthesis patch (for example, a variable assigned to an oscillator's volume value).

6. INITIAL RESPONSES

Development and testing of the sonification system is being carried out as this is being written. But it is apparent that two base conditions have been met. One is that coarse changes are audibly obvious: for example, repeated Requests from a single location are readily apparent, even to the least musically trained ears. Another is that of persistence: the sound quality makes a pleasing and unobtrusive backdrop. This was informally assessed during a wine and cheese event commemorating the tenth year of the College of Information Sciences and Technology. Visitors were invited to visit the EEL and other facilities. As people wandered in throughout the evening, the initial reaction of visitors was quite favorable. Granted, this soft anecdotal evidence hardly constitutes academic merit in and of itself. However, a pleasant listening experience is an essential component of a successful sonification, and is therefore an essential criterion for evaluating merit at this initial stage of the work.

7. FUTURE WORK

The work to date has been on design issues, exploring how the information can be mapped effectively. The next step is to establish proof of concept by evaluating our design with data sets containing known intrusion attempts. Well-documented case sets are publicly available [17], which allow for initial proof of sonification concept.

It is likely that other characteristics of a Web log data set can be usefully sonified, beyond simple renderings of the requesting IP address. Subsequent iterations of this work will explore anomalous log entries, such as unusually long requests, which are often associated with attempted database intrusions.

Larger-scale analyses will likely need some higher levels of abstraction in rendering, as the density of data may become unwieldy or incoherent when each point is sonified. While having this microscopic level present is desirable to ensure the integrity of the rendering, it is also desirable to be able to introduce various types of averaged and statistical data.

We also project creating a real-time renderer. This will likely be in the form of a daemon written in Python that receives copies of Web log entries, parses them, and reformats them as OSC messages, which it then sends to SuperCollider.

8. CODA

It is unlikely that Web system administrators will feel inclined to purchase octaphonic sound systems and Lemur interfaces. However, there has been interest expressed in using some form of auditory monitoring of server traffic at Penn State, and certainly scaled-down versions could easily be created for practical implementations. But it is more to the point to bear in mind that NC^{2}IF functions as a collective, whereby ideas are regularly shared among people working on a variety of projects. It is entirely possible that an interesting idea created for one project may turn out in practice to be more suitable for another project, as various forms of data renderings and data fusion are explored. It is thus in our interests to explore all possibilities for representation, rather than potentially limit ourselves by setting out to create a fixed product for this rendering.

9. REFERENCES

[1] http://supercollider.sourceforge.net/

[2] *Visualizing Science: Image-Making in the Constitution of Scientific Knowledge (an interdisciplinary symposium)*. October 24, 2007, Brandeis University. http://culturalproduction.wikispaces.com/visualizing_science

[3] McNeese, M. D. and Vidulich, editors, *Cognitive Systems Engineering in Military Aviation Environments: Avoiding Cogmenutia Fragmentosa,* Dayton, Ohio, Wright Patterson Air Force Base, CSERIAC Press, 2002.

[4] http://nc2if.psu.edu/

[5] Hall, D., C. Hall, S. McMullen, M. McMullen, and B. Pursel, "Perspectives on Visualization and Virtual World Technologies for Multi-Sensor Data Fusion," *Proceedings of the 11th International Conference on Information Fusion,* Cologne, Germany, June 30- July 03, 2008.

[6] Hall, D., B. Hellar, and M. D. McNeese, "The Extreme Events Laboratory: A Cyber Infrastructure for Performing Experiments to Quantify the Effectiveness of Human-Centered Information Fusion." *Proceedings of the 2009 International Conference on Information Fusion (Fusion 2009),* Seattle, Washington, July, 2009.

[7] Evans, S.C. and B. Barnett. "Network Security through Conservation of Complexity." *Proceedings of MILCOM 2002 Military Communications Conference.* October 7-10, 2002, Anaheim, California, IEEE.

[8] Eiland, E. E., and L.M. Liebrock, "An Application of Information Theory to Intrusion Detection." *Proceedings of the Fourth IEEE International Workshop on Information Assurance.* IEEE Computer Society, 2006.

[9] Gilfix, M. and A. Couch, "Peep (The Network Auralizer): Monitoring Your Network with Sound." In *2000 LISA XIV.* December 3-8, 2000 – New Orleans, LA, pp. 109-117.

[10] Chafe, C., and R. Leisikow, "Levels of Temporal Resolution in Sonification of Network Performance," Proc. 2001 Intl. Conference on Auditory Display, Helsinki, 2001.

[11] Chafe, C., and S. Wilson, D. Walling, "Physical Model Synthesis with Application to Internet Acoustics," Proc. 2002 Intl. Conference on Acoustics, Speech and Signal Processing, Orlando, 2002.

[12] Moore, F.R. *Elements of Computer Music*. Englewood Cliffs, NJ: PTR Prentice Hall, 1990.

[13] Kramer, G., ed. *Auditory Display: Sonification, Audification, and Auditory Interfaces*. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XVIII. Reading, MA: Addison Wesley, 1994.

[14] Hermann, T. *Sonification for Exploratory Data Analysis.* Ph.D dissertation, Bielefeld University, 2002.

[15] Ballora, M. and B. Pennycook, P. Ch. Ivanov, L. Glass, A. L. Goldberger. 2004. "Heart rate sonification: A new approach to medical diagnosis." *LEONARDO* 37 (Feb. 2004): pp. 41-46.

[16] http://www.jazzmutant.com/lemur_overview.php

[17] Lippmann, R. P. [et al.]. "Evaluating Intrusion Detection Systems: The 1998 DARPA Off-Line Intrusion Detection Evaluation." *Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX), Vol. 2. 2000*: IEEE Computer Society Press: Los Alamitos, CA. p. 12-26.

IDENTIFYING AND COMMUNICATING 2D SHAPES USING AUDITORY FEEDBACK

Javier Sanchez

Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University
The Knoll, 660 Lomita Dr.
Stanford, CA 94305, USA
jsanchez@ccrma.stanford.edu

ABSTRACT

This research project shows a technique for allowing a user to "see" a 2D shape without any visual feedback. The user gestures with any universal pointing tool, as a mouse, a pen tablet, or the touch screen of a mobile device, and receives auditory feedback. This allows the user to experiment and eventually learn enough of the shape to effectively trace it out in 2D. The proposed system is based on the idea of relating spatial representations to sound, which allows the user to have a sound perception of a 2D shape. The shapes are predefined and the user has no access to any visual information. While exploring the space using the pointer device, sound is generated, which pitch and intensity varies according to some given strategies. 2D shapes can be identified and easily followed with the pointer tool, using the sound as only reference.

1. INTRODUCTION

The aim of this research project is to use sound as feedback with the aim of recognizing shapes and gestures. The proposed system has been designed with the idea of relating spatial representations to sound, which is as a way of sonification.

Sonification can be defined as the use of nonspeech audio to communicate information [6]. Basically, our proposal consists on relating some parameters of the 2D shape that we want to communicate, with some sound parameters as pitch, amplitude, timbre or tempo between others. By nature, sonification is an interdisciplinary field, which integrates concepts from human perception, acoustics, design, arts, and engineering.

The best-known example of sonification is the Geiger counter, invented by Hans Geiger in the early 1900's. This device generates a "beep" in response to non-visible radiation levels, alerting the user of the degree of danger. Frequency and intensity vary according to the existing radiation level, guiding the user.

Another example of sonification is given by the Pulse-oximeter, which was introduced as medical equipment in the mid-1980's. This device uses a similar concept that the Geiger counter. It outputs a tone, which varies in frequency depending on the level of oxygen in the patient blood.

Other known example of sonification is the Acoustic Parking System (APS) used for parking assistance in many cars. It uses sensors to measure the distance to nearby objects, emitting an intermittent warning tone inside the vehicle to indicate the driver how far the car is from an obstacle.

Sonification has been used to develop navigation systems for visually impaired people [8] allowing them to travel through familiar and unfamiliar environments without the assistance of guides.

Other works [2],[11] are focused on creating multimodal interfaces to help blind and impaired people to explore and navigate on the web. The design of auditory user interfaces to create non-visual representations of graphical user interfaces has been also an important research activity [1], [9].

Some systems have been developed to present geographic information to blind people [5], [7], [10]. It allows the user to explore spatial information.

In some works the aural feedback is added to an existing haptic force feedback interface to create a multimodal rendering system [3], [4].

Although our system would be used to assist visually impaired people in the recognition of shapes and gestures, we do not want to limit its scope to this field of application.

2. SYSTEM DESCRIPTION

In this section is described our proposal, that consists on using auditory feedback to help users in the identification and communication of 2D shapes in those situations where they have no access to any visual feedback.

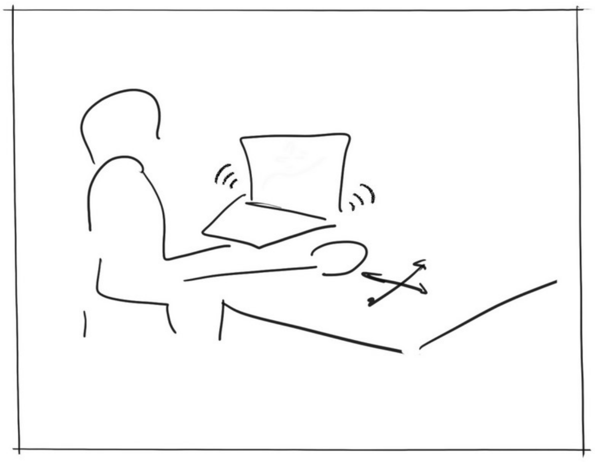


Figure 1. Using an universal pointer device to interact with the system.

Although the system would be conceived as a stand-alone product, the first prototype is designed as a piece of software that runs in any computer.

As the idea of the proposed system is to communicate a 2D shape to other users using auditory feedback, the first thing that has been implemented is a simple drawing interface to generate a 2D shape.

Once the 2D shape has been created or imported into the system, the system is ready to communicate the shape to the user. This communication is made possible by emitting some sounds while the user gestures using a universal pointer device as a mouse, a pen tablet, a pen display or a touch screen of a mobile device. This has been an important design specification of the system, which allows the user to interact with the system using any universal pointer device.

Figure 1 shows how the user interacts with the system using a pointer device. Although the user is sitting in front of a computer, it must be clearly stated again that the user has no access to any visual information.

In order to identify the 2D shape, the user should start exploring the space around him by moving the pointing device. The movement of the user pointer tool is directly associated with the movement of a virtual point in a virtual 2D space where the shape is located.

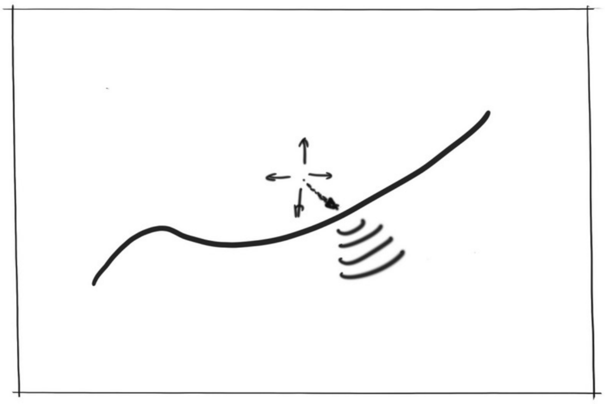


Figure 2. The user has not access to any visual information. A sound is generated when the user approaches to the shape.

As the user approaches to the shape, a sound is generated which pitch, timbre and intensity can vary according to a specific spatial to sound mapping strategy. Figure 2 shows how sounds are generated when the user approaches to the shape.

Once the user has located the 2D shape, the following step consists on trying to follow the shape using the sound as only feedback. If the user moves away from the curve, the sound disappears and the user can get lost into the silence. Anyway, the user can easily move the pointer back to the last position where the sound appeared, to continue tracking the position of the shape.

The size of the user workspace while moving the pointer matches with the size of the screen where the shape is located. When the user moves the pointer further the limits of the workspace, a different sound tells him that he has reached the workspace limits. This is very useful when using the mouse as pointer device.

The proposed system is based on the proprioception sense, which provides a relation between the gestures made by the user while following the sound, with the spatial representation of these gestures.

Thanks to the proprioception sense, the hand gesture made while following the sound is transformed into a spatial representation of the shape. Figure 3 shows how the user can reconstruct mentally the 2D shape using the auditory feedback.

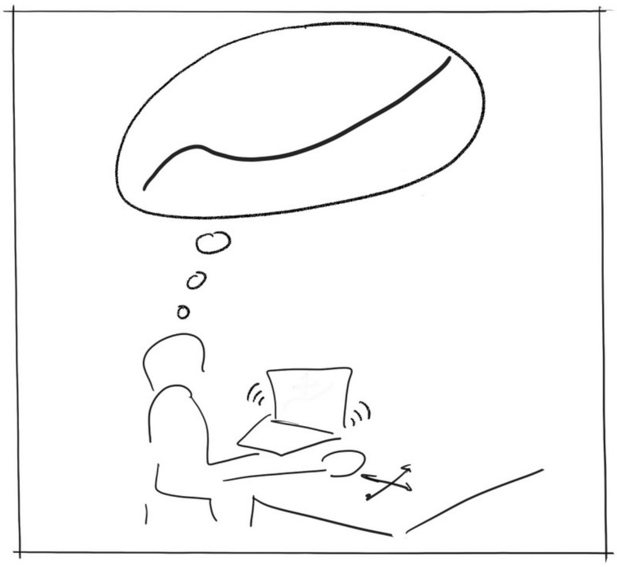


Figure 3. Users transform the gesture made while following the sound, into a spatial representation of the shape.

There are several ways of identifying the 2D shape using the pointer device. Some users would prefer to keep following the 2D shape slowly without loosing the sound. On the other hand, other users would prefer to start moving around the whole workspace, from side to side, getting some scattered points, which can be later connected mentally to form the 2D shape (see Figure 4).

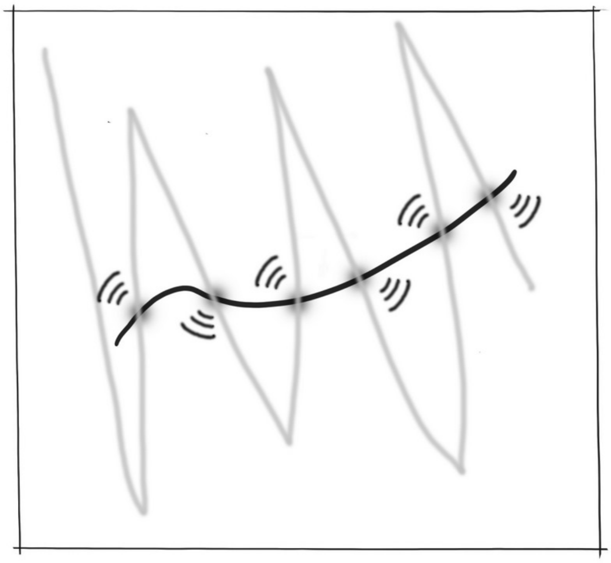


Figure 4. User movements from side to side of the screen trying to find a 2D shape using the sound as feedback

In order to provide a relation between the gesture made and the sound feedback, a perfect synchronization of perceived audio events with expected tactile sensations is needed.

The user workspace is divided into two different areas: *sound* areas and *no sound* areas. Figure 5 shows how the limits between sound and silence are located at certain distances at both sides of the 2D shape. The transition between silence and sound is made gradually, as shown in figure 5 where the sound intensity increases as the distance to the curve decreases.

The value given to this distance is not trivial and its appropriate selection will ensure that the user will be able to identify adequately the 2D shape using auditory feedback. If the distance were greater than needed, the sound area would be too wide. This would imply the possibility of finding multiple solutions, which would be far from the 2D shape that the user was trying to identify.

In the other hand, if the distance were too close to the 2D shape, it would be difficult for the user to locate a 2D shape, due to its thin thickness. The 2D shape would become invisible.

The value of this distance depends also on the pointer device used. For example, it is not the same to use the small track pad of the laptop that using a 15’’ pen tablet. In this example, the ratio between the size of the finger and the track pad area is much bigger that the ratio between the stylus diameter and the area of the 15’’ pen tablet. The value of this distance is also related to the resolution of the pointer device.

So, further studies should be carried out to find the optimum distance that delimits the sound area around the 2D shape.

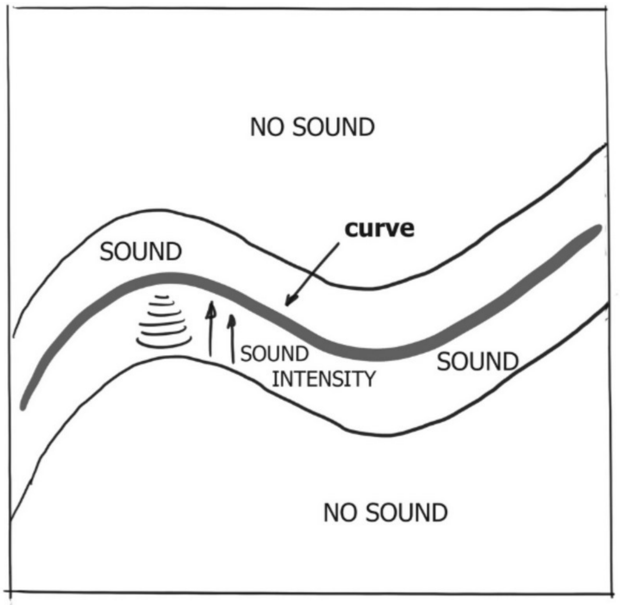


Figure 5. Sound to spatial relationship. Sound intensity increases as the distance to the curve shape decreases.

When working with pointer devices, it is necessary to be aware of the differences between relative and absolute referencing. In our system, it is much better to work with absolute references. Most of the pen displays and touch screens use absolute references. It is not the same case when dealing with a mouse or a track pad, where the referencing system is relative.

As example, if the user lifts the mouse, moves it away, and places it again on the surface, the pointer stays in the same position on the screen. This is not useful for our system, since the user would lose the spatial reference while trying to locate a 2D shape.

On the other hand, if the user uses a pen tablet, the whole area of the tablet is mapped to the whole area of the screen. So, if the user lifts the stylus, moves it away and places it again on the surface, the pointer moves to another position on the screen. This is exactly what we need.

3. STRATEGIES TO MAP GEOMETRY TO SOUND

An application has been built with the aim of studying how easy would be for a user to identify a 2D shape using the sound as feedback. Some parameters will be set to adjust the process. In this section are given some technical details and strategies used to develop the application.

As stated in the previous section, the sound intensity increases as the distance from the pointer to the 2D shape decreases. In addition to this, some parameters of the 2D shape, as position, slope or curvature, are used to enrich the sound information given to the user.

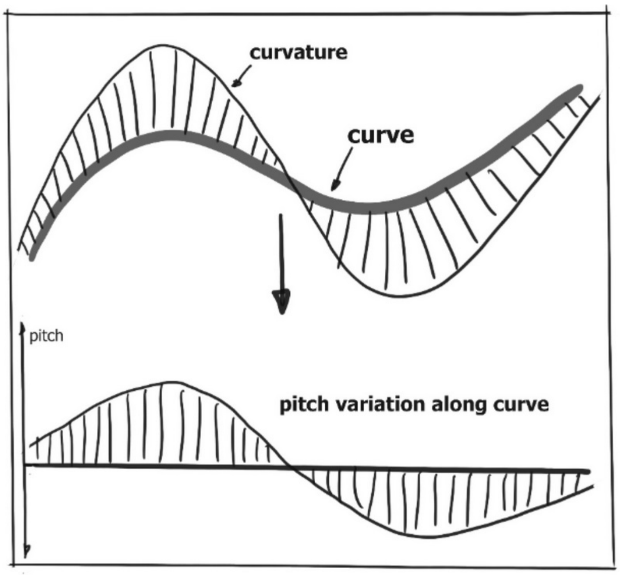


Figure 6. Sound to spatial relationship. Some properties of the 2D shape as slope or curvature are associated with the sound parameters to enrich the sound feedback.

As example, a pitch variation of the sound feedback would tell the user about the curvature of the shape at each point. So, a possible strategy would consist on varying the sound pitch along the 2D shape, according to the curvature at each point of the shape.

According to this, a straight line will generate a constant pitch. The curve represented in figure 6 has a variable curvature, so the user will have different pitch perceptions while moving along the shape.

Another useful strategy would be to use the slope of the 2D shape at each point to generate different sound pitches along the shape. Depending on the shape, it would be more appropriate to use one or other strategy.

So, the position of the pointer together with some geometric properties of the 2D shape would help to enrich the sound information given to the user.

There are other sound parameters that would be used to enhance the auditory feedback. As example, the duration of the sound can be related to the thickness of the 2D shape. This strategy would allow the users to make a difference between shapes with different thickness. It would be even possible to identify changes in thickness within the same shape, using the sound as feedback.

Depending on the pointer device used, it would be more convenient to relate the thickness of the shape to the loudness of the sound generated.

What about adding some effects to the original sound to express other variations that would appear on the geometry? We would distort the original sound using some filter, as a reverbs or an echo, to relate the new sound to the style of the pencil used to draw the 2D shape. Other parameters of the 2D shape as the transparency or the applied pressure while creating the stroke would be associated with some distortion of the generated sound.

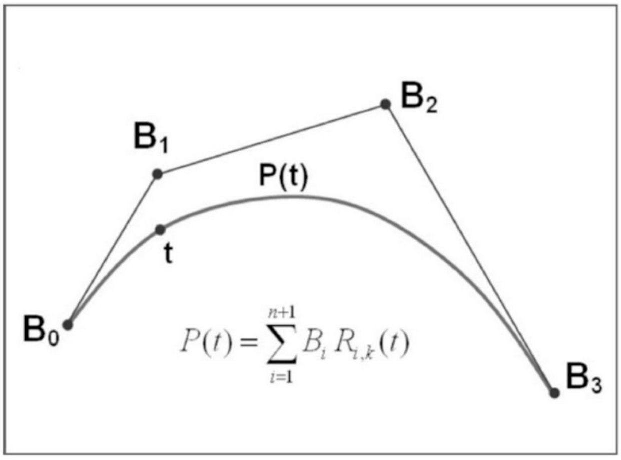


Figure 7. Parametric curves are used to define shapes.

Color is another property that would be associated with some sound property. We can start thinking in a system with 8 basic colors, which are associated with 8 different sound timbres. This relation has been established since timbre is considered as the color of music. Both terms timbre and color are used indistinctly traditionally to represent the sound quality.

Other possibility that have been included in the system is to represent a 2D closed shape. Imagine that the user is trying to follow a 2D rectangular shape. We can use the same previous strategies to identify the edges of the rectangle, relating them to a specific sound, and add a new sound to the area that is contained inside the rectangle. This strategy will enrich the sound feedback and will help the user to identify a shape.

Some primitive shapes as, circles, ovals, rectangles, triangles, etc, can have a secondary sound associated to the shape, which would indicate the user that he is trying to identify one of these singular shapes. This secondary sound doesn't need to be always active; it can appear slightly every few seconds to avoid excessive noise in the scene.

The idea of including several channels at the same time to express several shape properties would really facilitate to the user the identification of 2D shapes and enrich the sound feedback. Other sound parameters that would be used in the proposed system can be panoramization effects, changes in tempo and rhythm, or fade-in and fade-out transitions between sounds.

Auditory feedback should not be reduced only to sound. Music, voice or noise would be also used in the proposed system. A voice can be mapped to a linear shape and be triggered depending on the position of the pointer along the shape. The user would control the voice or some music back and forward at the desired speed as if controlling a music player. Following a music score would be also associated with the movement of the pointer device.

Special care should be taken with the selection of the generated sound. Using the same kind of sounds can be hard and tedious for the user, or even painful, depending of the ranges of pitches used. A library of sounds can be included to allow users to choose their own sounds. Random sound selection is another option. Ambient sound can be used to fill the background and some atmosphere sounds can be associated with the internal area of closed shapes. Textures can be associated with some noise added to the original sounds.

The 2D shapes are represented by means of parametric curves, which are a standard in 2D drawing representation. Since Drawing Exchange Format (DXF) is used to store the graphic information, it is very easy to generate curve shapes using any commercial CAD application and import them into our system. Figure 7 shows an example of a parametric curve.

Multiple curve shapes can be defined into the same scenario using different sound for each curve (see figure 8). Distances to the curves are evaluated as the user interacts with the model. Including too many entities in the same scene can be not the best idea, especially if using the track pad or the mouse as pointer device. A bigger workspace would be needed. A pen tablet or a pen display are preferred when working with multiple shapes.

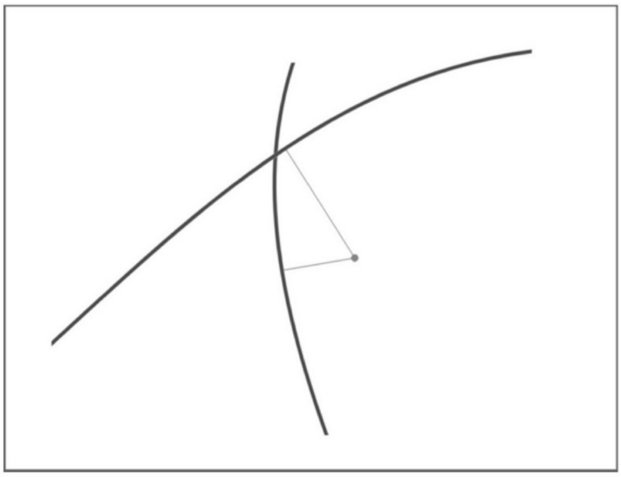


Figure 8. Multiple shapes are associated with different sounds.

4. SYSTEM IMPLEMENTATION

The analysis of the user motion, the curve representation and the output sound has been computed using MAX/MSP, a visual programming environment specifically designed to simplify the creation of acoustic and control the application.

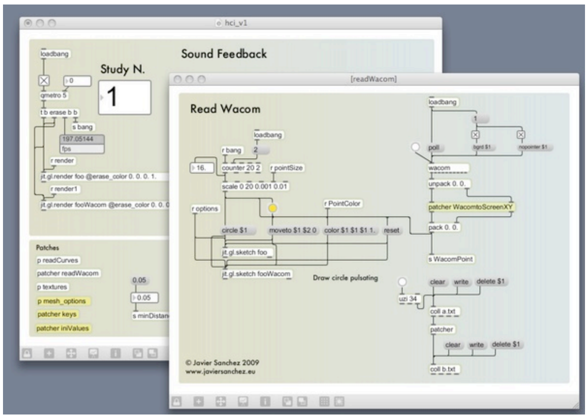
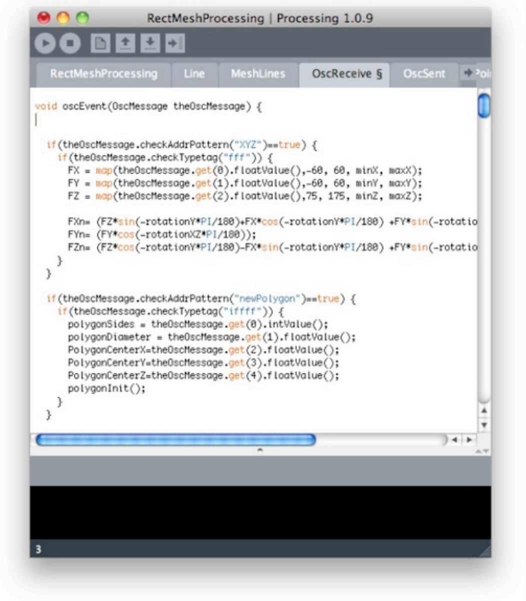


Figure 9. MAX/MSP is an excellent programming environment to test a prototype system, adjust sound parameters or communicate with any universal device.

Controlling external devices as the mouse, a pen display, an iPad or and iPhone is very easy to do using MAX/MSP. The visual programming environment facilitates the control of the process and the communication with other systems. Figure 9 shows a MAX/MSP snapshot.

The Processing programming environment has been chosen for building the visuals of the application (see Figure 10). Processing is an open source programming language and environment to work with images, animation, and interactions. It is also an ideal tool for prototyping.



```
void oscEvent(OscMessage theOscMessage) {

  if(theOscMessage.checkAddrPattern("XYZ")==true) {
    if(theOscMessage.checkTypetag("fff")) {
      FX = map(theOscMessage.get(0).floatValue(),-60, 60, minX, maxX);
      FY = map(theOscMessage.get(1).floatValue(),-60, 60, minY, maxY);
      FZ = map(theOscMessage.get(2).floatValue(),75, 175, minZ, maxZ);

      FXn= (FZ*sin(-rotationY*PI/180)+FX*cos(-rotationY*PI/180) +FY*sin(-rotatio
      FYn= (FY*cos(-rotationXZ*PI/180));
      FZn= (FZ*cos(-rotationY*PI/180)-FX*sin(-rotationY*PI/180) +FY*sin(-rotatio
    }
  }

  if(theOscMessage.checkAddrPattern("newPolygon")==true) {
    if(theOscMessage.checkTypetag("iffff")) {
      polygonSides = theOscMessage.get(0).intValue();
      polygonDiameter = theOscMessage.get(1).floatValue();
      PolygonCenterX=theOscMessage.get(2).floatValue();
      PolygonCenterY=theOscMessage.get(3).floatValue();
      PolygonCenterZ=theOscMessage.get(4).floatValue();
      polygonInit();
    }
  }
```

Figure 10. Processing is the programming environment used to control the application visuals.

The connection between MAX/MSP and Processing is made using the OSC (Open Sound Control) protocol, which bring the benefits of using modern networking technologies. It provides also everything needed for real time control of sound and other media.

Some other devices as the iPhone or the iPad Touch can be used as pointer devices. The OSC protocol can be used to communicate the mobile device with MAX/MSP using the wireless network. The TouchOSC application [12] has been used to connect the iPhone with MAX/MSP.

Figure 11 shows the appearance of the implemented application. As the idea of the system is to recognize shapes using the sound as feedback, the first step consists on drawing something on the screen. A schematic shape of a car has been represented using 5 lines: one for the external profile, two for the wheels, one for the door and another one for the bottom line. This drawing can be drawn by another user or can be loaded from a collection of drawings stored in the computer.

Once the drawing is completed, the next step consists on recognizing the shapes using the sound as feedback. It is evident that the user has no access to any visual information. As the user moves the pointer device, some lines appear on the screen, which represent the shortest distance from the pointer device to the drawing lines. These lines are updated as the user navigates around the screen.

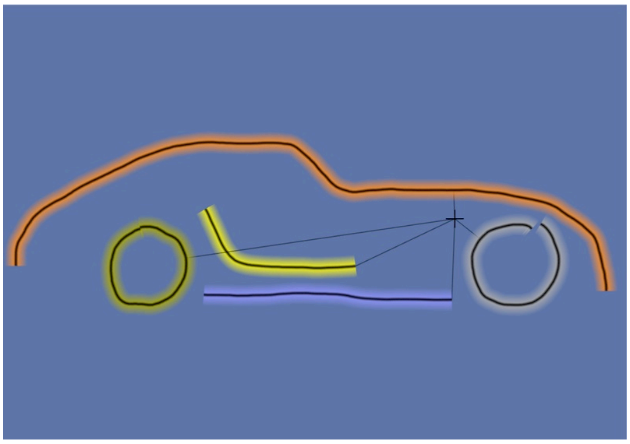


Figure 11. Snapshot of the implemented application, showing a schematic shape of a car, which is recognized by the user using the sound as feedback.

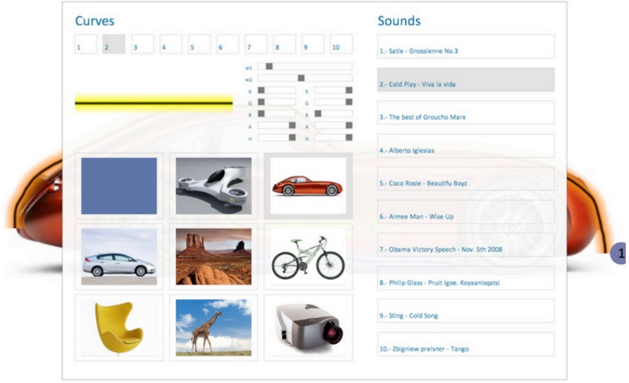
When the user approaches to any of the lines, a sound appears. This sound is related to the geometry by means of some mapping strategies, which are described in the previous section.

A new mapping strategy consists on the use of music as auditory display, instead of sound. The reason of this is that it is much more comfortable for the user to use his own library of music, that synthesized sound. Each curve can be related to a different music theme of the user library. So, when the user approaches to a line on the screen, a specific music theme is played.

In Figure 11 can be shown how each curve is made of two different sub-curves: a thin black curve inside and a ticker colored curve outside. These two different curves are associated with two different audio channels: music and white noise. Let's explain this. When the user approaches to the curve and the pointer is touching the colored area, a white noise appears, which tells the user that he is approaching to the curve. As the user moves closer to the black curve, the white noise disappears gradually, and the music appears clearly. When the user moves away of the thin black line, the music disappears gradually, and the white noise appears again.

The metaphor used in this system is based on the idea of tuning a radio. When the user approaches to a radio station, a clear sound appears. The white noise is telling the user to move the dial until he gets the desired radio station. So, our system can be seen as 2D radio tuner. The user can navigate in the 2D space identifying the curves and following them using the music and the white noise as feedback.

Figure 12 shows a control panel in which the user can associate each curve with a specific theme from his music library. The color and the width of each of the two sub-curves associated with each curve can be also adjusted easily from this control panel.



1.- Satie - Gnossienne No.3
2.- Cold Play - Viva la vida
3.- The best of Groucho Marx
4.- Alberto Iglesias
5.- Coco Rosie - Beautifu Boyz
6.- Aimee Man - Wise Up
7.- Obama Victory Speech - Nov. 5th 2008
8.- Philip Glass - Pruit Igoe. Koyaanisqatsi
9.- Sting - Cold Song
10.- Zbigniew preisner - Tango

Figure 12. Control panel of the implemented system.

A background can be used as reference to trace easily the curves of the model. Figure 13 shows how a picture of a car has been used to sketch the five curves of the model. Users can have their own library of pictures to be used as background.

Finally, it is important to emphasize that each line has an identifier and can be edited or deleted if desired.

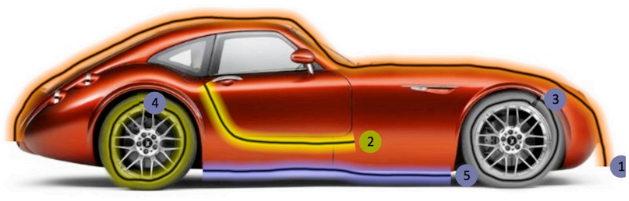


Figure 13. Users can use their own picture library as background to trace the curves of the model.

5. CONCLUSIONS

This paper proposes a novel method that consists in the use of auditory feedback to identify a 2D shape while the user gestures using a pointer device.

Several universal pointer devices, as a mouse, a pen tablet or a mobile device can be used to interact with the system, facilitating the human computer interaction.

Parametric curves are used, as they are a standard in 2D drawing representation. Some of the curve parameters, as slope, curvature or position, are related to the sound output, helping the user to identify the 2D shape.

Other parameters of the 2D shape as color, thickness can be associated to different timbres or loudness. Multiple sound channels can be included to add extra information of the background or to identify some closed areas.

Multiple 2D shapes can be defined in the same scenario using different sounds for each shape.

As it occurs in any interaction device, the user needs certain time to become familiar and confident with the new environment. Users can become skilled in a short time since the application is very intuitive and easy to use.

Current work is related with the use computer vision techniques to track the hand movement of the user. By means of this, the user can interact directly with the system, using the webcam of the computer.

It is also being evaluated the possibility of using the system as an extension (add-on) of some existing computer application.

Other applications are also been studied in which the sound can be related to a gesture to assist the user in common tasks.

The overall low cost of the system and its easy implementation is also an important point in favor.

A collection of applications based on the idea of using sound as feedback has been implemented for the new iPad. Applications for visually impaired people and collaborative games are the most important.

6. REFERENCES

[1] W. Buxton, “Using Our Ears: An Introduction to the Use of Nonspeech Audio Cues” in Extracting meaning from complex data: processing, display, interaction, edited by E.J. Farrel, *Proceedings of the SPIE*, Vol. 1259, SPIE 1990, p. 124-127.

[2] H. Donker, P. Klante, P. Gorny, “The design of auditory user interfaces for blind users” *in Proc. of the second Nordic conference on HCI*, pp. 149-156 (2002)

[3] N.A. Grabowski, K.E. Barner, “Data visualization methods for the blind using force feedback and sonification.” *in Prceedings of the SPIE Conference on Telemanipulator and Telepresence Technologies*, 1998

[4] [IFeelPixel: Haptics & Sonification http://www.ifeelpixel.com/faq/#whatitwill

[5] H. Kamel, J. Landay. “Sketching images eyes-free: a grid-based dynamic drawing tool for the blind.” *In Proc. of ACM SIGCAPH Conference on Assistive Technologies* (ASSETS). pp. 33-40. (2002)

[6] G. Kramer, B. Walker, T. Bonebright , P. Cook, J. Flower, N. Miner, J. Neuhoff “Sonification Report: Status of the

Field and Research Agenda." *In International Community for Auditory Display*, ICAD (1997)

[7] M. Krueger, "KnowWare™: Virtual Reality Maps for Blind People." *SBIR Phase I Final Report, NIH Grant #1 R43 EY11075-01*, (1996)

[8] J.M.Loomis, G. Reginald, L.K. Roberta "Navigation System for the Blind: Auditory Display Modes and Guidance. " *in Presence*,V.7,N.2,193–203(1998)

[9] E. Mynatt, G. Weber. "Nonvisual Presentation of Graphical User Interfaces: Contrasting Two Approaches." *in Proc. of the Computer,* CHI 94. (1994)

[10] P. Parente, G. Bishop "BATS: The Blind Audio Tactile Mapping System." ACMSE. (2003)

[11] W.Yu, R. Kuber, E. Murphy, P. Strain, G.A. McAllister " Novel Multimodal Interface for Improving Visually Impaired People's Web Accessibility." *in Virtual Reality*, Vol 9: 133-148 (2006)

[12] Touch OSC.

http://www.creativeapplications.net/iphone/iosc-iphone/

SPIN QUARTETS.
SONIFICATION OF THE XY MODEL.

Katharina Vogt, Robert Höldrich,
David Pirrò

Institute for Electronic Music and Acoustics,
University of Music and Performing Arts Graz,
Austria
vogt@iem.at, hoeldrich@iem.at,
pirro@iem.at

Christof Gattringer

Institute for Physics,
University of Graz,
Austria
christoph.gattringer@uni-graz.at

ABSTRACT

We present an intuitive sonification of data from a statistical physics model, the XY-spin model. Topological structures (anti-/vortices) are hidden to the eye by random spin movement. The behavior of the vortices changes by crossing a phase transition as a function of the temperature. Our sonification builds on basic acoustic properties of phase modulation. Only interesting structures like anti-/vortices remain heard, while everything else falls silent, without additional computational effort. The researcher interacts with the data by a graphical user interface. The sonification can be extended to any lattice model where locally turbulent structures are embedded in rather laminar fields.

1. INTRODUCTION

In our research, the usefulness of sonification for the display of numerical physics results is being studied.

One interesting model is the so-called *XY-model.* The model is programmed as a lattice of single spins, that may point in any direction in the 2-dimensional XY-plane. It will be explained in more detail below, but in this introduction, we want to point out why it is useful to sonify a 2-dimensional model, that may as well be visualized.

XY-models exhibit a special *topology*, i.e. we may find structures formed by the spins. These are vortices and anti-vortices. A vortex is defined as an arrangement of 4 spins (that we will refer to as spin quartet), which turns around $+2\pi$, if you follow it in counterclockwise direction. The anti-vortex turns by -2π (see Fig. 1). There is always the same amount of vortices and anti-vortices on a lattice with periodic boundary conditions.

We chose to sonify this model for various reasons. With the naked eye, the topological structures are hard to find (Fig. 2(b)). In a method called *cooling*, the overall energy of the configuration is lowered and only the most stable structures stay in an else laminar field (Fig. 2(c)). As a drawback, cooling destroys all topological structures, if applied long enough. And at any step, physical information is lost.

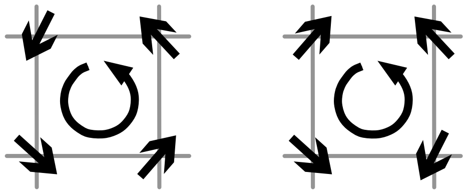


Figure 1: Scheme of an ideal vortex (left) and an anti-vortex (right). If one follows the spins in counterclockwise direction and adds up the angle differences, the vortex turns by $+2\pi$ and the anti-vortex by -2π.

Depending on the temperature as model parameter, the behavior of the vortices and anti-vortices changes. The exact changing point is called the Kosterlitz-Thouless phase transition. This cannot be *seen*, when observing the model, and may only be calculated with the help of nonlocal observables (measures of the whole configuration). A simple nonlocal observable is the vorticity, the number of vortices and anti-vortices in the configuration, but the phase transition can not be deduced by this analysis. Different sound properties at different states - below or above the phase transition - is what we wanted to achieve by utilizing sonification.

Finally it was a very interesting task to find appropriate, easy-to-hear sonifications for a problem that *is* visually displayable. The sonification has to make use of acoustic principles and thus leave visual thinking concepts behind. We believe, that this was achieved in the sonification we present in this paper. It can now also be extended to other models and higher dimensional data, where the hidden structures are still unknown and the data is not visualizable.

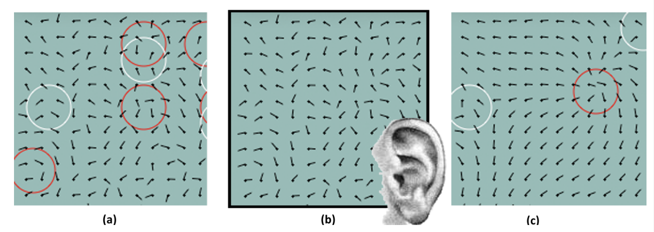


Figure 2: Detail of the XY-model. (a) shows a typical configuration, where also the positions of the vortices and anti-vortices have been calculated and are shown as red and white circles. If only raw data is shown (b), these structures are very hard to find visually. On the right hand side, (c) shows the same detail after several steps of cooling, an algorithm that lowers the overall energy and leaves only the most stable vortices and anti-vortices. In this kind of averaging procedure a lot of physical information is lost. Even the exact position of the structures changed during cooling. In our approach, we work only with the raw data (b). The sonification allows a quick overview over the state the system is in and detailed information on the position of the vortices and anti-vortices *without* calculating them.

2. SONIFICATION IN PHYSICS

In physics, many examples are known where sonification was used before. The most common cited are probably the Geiger counter or the Sonar. Physics is a huge and diversified field, and there are various sonification approaches. therefore an overview of the state of the art cannot be given in this short paper. Some approaches can be found in [1].

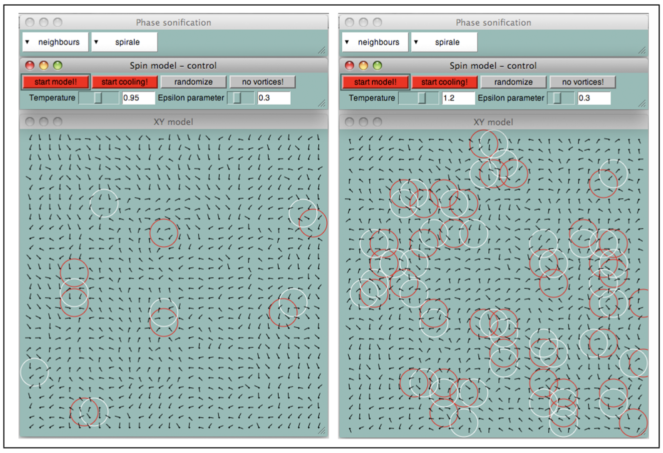
Many projects can be found at the border between science and arts, using sonification of physics data for exploratory and artistic reasons. For instance, algorithmic composition tools are based on physical event generation as the fission model (e.g., [2]) or scientific experiments become music (e.g. the piece *50 Particles*, [3]). In the AlloSphere, a 3-storey high sphere for virtual environments, also theoretical physics data shall be explored [4].

Two research projects ([5], [6]) have studied the sonification of numerical physics models, of which the current paper is a continuation.

3. SPIN MODELS - THE XY-MODEL

3.1. Spin models

Spin models are simple computational models, that use discrete or continuous data on a lattice (a discrete structure). Usually, the lattice forms a (hyper-) torus, i.e. one uses periodic boundary conditions. The degrees of freedom are called spins. In the simplest model, the Ising model, only 2 values are possible, spin up or spin down. A straightforward extension is the XY-model, where the spin is continuous and may point in any direction in the plane.



Temperature 0.95 Epsilon parameter 0.3

Temperature 1.2 Epsilon parameter 0.3

Figure 3: Visualization of typical configurations of the XY model below and above the phase transition (respectively on the left and right hand side). The higher the temperature, the more (red) vortices and (white) anti-vortices can be found. Vortices and anti-vortices usually form pairs, that appear more tightly bound above the phase transition.

The spins are linked to each other through an energy functional, depending on the temperature. The higher the temperature, the more the spins flip randomly due to heat induced fluctuations. The lower the temperature, the more they try to align with their neighbors. Between the low- and high-temperature *phases*, depending on the model, we find a phase transition at a critical temperature.

The Ising model describes, e.g., ferro-magnetic behavior (changing from the high-temperature non-magnetic to the magnetic phase). The XY-model is richer in structure. It exhibits topological objects, vortices and anti-vortices. They are defined as special configurations of the four spins located at the corners of an elementary square of the lattice (a plaquette). If the differences of the angles θ_i at the corners sum up to 2π when visiting them in the counterclockwise direction, we speak of a vortex. For the case of -2π, an anti-vortex sits at the plaquette (compare Fig. 1). In case the sum is 0, no topological object is present at the plaquette. The structures are topologically stable and change their behavior depending on the temperature. The XY-model exhibits a Kosterlitz-Thouless phase transition [7]. At this transition, the vortex - anti-vortex pairs, that are close together at low temperatures, become unbound (compare Fig. 3).

3.2. Software implementation

Numerically, spin models can be treated with *Monte Carlo algorithms*. For each update of the spin configuration, a lattice site is chosen, and a random candidate spin proposed. If the overall energy decreases, the new spin value is accepted. If the energy of the candidate configuration is higher, the new spin is accepted only conditionally based on a random decision. Else, the old spin value is kept. (For a description of the algorithm, see [8].)

In order to obtain smooth configurations of spins, we apply a *cooling algorithm*. In this case, the overall energy always decreases . The problem is, that cooling will average all vortex pairs out to establish a configuration of minimum energy, i.e. a laminar lattice with all spins aligned. Thus it depends on experience when to stop the cooling process. (See Fig. 2.)

We implemented the model and the sonification package (see below) in SuperCollider3, a free programming language developed for real-rime audio synthesis [9]. A graphical user interface (GUI) allows to visualize the spins, and was used to produce the plots in this paper. The temperature can be changed dynamically. Also the location of vortices and anti-vortices can be computed, and is indicated in the GUI. The whole package runs well with a 38x38 lattice in real time, which is rather a small size for an up-to-date simulation of this system. Still, the behavior of the phase transition is correctly reproduced.

4. SPIN QUARTET SONIFICATION

For the sonification approach the plaquettes are the starting point – four neighboring sites on an elementary square, that carry the topological structures. We refer to them as *spin quartets*:

$$s_{x,y,i} = (s_i, s_{i+\hat{x}}, s_{i+\hat{x}+\hat{y}}, s_{i+\hat{y}}) \qquad (1)$$

For each spin quartet, a sound grain y_r is played, where a sine oscillator is modulated depending on the spin values. The sonification operator[1] is given as:

$$\mathring{y}(\mathring{t}) = \sum_{r(R)} L_{n_s}^{\infty} \left[a_r(T, t_r) \cdot \mathcal{F}_{f_{0,r}}^{BRF,2} \left[y_r(\mathring{t}, d_{x,y}, r) \right] \right] \qquad (2)$$

In principle, the sonification signal $\mathring{y}(\mathring{t})$ is the sum over r spin quartets (within the range R). Each quartet is looped infinitely (or until the user changes the selection), $L_{n_s}^{\infty}$, over n_S samples. The signal of each spin quartet is $[y_r(\mathring{t}, d_{x,y}, r)$, as described below. It is filtered with a band reject filter $\mathcal{F}^{BRF}$ of second order by a frequency $f_{0,r}$.

$$y_r(\mathring{t}, d_{x,y}, r) = sin\left(2\pi f_{0,r}\mathring{t} + \phi_r(\mathring{t}, d_{x,y})\right) \qquad (3)$$

Each spin quartet data $d_{x,y}$ is used to modulate the phase of a sine oscillator.

$$a_r(R, t_r) = Env_{t_t, t_a, t_d, a_{max}(R)}[a] \qquad (4)$$

The phase is constructed in the following way: the data values s' are cubically distorted and normalized. The distorted phase of an anti/-vortex still yields a final value of $\pm 2\pi$. Other configurations are suppressed, see Fig. 4. Then they are up-sampled by a factor of S ($\uparrow_S$) samples, and interpolated with a cosine function over $S/4$ samples.

The resulting phase distorted ramp is looped L_S over S samples. This is the phase that controls the phase of a sine oscillator with the base frequency f_0. (This frequency is filtered out in the end, as described above, thus only the frequencies resulting from the phase modulation and their overtones remain in the signal.)

$$\phi_r(\mathring{t}, d_{x,y}) = L_S \left[\uparrow_S INT_{\frac{S}{4}}^{cos} \left[\frac{1}{4\pi^2} s'^3_{x,y,i} \right] \right] \qquad (5)$$

The differences between adjacent spin values in the plaquette $\delta s_{x,y,i} = s_{x,y,i+1} - s_{x,y,i}$ are calculated assuring that the cumulation is continued in always the same rotational direction.

The $\delta s_{x,y,i}$ are added up in counter-clockwise direction to form a cumulative sum of the angles' differences $s'_{x,y,i}$:

$$s'_{x,y,i} = \sum_{n=1}^{i} s_{x,y,n} \qquad with\ s_0 = 0 \qquad (6)$$

For an ideal vortex and antivortex, $\delta s_{x,y,i}$ is $+\frac{\pi}{2}$ and $-\frac{\pi}{2}$. The cumulative sum $s'_{x,y,4}$ is accordingly $+2\pi$ and -2π. Neutral spin quartets containing no anti-/vortex[2] show a total rotation of $s'_{x,y,4} = 0$. (Any other configuration than anti/-vortices has values between $-\pi$ and $+\pi$, but with a total rotation of 0.)

The resulting curve s'_0 to s'_4 is used for the sonification shown in Eq. (6).

In the case of a vortex, the phase raises by 2π and the resulting frequency increases. In the case of an anti-vortex, the frequency is lowered due to a negative phase slope. The number of samples between each of the spins is $S = 30$. This results in a frequency of $f_p = r_s/4S = 44100/120 = 367.5\,Hz$. We choose a basic frequency f_0 of $3f_p = 1102.5\,Hz$. Thus, $(f_0 + f_p) = 2(f_0 - f_p)$, and a vortex and an anti-vortex are one octave separated.

The sonification is used interactively. Many spin quartets are played simultaneously around a central clicking point; their range of neighborhood, R, can be chosen, and $r = r(R)$ gives the number of simultaneously played quartets.

[1] Recently, J. Rohrhuber [10] suggested the formalization of the sonification operator, to make the mapping between the domain science and the sound synthesis more explicit. We take up this idea and extend the formalization by notation suggestions, as used in Eq. (2-5)

[2] This notation is used for 'vortex or anti-vortex'.

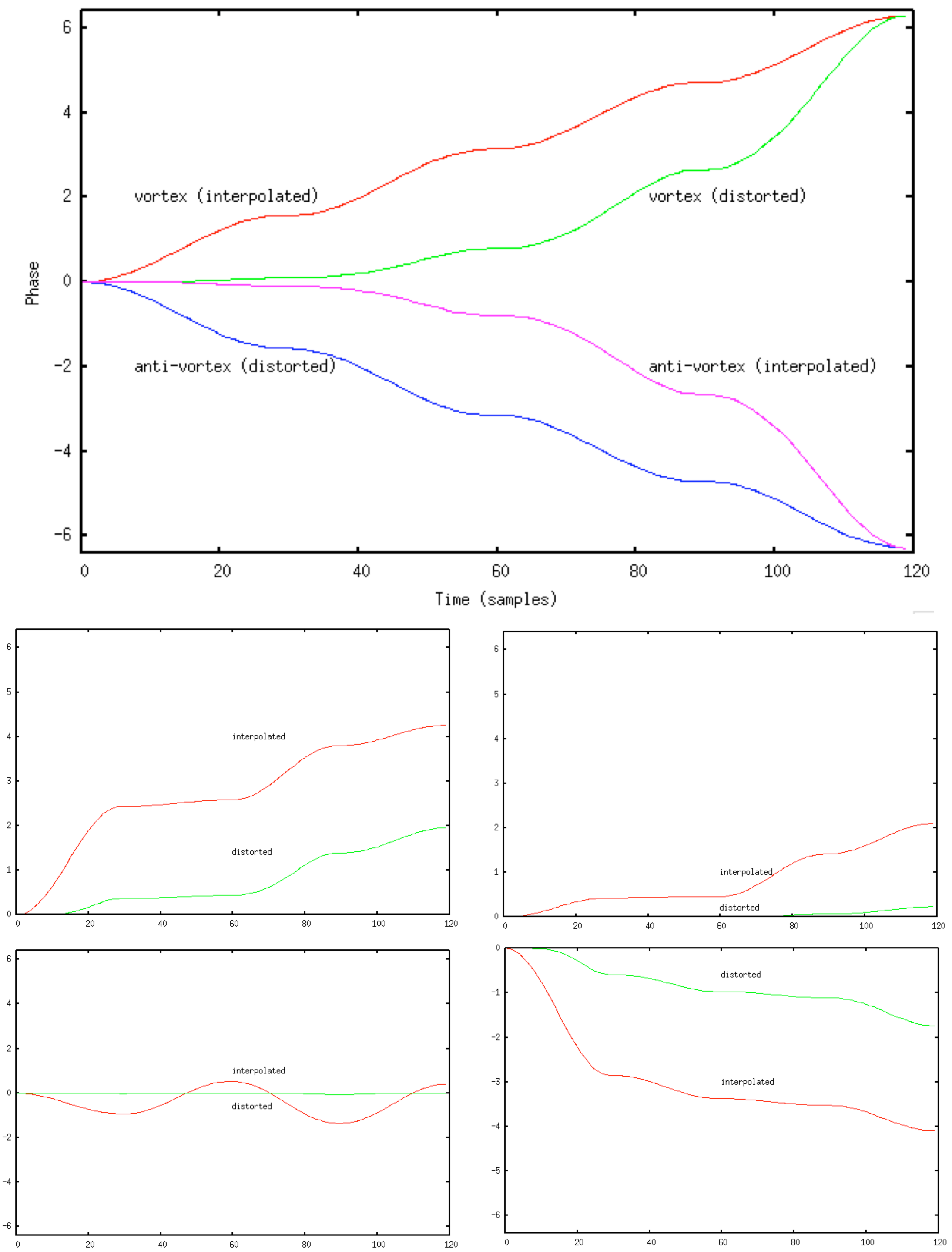


Figure 4: Phases of the ideal anti-/vortex (*upper figure*) and random configurations (*four lower figures*) for the XY sonification. The interpolated curves are depicted in red (and magenta), the distorted one according to Eq. 5 in green (and blue). The x-axis gives the number of samples. When the phase is looped, it is added smoothly to the last value, the base oscillation is periodic in 2π and does an effective modulo.)

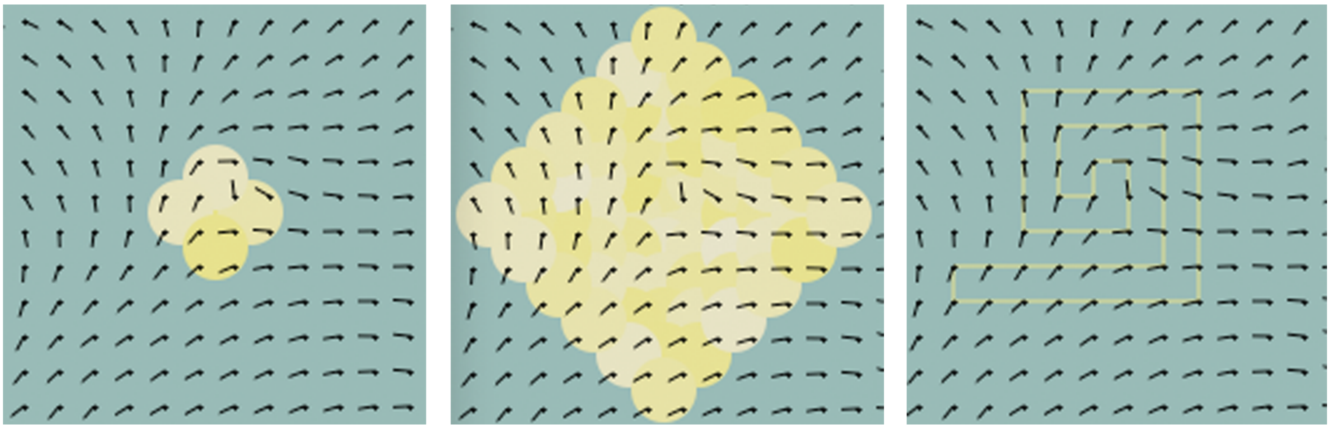


Figure 5: Interactive GUI of the XY model showing different listening range modes: In the neighbors mode (*left and medium figure*) a certain number of spin quartets around the clicking point is sonified. As an additional, more comprising mode of interaction, a spirale path was implemented. The spirale has variable length – order 1 gives a spin quartet. This setting is more exploratory, as the possible phase differences become more complicated. The sound of a vortex or anti-vortex depends on its position in the spirale and several anti/-vortices can be encompassed in one spirale.

A spotlight indicates all playing quartets in the GUI, see Fig. 5. Each sound is enclosed in an envelope Env and looped by the looping operator L, until a new site is chosen. The duration and loudness map the distance to the clicking point, thus closer neighbors sound louder and quicker, $a_r(t_r)$. This information is also encoded in a mistuning of the base frequency, thus $f_0 = f_{0,r}$. Very close anti/-vortex pairs will have nearly the same base frequency in octaves. If the pair is further shifted, the interval is mistuned, resulting in a beating of varying frequency. This is a key feature of the sonification that allows to distinguish the difference between bounded pairs and a vortex plasma. To give some orientation, left/right panning is applied, $\mathring{y}(\mathring{t}) = (\mathring{y}_L(\mathring{t}), \mathring{y}_R(\mathring{t}))$.

5. DISCUSSION

The presented sonification employs simple acoustic properties of a phase modulation. We believe that it is intuitive, as only the interesting parts sound and a laminar *field* of spins is silent. The difference between a vortex and an anti-vortex is absolutely clear, which is a major advantage over the visualization. The sonification gives also information where the eye has substantial problems of finding structures at all.

We did a short pre-evaluation. In a blind-test environment, the temperature had to be assessed in different settings: only-visual, only-audio and visual-plus-audio. The authors were the testing subjects. Even if the results are not statistically exploitable due to the small number of participants and test runs, a tendency was found that the audio-visual case had best results, shortly followed by the only-audio case. The results of the visual-only case were much worse in terms of temperature assessment. Still, it has to be admitted, that the visual case was by far the fastest way of assessment. This shows that in a sonification, one often has to invest more time, but it also can lead to a more precise results.

6. CONCLUSIONS

In this paper we described a sonification of a system of statistical physics, the XY-model. Its interesting topology exhibits vortices and anti-vortices. These are defined by a nontrivial accumulated rotation of $\pm 2\pi$ and used to control the phase of a sine oscillator. The sonification allows to hear only the interesting structures, *without* calculating them beforehand. Vortices and anti-vortices can clearly be distinguished, even for untrained listeners. Information on the closeness of the vortices and anti-vortices is encoded by using slightly mistuned base frequencies, that result in a beating for remote vortices. The sonification is controlled interactively via a graphical user interface.

For the future, we will use the experiences gained with this model to study more complex ones. The ultimate goal is lattice QCD, a huge and high dimensional model, with a rich structure of topological objects.

Listening examples/screenshot videos.

Listening examples and further documentation can be found at http://qcd-audio.at/results/xy.

Acknowledgments.

Our research project is funded by the Austrian Science Fund FWF within the Translational Research Program.

7. REFERENCES

[1] K. Vogt, *Erstausgabe*. Karl-Franzens Universität Graz, 2008, vol. 1, ch. Sonification and particle physics.

[2] S. Bokesoy, "Sonification of the fission model as an event generation system," *Proc. of the 7th Int. Conference on Digital Audio Effects (DAFX-04), Naples, Italy*, October 2004.

[3] B. L. Sturm, "Composing for an ensemble of atoms: the metamorphosis of scientific experiment into music," *Organised Sound: Cambridge University Press.*, vol. 6, no. 2, pp. 131–145, 2001.

[4] T. Höllerer, J. Kuchera-Morin, and X. Amatriain, "The allosphere: A large-scale immersive surround-view instrument," *IEEE MultiMedia*, vol. 16, no. 2, pp. 64–75, Apr.-June 2009.

[5] SonEnvir-Team. Sonification environment research project. [Online]. Available: http://sonenvir.at

[6] K. Vogt. Qcd-audio research project. [Online]. Available: www.qcd-audio.at

[7] J. M. Kosterlitz and D. J. Thouless, "Ordering, metastability and phase transitions in two-dimensional systems," *J. Phys. C*, vol. 6, pp. 1181–1203, 1973.

[8] H. Gould, J. Tobochnik, and W. Christian, *Introduction to Computer Simulation Methods*. Addison Wesley Pub Co Inc., 2002.

[9] J. McCartney, "Rethinking the computer music language: Supercollider." *Computer Music Journal*, vol. 26, no. 4, pp. 61–68, 2002.

[10] J. Rohrhuber, "Sonification variables," in *Proceedings of the Supercollider Symposium*, 2010.

A SONIC TIME PROJECTION CHAMBER. SONIFIED PARTICLE DETECTION AT CERN.

Katharina Vogt, Robert Höldrich,
David Pirrò, Martin Rumori

Institute for Electronic Music and Acoustics,
University of Music and Dramatic Arts Graz,
Austria
vogt@iem.at

Stefan Rossegger, Werner Riegler,
Matevž Tadel

CERN - European organization for
nuclear research,
Geneva, Switzerland
stefan.rossegger@cern.ch

ABSTRACT

In a short-term research project at CERN, an auditory display of elementary particle tracks has been developed. Data stems from simulations of the Time Projection Chamber (TPC) in ALICE experiment. Particle detection there is based on pattern recognition algorithms, but is still today double checked with visualization tools. The sonification works with cluster data of the TPC and was designed in analogy to the physics behind the measurement device. Thus it is possible to listen directly to the otherwise silent detector.

1. INTRODUCTION - SOUNDING CERN

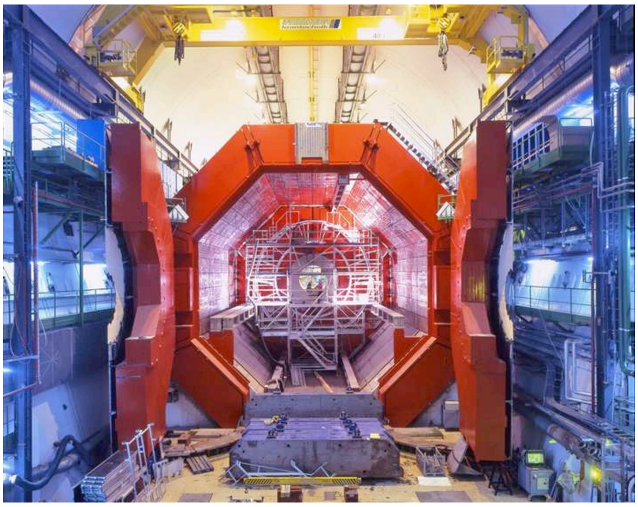


Figure 1: The cavern of the ALICE experiment - a 50 m high dome, 50 m under ground. The huge magnet doors are closed and the beam pipe is mounted and shielded today, as particle beams are circulated since November 2009 from where the photo was taken. In the middle of the detector, the TPC was installed - one read-out chamber at the front, where the doors are, the other at the back. *Photo: A. Saba, http://aliceinfo.cern.ch.*

During a three months research visit at the European Organization for Nuclear Research CERN, the principal author of this paper had the opportunity of getting insights into this scientific community. Around 3500 physicists, engineers, other scientists and staff are working at CERN. At the time of this stay the newest experiment was started, the Large Hadron Collider (LHC). The dimensions of CERN are impressive in every respect: Particle beams are accelerated in a tunnel of 27 kilometers length to nearly the speed of light. Two counterrotating beams collide at four possible sites, where different detectors are mounted. They detect traces of the particles they are specialized at, that have been produced in the collisions. (One of the collision points is ALICE, A Large Ion Collider Experiment, see Fig.1). The LHC experiment has been planned for two decades and will run for 10 to 20 years. In the planning of the beam acceleration and the various detector facilities, simulations were done for experiments that have been realized much later.

CERN is naturally a very open community, as there is a high fluctuation of scientists from all over the world and a melting pot of all kinds of technologies. Sonification has not been known to any of the physicists we spoke to before. Often, sonification was put into context with sounds occurring in experiments or simple analysis: from alarms in the control room to the fine-tuning of parts of detectors to listening to the beam spectra *"because there was nothing else to do and you just had to plug in headphones"*. Thus the idea of using sound was not so new to many, though systematic studies of sonification have not been conducted at CERN.

In our short term project, we gained an overview over sounds and CERN. It can never be complete, as the organization is wide spread and diverse, with thousands of collaborations with external institutions. Still, a few interesting projects have been found: the sonification of beam spectra, referred to above; or for instance the mounting of microphones in the LHC tunnel, in order to monitor the performance of collimators, devices for the so-called cleaning of the beam (for a pilot project on sound measurements on a prototype for the collimator, see [1]). The analysis of beam spectra is worth mentioning in some detail, as it is an ideal application of sonification – while their makers didn't even know this term.

Parameters of both LHC particle beams, like horizontal and vertical position in the vacuum chamber, are measured at many places in the tunnel. The particles are grouped in bunches, that have transversal and longitudinal oscillation modes. These oscillations can be described in phase space and should stay outside resonances, otherwise beam oscillations grow and the beams might get lost. The oscillation modes are measured accurately, as they are very important for keeping the beam on a stable orbit for the 27 kilometer circumference. The resulting transversal (β-tron) and longitudinal (synchrotron) oscillation frequencies range from a few tens of Hz to a few kHz, therefore they are audible without any further processing. With this sonification, many details of beam dynamics can be monitored by listening, in parallel to standard observation usually done in the frequency domain by performing real-time Fourier analysis of the beam signals. As full performance of the LHC is expected only in 2010, sonification results from this machine will be published in the future. Sonification of beam oscillation signals from other machines - the Super Proton Synchrotron at CERN - have also been tried out, but without any regular studies. Information and soundfiles can be found at [2, 3].

In our sonification approach, we chose data from one experiment, ALICE, and focused on its main detector, the TPC. The data we worked with is still simulated proton-proton collisions, but the sonification can be used for real measurement data as well. For this data, a well-developped visualization tool exists, which is called AliEve [4].

2. PARTICLE DETECTION AS A PATTERN RECOGNITION PROBLEM

The search for subatomic particles always depends on indirect measurement. Obviously, no direct perception is possible. Still, since the electron was detected in 1897, physicists tried to find traces of ever harder detectable particles. Usually, a huge amount of noisy raw data has to be assessed, and patterns in the data - tracks - give evidence of elementary particles having passed.

The history of particle detection is one of technological and theoretical achievements on the one hand side and the (re-) organization of scientific labour on the other, starting from single physicists in small laboratories to research groups of hundreds of scientists. But, for a big part, it is also the story of human pattern recognition. P. L. Galison argues in his book *Image and Logic* [5], that two rivaling methods have been developed in the 19th and 20th century. Logic, in the form of logic circuits, with many events treated statistically, stood against the golden event of the image tradition, where one single picture could proof a new particle. Particle detection is an inherently stochastic process, thus was the argument of the statistical approach of the logic community. During the last century, the big experimental particle research organizations were still largely led or influenced by individuals, who supported one or the other school.

The image tradition always relied mainly on the human being. “Alvarez [a leading proponent of the image tradition] wanted the ‘human operator’ to be ‘the black box pattern recognizer”’ [5, p.391]. But they had to struggle with the ever rising amounts of data being produced by the latest detectors. In the 1940s human operators (the ‘scanner girls’) assessed the pictures of bubble chambers according to certain criteria, before physicists would analyze only the more interesting events. Different technologies were developed in order to aid and accelerate the human scanning process. The logic tradition was pursued at CERN (amongst other institutions). One of its leading persons, L. Kowarski, put the idea in 1960 as “[t]he evolution is towards the elimination of humans, function by function.” [5, p.371]. The more measurement devices produced data, and the more pattern recognition algorithms were refined, the logic tradition prevailed. Still, a combination with the fine-coarse measurement data of the image tradition with obvious and pervasive results could not be achieved for a long time.

During the long development of detectors, a few times sound was used implicitly or explicitly. Very early versions of the Geiger-Mueller counter had such a large voltage supply that a sparkover caused a bang as well. Today still, the typical Geiger counter display is an auditory one. Eyes-free conditions in radio-active environments has of course huge advantages for physicists and engineers. When they work on the machinery, they get information on what humans cannot perceive with their senses.

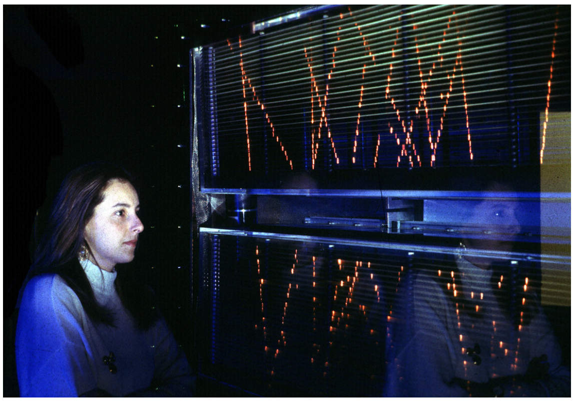


Figure 2: *Spark chamber at CERN’s permanent exhibition Microcosm, source: http://cdsweb.cern.ch/record/39277.*

The logic of the Geiger counter was pursued in spark chambers. The sonic chamber was a device used in the 1960s at CERN and one example is still shown in their main exhibition called *Microcosm* (Fig.2). In a spark chamber, an energetic retort is produced between two plates. If a spark crackles through air, a loud *bang* is produced, which is recorded by microphones and thus can be counted. These detectors were called sonic chambers. A next development step was the wire chamber, which uses a similar but ‘silent’ technique. An external electrical field accelerates electrons towards highly charged wires, where they can be detected.

The first detector finally fusing the logic and image tradition is the TPC, invented in 1974. It is an extension of the (logic) wire chamber for the read out part, but with all benefits from the image tradition chambers. Many events can be recorded and studied by statistical means, while the tracks are reconstructed 3-dimensionally. Details are discussed in the next session.

While any particle detection is today based on algorithmic logics and statistics, visualization has still a standing. Data from the ALICE experiment is, e.g., finally cross-checked by human ‘scanners’ in organized shifts. Today, the goal of visualization is different than 50 years ago. Particle measurements became even more complicated with the reliance on computers. Thus the human scanners double check the functionality of the detection machinery and pattern algorithms and help debugging the extensive code. Furthermore there are applications for non-experts: firstly, newcomers -future experts- can become acquainted with all parts of the experiment. Secondly, outreach becomes more important also in high energy physics, where the energy (in terms of electron volt) as well as funding (in terms of money) are of high orders of magnitude.

Arguments of the image tradition, some 50 years ago, are remarkably similar to sonification arguments of today. Humans shall be presented with data with the least hypothesis applied in the display and search them for patterns in order to allow for the formulation of new hypothesis. Pattern recognition is very quick when comparing for instance real data to simulated one. All those arguments can as well be applied in favor of sonification, with all known additional advantages, but also drawbacks, of hearing vs.

seeing [6].

3. THE TPC - A PARTICLE DETECTOR

At the heart of the ALICE experiment there is a TPC installed – the most exact and with 5 m diameter the biggest which has ever been built. It is a detector consisting of a cylindrical gaseous volume mounted around the collision spot. If particles are produced in the collision, they cross the gas and ionize it by hitting the gas molecules. The freed electrons are lead in an electric field parallel to the beam direction, to the left or right. At both sides, read out chambers are situated. They consist of different layers of wires at high potential producing an electrical field and accelerating the electrons towards them. In an avalanche process electrons are multiplied and the induced current can be read out. From the time, they are hit by the collision particles, the electrons move (ideally) straight and with a constant drift velocity towards the read-out chambers. Thus the information on their impact time and location on the circular read-out chamber suffices to reconstruct the particle path exactly.

Depending on the energy deposit on the wires and the shape of the tracks, physicists and algorithms can deduce which particles were produced in the collision. This is not as straight forward as simple plots or sonifications of the raw data suggests. our perception groups single events that form a shape automatically. In the measurement, single signals from the read out pads behind the wires have to be combined to find the center of a freed electron cloud (cluster). These clusters are then grouped to a complete track. Analysis of the shape of the track makes it possible to associate a certain particle with it – the one that must have caused it. In a second step, 'the physics' can be studied and interpreted. Each collision is only measured partially, as very short lived particles and decay products do not leave a trace.

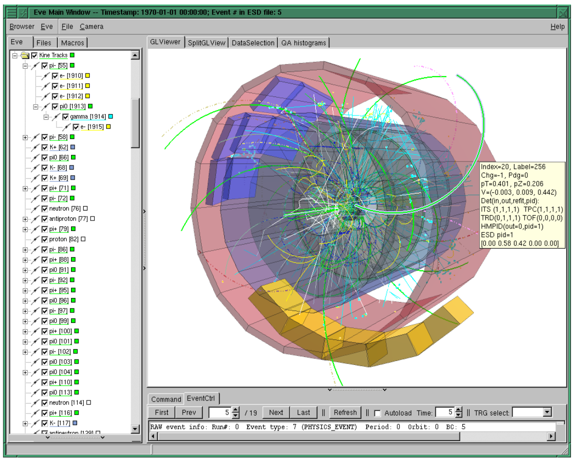
4. SONIFICATION

4.1. Methodology and goals

This sonification was developed in a short-time visit at CERN. We wanted to achieve an intuitive sonification of basic cluster data, not yet grouped tracks, which is based on analogies to the measurement.

One goal of the sonification is to extend the visualization. The primary visualization tool of ALICE is AliEve [4]. It allows 3-dimensional display of all the detector's data. The display is freely moveable. AliEve is a full software package, and our sonification can of course not compete with the functionality. Still, it could be a first step towards an additional auditory display.

The provided data sets are simulated *events* of p-p collisions, containing up to 35 tracks comprising a few 100.000 single electron impacts, each given at a certain time (t_i) and location ($\phi_i,\ r$), with an energy deposit (e_i). These are the simulated raw data expected in the measurements; further information is given from a second level of pattern recognition, i.e. which single electron impacts form a track caused by one particle. The data files are usually smaller than the ones of lead-lead collisions, that contain each around 80 MB of data or 60000 primary tracks. Still, it was challenging to stick with the raw data: hundreds of thousands of single electron impacts, each with a certain time, location and energy deposit.



Index=20, Label=256
Chg=-1, Pdg=0
pT=0.401, pZ=0.206
V=(-0.003, 0.009, 0.442)
Det(in,out,refit,pid):
ITS (1,1,1,1) TPC(1,1,1,1)
TRD(0,1,1,1) TOF(0,0,0,0)
HMPID(out=0,pid=1)
ESD pid=1
[0.00 0.58 0.42 0.00 0.00]

```
RAW event info: Run#: 0  Event type: 7 (PHYSICS_EVENT)  Period: 0  Orbit: 0  BC: 5
```

Figure 4: *A screenshot of AliEve [4], the visualization tool of the ALICE offline group. The reddish surface gives the volume of the TPC (yellow and blue are other detectors). Each line is the track of this one event -a certain time of measurement after a collision with a certain amount of particles produced.*

4.2. A sonic time projection chamber

The sonification is a parameter mapping that uses the raw data of single electron hits, allowing for a perceptual grouping into tracks, following auditory grouping principles [9].

Based on the fact, that 'electrons' (in fact electron clouds) hit the wires with a certain charge (the number of electrons), the wires are taken as analog to *strings*, which are hit and resonate with their basic frequency depending on their length.

It was a natural choice to place the listener at the collision point and let the time evolve towards the left and right read-out chambers. The time in the raw data is given *inversely*, as it is the time of the electrons freed by the particles passing nearly at the speed of light. Those electrons reach the read-out chambers first that are closest to them, and the time in the raw data thus evolves from outside back to the collision point. The sonification time is inverse to the data time, as it is more natural to follow the tracks from the collision point outwards.

In order to enhance the perceptual grouping and separation of tracks, we had to disambiguate those which are in the same height of radius but at a different azimutal position (given by ϕ in spherical coordinates). Determined by this angle, we add different sets of overtones to the base frequency. In order to achieve different timbres, the base frequency is either played solely for $\phi = 0^\circ$ (where the amplitudes of all even and odd overtones are 0), or with just one set of overtones (odds = 0, evens =1 for $\phi = 90^\circ$ or vice versa if $\phi = 270^\circ$), or as a full sound at $\phi = 180^\circ$ (evens and odds = 1). See Fig. 6.

For angles in between these extreme positions, there is a linear mapping of rising or falling of overtones, introduced as a weighting factor $w_i(\phi)$. The amplitudes are in the first place weighted with $w_k = 1/n$ for n being the harmonics. Finally, the sum of all amplitudes was normalized to 1 in order to avoid clipping.

This differentiation of timbres allows the correct grouping in human perception: following the *gestalt* psychological principle of *similarity*, similar sounds are grouped together and believed as coming from the same track. Pitch is a very strong grouping fac-

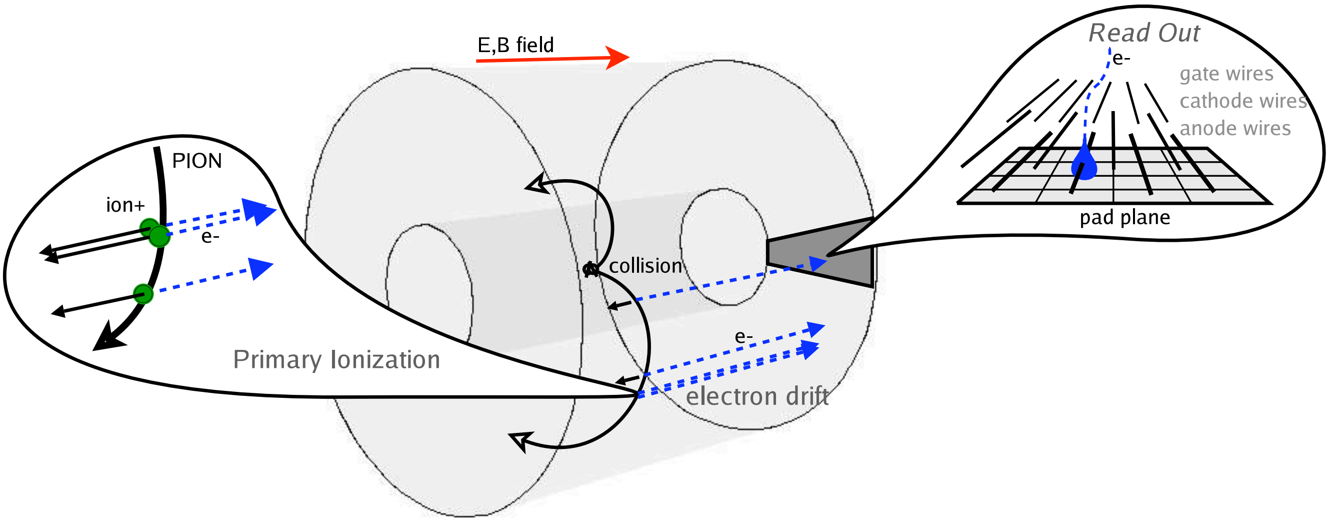


Figure 3: *Schematic plot of the working principle of a TPC, at the example of a pion track. Charged particles transverse the volume of the drift chamber and ionize the gas. The created electrons follow the applied electrical field (E) and are collected in the wire chambers, where they are read out. Three layers of fine copper and wolfram wires are strained on top of each other with some mm between the layers (gate wires, cathode wires and anode wires). The triangular elements are mounted on each of the two read out chambers. The inner wires are shorter than the outer ones, ranging from 27 cm to 84 cm (in total, there are 656 wires from inside to outside.). Calculating the real frequency range for these wires results to 0.0028 and 0.0089 Hz. All technical details of the ALICE TPC can be found in [7]. Source: [8]*

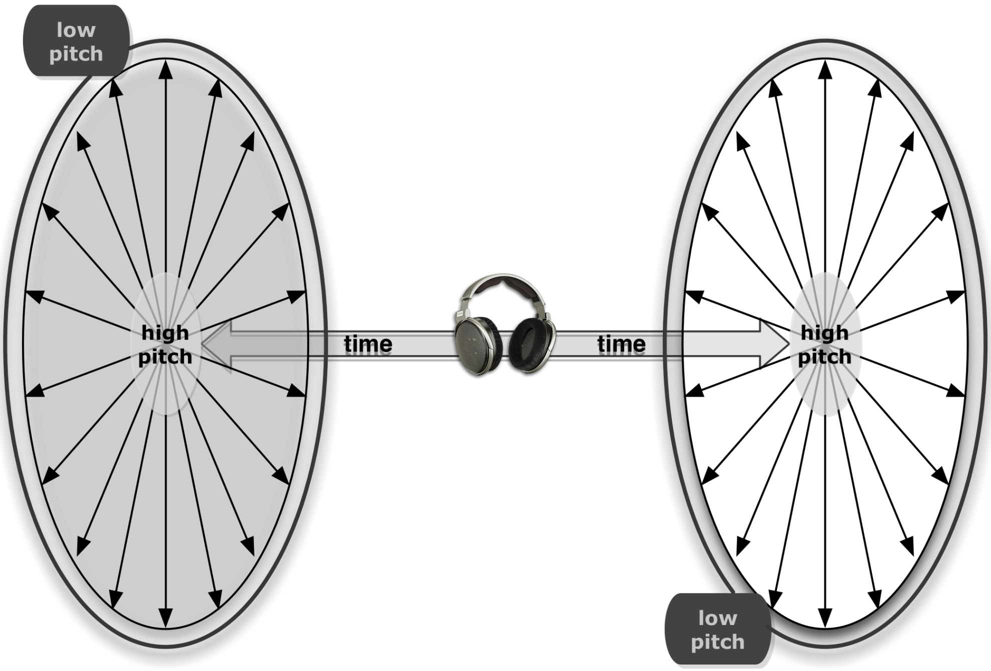


Figure 5: *Scheme of the sonification: the listener is virtually placed in the center, where the beams collide. The two read-out chambers are situated to the left and right hand side. The strings in the center of the read-out chamber are shorter and higher pitched. The tracks start playing in the point of collision and evolve simultaneously to the left and right hand side. The volume of the sounds represents the charge deposit of the electrons. And finally, in order to not confuse tracks that are close to each other, they sound slightly different - as different instruments of an orchestra. If more electrons are hitting wires within a short time, the sounds overlap and auditory grouping happens. Thus we hear a continuous and coherent sound for each track rather than single tones for each single hit.*

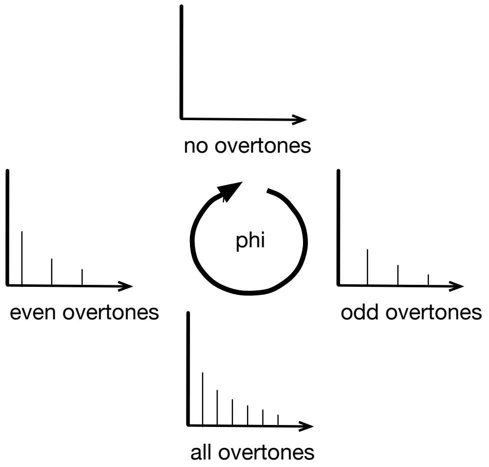


Figure 6: Simplified scheme of the overtone structure in the TPC sonification depending on the angle ϕ. It helps disambiguating the correct grouping and separation for single sounds into coherent tracks, even if these tracks have similar pitches.

tor. As additional cues, similar sounds always follow close to each other (principles of *proximity* and in *good continuation*).

Each sound consists of a bank of resonators $\mathcal{F}_{band}$ with frequencies $f_{i,k}$ specified according to the pitch mapping. The filter bank is excited with an impulse, and enclosed by an envelope $a(\mathring{t}, e_i)$. The level of the impulse and thus the amplitude of the resulting sound are determined by the charge deposit of the electron. Tracks with only few single electron impacts or very weak ones fall silent.

The sonification operator[1] is given as:

$$\mathring{y}(\mathring{t}) = \begin{pmatrix} \mathring{y}_L(\mathring{t}) \\ \mathring{y}_R(\mathring{t}) \end{pmatrix} \tag{1}$$

The $\mathring{y}(\mathring{t})$ denotes the sonification signal, depending on a sonification (listening) time. $\mathring{y}_L(\mathring{t})$ and $\mathring{y}_R(\mathring{t})$ denote the left and right channel of a stereo or binaural rendering.

$$\mathring{y}(\mathring{t})_{L,R} = \sum_i Trig_{t_i}\left[y_i(\mathring{t}, d_i)\right] \tag{2}$$

The sonification consists of a sum over all i (each a single electron impact) of single sounds $y_i(\mathring{t}, d_i)$, that are triggered at respective times t_i as given in the data.

$$y_i(\mathring{t}, d_i) = a(\mathring{t}) \sum_k w_k w_i(\phi) \mathcal{F}_{band, f_{i,k}}[\mathcal{I}(e_i)] \tag{3}$$

Each of these single sounds is a weighted filtered impulse.

$$f_{i,0}(d_i) \in [200, 800]^{exp}\ Hz \tag{4}$$

The frequencies are mapped exponentially between 200 and 800 Hz.

We rendered stereo files and also a binaural version, but the latter seemed not to work well with 'imaginary' paths (as the perception has no fix references, but virtual 'flying' objects close around the head). Simple stereo panning was less effort to render but even clearer perceived in addition to the visual cues of the screenshots.

In the current setting, each event takes 10 seconds of sonification time. This span can be shortened, of course, but is a good length to disambiguate tracks even in the more complicated events.

As it is difficult to listen to many tracks at once, all tracks can be chosen and played individually. In Fig.7, you see a screenshot for one sound example on the homepage. One track is marked, which is played solely. In this case, the particle did not come from the collision, but stemmed from background radiation or some secondary process of disintegration. It is a charged particle, as it is whirling around in the exterior magnetic field of the ALICE experiment. The pitch is rising and falling, which matches the idea of a turning flying object.

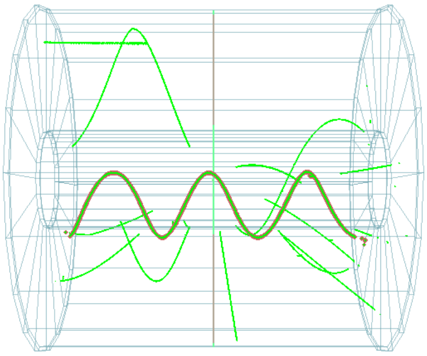


Figure 7: Screenshot of a sound example of the TPC sonification (event number 6), which can be listened to at www.qcd-audio.at/tpc.

Some remarks have to be made with regard to sound files of single tracks. Some tracks are clearly visible but nearly silent in the sonification. This is because the charge deposit of the electrons was very small, and indicates a low energy particle. Another reason may be that only very few electrons constitute a track - either this is a measurement or track counting error or the passing elementary particle really only kicked out a few electrons.

Sound examples can be accessed at: www.qcd-audio.at/tpc.

5. DISCUSSION AND OUTLOOK

The sonification as described in this paper has advantages as well as short-comings. First of all, the outcome is very simple - which we regard as a huge benefit. The mapping is made in analogy to the measurement, thus easy to understand. Also, raw data is sonified, and all pattern recognition of grouping tracks is done automatically by auditory perception.

A 'normal' event will mainly include tracks from the collision. If they start at the height of the beam axis, their pitch is falling.

[1] Recently, J. Rohrhuber [10] suggested the formalization of the sonificaiton operator, to make the mapping between the domain science and the sound synthesis more explicit. We take up this idea and extend the formalization by notation suggestions, as used in Eq. (1-4)

This sound matches very well a real-world sounding object which is thrown away. Still, many other things might happen, where the real-world association does not make sense any more. As it was shown in Fig.7, particles might also stem from background radiation or secondary derivation (the particles stemming from the collision were e.g. neutral and thus could not be detected, but they launched other particles at different places than the collision spot). This creates also problems for the timing, as it is assumed to start in the collision point and evolve towards the read-out chambers. Still, the time is the reversed but otherwise 1:1 measured time of the TPC and thus represents the measurement.

The example shown above, and also most examples on the webpage have rather few tracks. There is also data for other types of collisions, e.g., between two heavy lead nuclei, as pb-pb, which produce thousands of particles in a collision. For such a data set, we did not apply the sonification. We assume, that a sonification of a full such event would not be of much use. This is also true for a visualization. In such a case both sonification and visualization can only provide a very rough overview and tell the scanning person that *a lot* was going on. In AliEve, such plots for raw data are only used for outreach pictures, otherwise some kind of automatized data reduction is done instead of visualizing the raw data. Such a strategy would also be applicable for the sonification.

During the configuration of the sonification, we remarked that 3d plots are often very counter-suggestive. At least in the case, where one sees a bunch of lines on an else empty surface, that gives only scarce cues of dimensionality and perspective, one often interprets the real position of a track wrongly. Different viewpoint angles that can be interactively rotated, as it is possible with AliEve, make the estimate much more accurate. Sound, on the other hand, has other disambiguities. Even with binaural recordings, a cone of confusion stays at the very left or right hand side, and frontal and rear sound events are confused as well. This is addressed as tracks usually are not always in such a confusing area, but rather evolve somehow. This gives our brain an additional cue about the movement and real location of the particle.

The outcome of the project has been presented in two meetings at CERN (weekly meetings of the Offline group and the TPC group) in November 2009. For the physicists attending the meetings, this project clearly presented an interesting variety. Furthermore the sonification is presented as interactive installation in the permanent exhibition of the ALICE experiment at Point 2 of the LHC. This shows, that the idea of sonification was taken serious for didactic and outreach reasons. A continuation in research would imply an interactive bridge to AliEve, which would make *real-time* audio synthesis necessary. Until now, the rendering of an event takes between 10 to 20 minutes.

This project was ended with the stay of the main author at CERN. A continuation depends on time and financial resources. A second project was implemented, where physicists were questioned about their expectance on how particles *should* sound like. This would allow a different mapping, which is also intuitive to the people who can work with it, but provides much more information. As the methodology was completely different, this project was treated in another paper.

Remark on nomenclature.

We used the software SuperCollider3 [11] to sonify the data from the LHC. The LHC is a *super collider*. It was decided to build it only after the US government had abandoned the Superconducting Super Collider (SSC). The SSC would have studied even higher energetic particles. The programming language was called *super collider* after the SSC, as its initiator James McCartney lived in the same region where the SSC was built. In this project, we used super collider to sonify super collider data.

Acknowledgments.

Thanks to Till Bovermann for very SC3 hints. Also, we would like to thank many other people at CERN who contributed through discussions and feedback: These are Despina Chatzifotiadou, Kajetan Fuchsberger, Marek Gasior, Betev Latchezar, Luciano Musa, Ralph Steinhagen and Alja Tadel.

6. REFERENCES

[1] S. Redaelli, Olliver Aberle, Ralph Assmann, A Masi, and G Spieza, “Detecting impacts of proton beams on the lhc collimators with vibration and sound measurements,” in *Proc. of Particle Accelerator Conference, Knoxville, Tennessee*, 2005.

[2] M Gasior and R Jones, “The principle and first results of betatron tune measurement by direct diode detection,” 2005.

[3] Marek Gasior, “Homepage 3D-BBQ,” .

[4] Matevz Tadel and Alja Mrak-Tadel, “AliEVE - ALICE Event Visualization Environment,” *Proc. Computing in High Energy and Nuclear Physics Conf. 2006 (CHEP 2006) Mumbai, India*, pp. 398–401, 0000.

[5] Peter Galison, *image and logic. A material culture of microphysics.*, The University of Chicago Press, 1997.

[6] Gregory Kramer, Ed., *Auditory Display. Sonification, Audification and Auditory Interfaces.*, Proceedings Volume XVIII. Santa Fe Institute, Studies in the Sciences of Complexity, 1994.

[7] Alice Collaboration, “Technical design report of the Time Projection Chamber,” in *ALICE TDR 7*, 2000.

[8] Stefan Rossegger, *Simulation and Calibration of the ALICE TPC including innovative Space Charge Calculations*, Ph.D. thesis, Graz University of Technology, 2009.

[9] Albert S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound.*, MIT Press, Cambridge, Massachusetts, 1990.

[10] Julian Rohrhuber, “Sonification variables,” in *Proceedings of the Supercollider Symposium*, 2010.

[11] James McCartney, “Rethinking the computer music language: Supercollider.,” *Computer Music Journal*, vol. 26, no. 4, pp. 61–68, 2002.

HURLY-BURLY: AN EXPERIMENTAL FRAMEWORK FOR SOUND BASED SOCIAL NETWORKING

João Cordeiro, Nicolas Makelberge

Research Center for Science and Technology of the Arts (CITAR)
Portuguese Catholic University - School of the Arts
Rua Diogo Botelho 1327, 4169-005 Porto, Portugal
{jcordeiro; nmakelberge}@porto.ucp.pt

ABSTRACT

This project deals with the topic of social interrelations; its aim is to achieve a deeper understanding of the underlying mechanisms of these relations through the use of sound and mobile devices/ubiquitous computing. The proposed framework follows two interdependent directions: 1) using environmental sounds as input data for context analysis, 2) using sound as an output to express results (sonification).

This project is part of a long-term research project concerning sound based social networks, conducted at the Research Centre for Science and Technology in Arts (CITAR). The aim of this paper is to share some initial results, both practical and conceptual in form of a related work overview on social networking technologies, a conceptual design for a Facebook® application based on the project initial idea (including an IPhone® graphic interface proposal) and last but not least, an experimental framework for data communication between an IPhone® and a computer (using Pure Data through RjDj).

Keywords - ubiquitous computing, mobile devices, social networking, soundscape, virtual community, data visualization, sonification, field recording, sound classification, interface design, auditory scene analysis.

1. INTRODUCTION

The average person, normally operating on a casual listening mode [1], has an overall low awareness of his or her sonic milieu. It often takes a disruptive sound or one out of the ordinary to perceive and become conscious of changes in a personal soundscape. These disruptive sounds often owe their conspicuousness to their lack of reference to any visual cues which can be true both with sounds operating clearly out of context and/or because a sound source is out of one's sight. Other forms of disruptive sounds include those that listeners dislike but are forced to live with, while not being completely unaware of its source (ex. a neighbor's dog barking outside your window at 6 o'clock in the morning).

However, we as sound designers are more fully aware of our sound environment and subtle changes in it. We also are aware of its potential to convey information and portray not only places themselves but also specific situations. In cinema for example, any background noise often sets the sonic context where actions are to take place, which in turn is many times manipulated to induce certain emotions in the spectator, often relating to narrative tensions in the script. Sound designers in general are very well aware of background sounds psychoacoustic possibilities, and are not afraid to explore its potential. The same meticulous attentiveness is not only suitable for sound design in cinema but also, as we have seen, for public health, marketing, architecture, urbanism, environmentalism, videogames, automobiles or product design etc.

As far as we know, soundscapes - as an integral element of social-network-building is still a quite unexplored area. Hence, we would like to investigate this same potential (as we see in all aforementioned examples) to characterize people's auditory milieus. We see soundscapes as yet another untapped source, which could provide unforeseen, interesting and valuable information about a person's life-style. Information, which we hope, may enrich and complement a social networkers profile.

The value added as we see it, is that by using sound as a tool within social networking, we not only allow for a new understanding of social interrelations within a group (a network) but also an different, yet efficient way for group awareness.

Beside the systems social function, it could also be applied as a research tool where one is enabled through the dissemination of such a technology to gather sound data on a wide scale. Researchers from various areas such as sound ecology, sociology, social geography could all benefit from this wealth of soundscape information to draw new conclusions concerning the constitution of various environments or sounds vis-à-vis other user variables such as location, age, gender, etc.

2. SOCIAL NETOWRKING AND SOUNDSCAPES

It is difficult to conceive the subsistence and development of human society without the act of communication among its constituent parts, humans. Nowadays, this idea as been extrapolated to such a degree, that we use the term information society when referring to developed societies. A society only possible by the emergence of technological information networks (mainly CMC) that completely subverted time and space dimensions turning them to "immediacy" and "space fluxes", respectively [2]. Computer mediated social networking could be said to be the byproduct of this technological development applied to the social interrelations realm. With this, new processes came forward to supply old needs on a wide range of human activities, such as work, entertainment, love or knowledge.

> *“One cannot, for example, be friends with absolutely anybody. People are constrained by geography, socioeconomic status, technology, and even genes to have certain kinds of social relationships and to have a certain number of them. The key to understanding people is understanding the ties between them; (...)” [3].*

By the annulment of time and space in computer mediated communications, relations between people in the network became easier to establish and easier to cease. Chat rooms took worldwide proportions and fake identities with stylish avatars circumvented embarrassing issues, like shyness or ugliness. All these issues seemed to promote the right context for the proliferation of social networks over the Internet, as one can verify on World Map of Social Networks (information on the map retrieved with Alexa & Google Trends for Websites traffic data) [4].

Users of social networks are normally invited to build a profile that others can see, sharing factual information (name, age, place of birth, look, etc.) and less-factual information (thoughts, convictions, whishes, moods, etc.) in order to build a virtual identity on the network. According to our research, the privileged type of information conveyed by each network will determine its character and, in some extent, the character of its users virtual community. Thus, a social network like Myspace, where users can easily upload music onto their profile page, tends to (not surprisingly) grow within a musical community made up of bands, producers and DJ’s. This same tendency, in general, can be seen among other social networks oriented towards all different types of relations, such as friendship, professional, artistic or touristic etc. These very same concerns are further developed in our research, although for now we will remain at the more practical questions concerning our project. As previously mentioned, our project privileges soundscape information rather then any other kind of information (including music and speech in a strict sense). Paradoxically, we don’t assume this network to be tailor-made exclusively for musician or sound designers (even if we believe these will interested) but for everyone engaged in social networking, especially those interested in friendship (on Myspace, for example, we observe a community not only formed by bands and musicians but also by their fans and friends).

We also recognize that sound is an integral part of people’s lives and might be somewhat overlooked in our *image-preponderant* society, which might warrant a method to rescue and update the concept of acoustic community postulated by Schafer [5].

> *“Community can be defined in many ways: as a political, geographical, religious or social entity. But I am about to propose that ideal community may also be defined advantageously along acoustic lines.”*

To support the idea of acoustic community we suggest the creation of “acoustic link” throughout the users network connections, allowing the expansion of each user “acoustic space” beyond the inherently physical constraints.

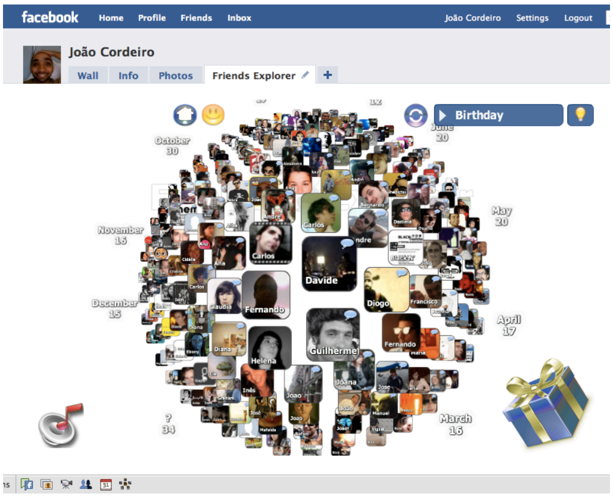


Figure 1 - GUI for Friends Explorer (visualization by birthday criterion).

3. RELATED WORK

The reduced size of this short paper is incompatible with the expected long and detailed description of a *related work* chapter. Thus, we will point out solely three examples, which we consider significant to our research for different but complementary reasons, as we explain later.

3.1. Friends Explorer

Friends Explorer is a Facebook application developed by Emonect (http://www.emonect.com/), which permits new visualizations and interactions within the user virtual community (Fig. 1), such as dynamic overview of all friends grouped by friends lists or other criterion (birthday, place of birth, actual place and alphabetic). Besides, this application is also designed for mobile devices interfaces, and pays special attention to usability issues like tactile navigation and scaling options.

This project is important to us once it also addresses network data visualization on mobile devices (though our regards auditory criteria).

3.2. FaceMic

FaceMic is a Facebook application developed by Voicetal LLC (http://www.voicetal.com/), to post audio updates from anywhere directly to the user Facebook wall using iPhone or iPod Touch. Like *Friends Explorer,* a suitable interface design for mobile devices as been achieved, with a simple and intuitive GUI (Fig. 2 - left).

Our project shares some characteristics with this application, once they both intent to work with pre-existent computer mediated social networks and make use of audio recordings with mobile devices (namely IPhone).

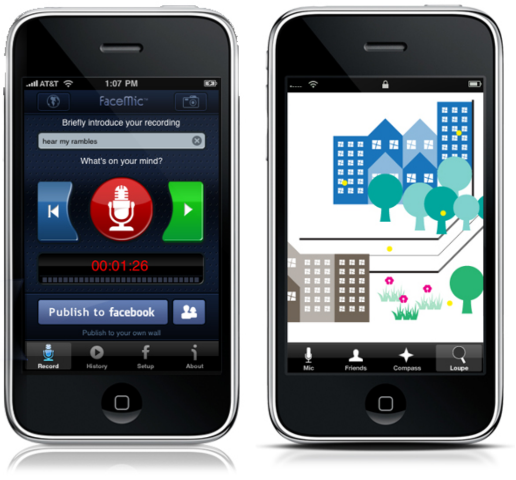


Figure 2 – IPhone GUI: FaceMic (left); Hurly-Burly - proposal for a graphic visualization (right).

3.3. Out To Lunch

Out to Lunch is a software prototype developed by Johnatan Choen, that "attempted to foster this sense of group awareness by using background sounds and an electric signboard to inform physically dispersed or isolated groups members about each other's presence" [6]. Succinctly, the software triggers background sounds (and visual sign) in each computer within a network, which relate to other users' presence. Thus, it emulates the physical presence by means of visual and auditory displays.

This project has much to do with our research, once both look forward to achieve a group awareness feeling with the use of displaced sounds over a computer network.

4. DESIGN WORKFLOW FOR A FACEBOOK APPLICATION

As described above, the goal of the project is to establish a social network based on sound, more precisely on users sonic environments or soudscapes. Therefore, we propose to capture and analyze the sonic environment of each user on a daily base, and distribute the data over the network, as described further on.

4.1. Name

Hurly-Burly is a popular expression used to describe a noisy and boisterous activity. We chose this name given the sonic result reached by each user when manipulating the application.

4.2. Social Network Application Framework

As a starting point we need a social network connected by means of a software application (virtual community). This can be accomplished by several means: 1) by setting up a new social network from scratch - people would have to sign up and invite friends; 2) by setting up a new social network based on contacts and profiles of an pre-exiting one; 3) by designing an application (app) for an existing software social network (like Facebook for example).

For now, we will assume the third method once it provides excellent starting conditions: 1) An established and well-known social community – users of Hurly-Burly could use their profiles and network community to run it; 2) Facebook policy encourages the development of new applications by external developers; 3) Facebook has a mobile platform.

However, we should note that by using Facebook as a platform is, yet, not a definitive option for Hurly-Burly framework. The perfect scenario would be an application suitable to be used with multiple social networks platforms as also as a standalone application (a new social network).

4.3. Device Framework

In order to serve the purpose of the project, mobile devices should be used to keep track on users daily moves. Beside mobility, this equipment should also live up to the following prerequisites: 1) Be part of users every day lives (non intrusive), 3) be able to compute data (ubiquity/pervasive computing), 4) be able to connect to the Internet, 5) be able to capture and convert audio to the digital domain, 6) have geo-reference capacity, 7) have a graphical display.

Despite these demanding specifications, most smartphones available nowadays in the market comply with these requisites. At this initial stage of the project, our choice falls to Apple IPhone which not only complies with all the requisites mentioned above but it's also a best seller (therefore, more suitable as an ubiquitous device), it's programmable, and its target market matches that which engages extensively in social networking.

Again, we assume that there may exist other devices that also fit the needs of the project, or even beat IPhone to the task. However, we will at this initial stage appoint the IPhone as the main hardware device of our experimental framework for now. Nonetheless, the tool should ideally not be exclusive, since the aim of the project is to reach an as broad audience as possible. Therefore, a cross-platform/cross-device application would ideally be desired.

4.4. Operation

Now that we covered the two main aspects of the system – social network application and device - we will explain how those two interrelate and how the proposed system works. Thereunto, we will describe the user experience:

First, the user A (our main user) should add the Hurly-Burly app to his Facebook page through the mobile device and be connected with friends (B, C and D) whom also are using both applications on mobile devices. User A's IPhone will record small samples of audio (about 3 seconds) every time the sound environment significantly changes. This audio is then uploaded to a server to be automatically classified according to

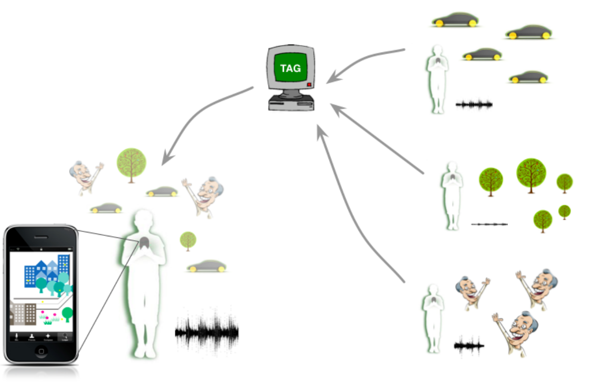


Figure 3: Data flow sonification (corresponds to the invisible status on regular instant messaging applications).

function and meaning (semiotics and semantics)[1]. We do know that the further we get with sound classification, the more accurate the system will be. However we also believe that an accurate classification of few soundscape classes can, thus, provide important information (noisy environment vs quiet, musical vs non musical, human vs non human). The audio and the data resultant from classification are then distributed to all of his network connections.

4.5. Data Visualization

Once the entire network is connected and exchanging data, a map will be displayed on the IPhone of each user. This map is based (primarily) on the received audio classifications, rather than on the geographical data of the other users (geo-localization). Thus, the resulting map is a cartographic iconic visualization designed according to the different sound classes (ex. people, traffic, animals, music.). Each map is unique and represents the sum of friends' soundscapes.

4.6. Auditory Display

The auditory display follows two different approaches (or *modes*): 1) soundscape composition and 2) sound mapping (sonification). Ideally, the user should be able to choose which one he prefers to use.

4.6.1. Soundscape composition

In this *mode* the collected data (audio) from each user is received by his/hers mobile device. The volume of each recording reproduction varies according to specific criteria such as geographic distance (from the user), friendship relevance (strong or weak connection) or network activity. The result is a mixture of background recordings.

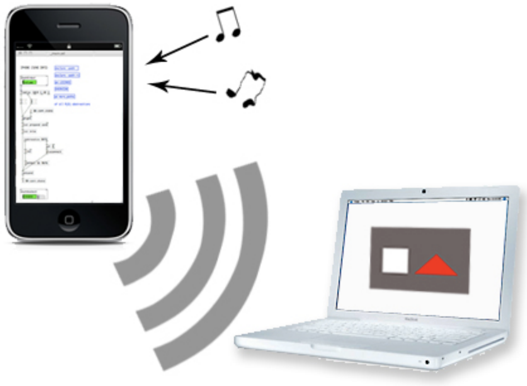


Figure 4 – Framework using an IPhone and a Macbook.

4.6.2. Sound mapping (sonification)

Sound mapping mode represents a sonification of the graphical displayed data. The process consists in associating different sound classifications (retrieved from recordings) to sound sources (synthesized sound; sampled sounds) and manipulate them according to the same criteria used in soundscape mode. Through this process, each user can personalize his "friendship composition" by defining his own set of sounds, mappings and manipulations (sound processing, volume, synthesis parameters, etc.).

5. EXPERIMENTAL FRAMEWORK

In order to test some basic procedures regarding audio capture and data transmission with IPhone, we developed an experimental framework application using an IPhone, a Macbook, Pure Data (vanilla) software and RjDj platform. The framework includes two different patches: one sending data (running on IPhone trough a RjDj scene) another receiving and displaying (running on a Macbook computer). The patch responsible for sending data, analyzes sound input and classifies it as pitched or non-pitched (using *fiddle~,* an PD object by Miller Puckette).

This information is then sent to the computer via UDP protocol (using *netsend*), where the "receiver patch" displays (using *GEM*) a triangle or a square, accordingly. The test clarified preliminary uncertainties: IPhone supports basic audio analysis and data transmission (without significant latency for us) over Internet. The experiment took place during a RjDj sprint; no further improvements have been made to date.

6. FUTURE WORK

For the future, our goal is to implement this yet conceptual project and extend the underlying paradigm through geo-location to specific places, in addition to people.

[1] Although we consider automatic classification a non-trivial function, at this point we are not able to describe the technology implied, nor the adopted sound classification method.

7. ACKNOWLEDGMENT

Hurly-Burly concept was conceived during the workshop *Interface Design for Mobile Applications* that took place at the Future Places digital media festival, organized by Oporto University at October 2009. Ana Parada, João Cordeiro and Kateina Marková composed the group responsible for the core idea. Figure 2 (right) was design by Katerina Marková.

8. REFERENCES

[1] M. Chion, *Audio-Vision. Sound On Screen.* (C. Gorbman, Ed., Eng. Tr.) New York: Columbia University Press, 1994.

[2] M. Castells, *A Sociedade em Rede. A Era da Informação: Economia, Sociedade e Cultura*, 3rd ed., vol. 1, Lisbon: Fundação Calouste Gulbenkian, 2007.

[3] N. Christakis & J. Fowler, *Connected: The Amazing Power of Social Networks and How They Shape Our Lives.* London: Harper Press, 2009

[4] V. Cosenza, (retrieved at 03/02/2010) http://www.vincos.it/world-map-of-social-networks/

[5] R. M. Schafer, *The soundscape: our sonic environment and the tuning of the world* (1994 ed.). Destiny Books, 1977.

[6] J. Cohen, "Out to Lunch: Further Adventures Monitoring Background Activity." in G. Kramer, & S. Smith (Ed.), *Proceedings of the Second International Conference on Auditory Display*, 1994, pp. 15 - 20.

GROOVING FACTORY – BOTTLENECK CONTROL IN PRODUCTION LOGISTICS THROUGH AUDITORY DISPLAY

Revealing Work Content Overloads

Katja Windt

Jacobs University Bremen gGmbH
School of Engineering and Science
Workgroup Global Production Logistics
P.O. Box 750 561, 28725 Bremen, Germany
k.windt@jacobs-university.de

Michael Iber

Jacobs University Bremen gGmbH
School of Engineering and Science
Workgroup Global Production Logistics
P.O. Box 750 561, 28725 Bremen, Germany
m.iber@jacobs-university.de

Julian Klein

Institute for artistic research / Radialsystem V
Holzmarktstrasse 33
D-10243 Berlin, Germany
JulianKlein@artistic-research.de

ABSTRACT

Grooving Factory is the name of an interdisciplinary research project in the fields of production logistics engineering and auditory display. It aims to reveal bottlenecks in industrial productions and to improve the achievement of logistic targets by using sonification in production planning and control (PPC). Since data sets derived from production processes are time related, processes in production can be displayed as oscillating complex sounds, e.g. via additive synthesis. In this study, the feedback data of operations at 33 workstations of a circuit board manufactory served as a model for auditory display. The workload of the workstations was compared to their actual performance, which indicated their work in process (WIP) level, i.e. the balance of their input and output. The results of the auditory display were compared to WIP related bottlenecks identified by the bottleneck oriented logistic analysis including logistic operating curves. The research project includes the development of a new PPC method realized as a prototype in a software tool.

1. INTRODUCTION

1.1. Complexity of logistic targets

The analysis of different types of bottlenecks in the production workflow [1] is a basic principle in preposition to the parameter setting and application of production planning and control methods. These bottleneck types relate to the four logistic main targets, which include the achievement of short throughput times, high delivery reliability, adequate degree of capacity utilization and low level of work in process (WIP). Some of these targets are contradictive and the identification of an ideal balance – which comprises the best possible achievement for the privileged target with the least drawbacks on the remaining targets - has become a highly complex task according to the increased requirements of flexibility and product variability in the global market.

1.2. Auditory Display and complex data

As a fairly young discipline at first defined in 1992 [2] auditory display is the non-speech acoustic representation of information within the human hearing range. One of its main features is the identification of single events in complex data, which would get lost in methods using rough resolutions like graphic displays or averaged calculations. In scientific fields such as space physics or stock market analysis, auditory display has been proved as a superior analysis tool in certain test arrangements. The research for an analysis tool based on auditory display to cope with the complexity of logistic data therefore seems a promising approach, especially since (once established) auditory displays provide results very quickly.

1.3. State of the art

The more it is surprising that apparently there have been only very few research projects in the field of production logistics involving auditory display: In the so named ARKola project, Gaver, Smith et al [3] developed a simulated soft drink factory, which consisted of an interconnected series of nine machines, eight of which were under user control. The processes of the machines were displayed by auditory icons representing the semantics of the specific machines, e.g. the bottle dispenser producing the sound of clanking bottles. These sounds were audible to all users to give them a forecast of what to expect as well as a report on the impact of their actions.

A further approach was conducted by Alicke [4]: According to him both, music and logistics rely on the order of sequences and in logistics, deviations from the sequencing rule “first in first out” (FIFO) of orders waiting in front of a machine influence the length of the lead time. He assumes that logistics can learn from musical principles and claims that there must be a common line between aesthetics and functionality. As an example Alicke links the logistic processes of a container terminal to a given piece of music, “Satin Doll” by Duke Ellington, and compares patterns of melodic motives and logistic sequences. In spite of this interesting approach, Alicke does not complete his research by developing a method for production planning and control.

Whereas the ARKola project emphasized on interactive aspects of Auditory Display and Alicke's approach focused on arguable theses about the portability of musical structure for general purposes, this project tackles a third route, which is more related to approaches based on non-linear dynamics [5] investigating the phenomena of oscillation and synchronization.

1.4. Research targets

The overall target of the project is the development of a production planning and control method for workflows in production logistics based on auditory display. To accomplish this, a methodic transfer of production based data into auditory display has to be explored and methods to identify process related bottlenecks have to be investigated and established. The results will be compared with the ones obtained from the Bottleneck Oriented Logistic Analysis (BOLA) [6]. This comparison will also serve as a first validation step for analyzing production logistic data based on auditory display. Since there has been no fundamental research yet combining logistics engineering and auditory display, the setup of the project, which includes the development of prototyping software, started by very basic means to be sufficiently flexible to adjust to upcoming results. This included impulse based and sinusoidal based sonifications as well as several mapping strategies [7].

Section 2 of this paper introduces the data set used and gives insight into the BOLA. Section 3 gives general considerations about the mapping of production data to auditory display, which will be specified in Section 4 on the practical experience. Section 5 gives insight into the actual results of the auditory display based on the exposure of work content. Section 6 concludes the state of research and provides an outlook on future steps.

2. BOTTLENECK ORIENTED LOGISTIC ANALYSIS

2.1. DATA SET OF CIRCUIT BOARD MANUFACTORY

The data set chosen for the research originates from a circuit board manufactory (figure 1).

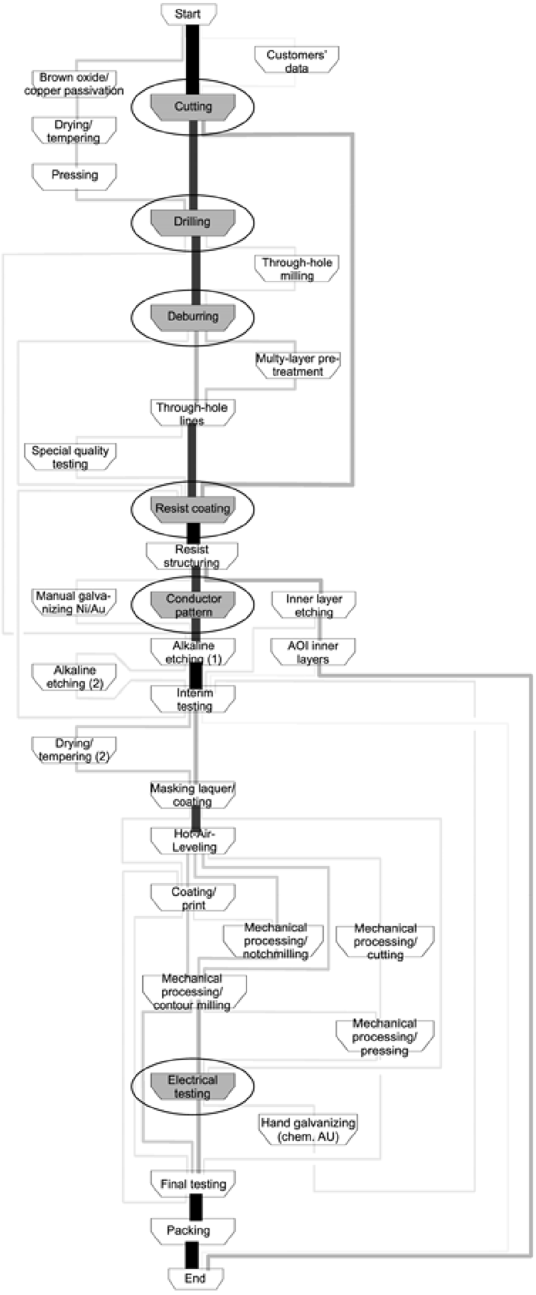


Figure 1: Workflow of circuit board manufactory. The individual workstations are displayed as funnels. Encircled workstations are bottlenecks identified by BOLA [6].

It monitors the workflow of an evaluation period of five month representing the processing of 4270 orders, each of which running through up to 31 operations at different workstations. The recorded data include information about the individual workstations with their daily capacities, order related information including the lot size of orders, as well as operation related data like operation sequence, work content, end of operation, or technology dependent waiting time. The operational feedback data was monitored on a timely precision in seconds.

The motivation to select precisely this data set was that it has been extensively analyzed and documented by Peter Nyhuis and Hans-Peter Wiendahl [6] as an example of an application of Bottleneck Oriented Logistic Analysis including the Logistic Operating Curves Theory (LOC). The results derived from auditory display thus can be reliably compared and verified.

2.2. Bottleneck Oriented Logistic Analysis (BOLA) including Logistic Operating Curves (LOC)

The basic elements of the Bottleneck Oriented Logistic Analysis (BOLA) are the funnel model, which describes the input and output relation at individual workstations (figure 2a), and the 2-dimensional throughput element, which integrates the several processes before the operation and the operation itself including the setup time. The work content of the order (measured in the time needed for the operation) is represented by the height of the element.

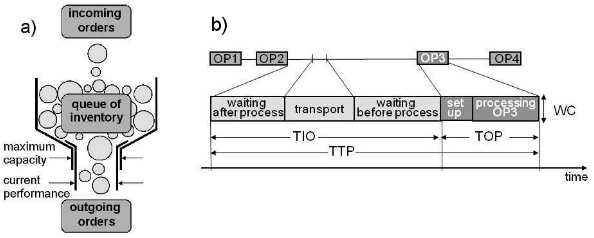


Figure 2: a) the funnel model describes the input/output relation at a workstation. b) the 2-dimensional throughput element describes the individual order at a workstation with OPx: orders, TIO: interoperation time [shop calendar days SCD], TOP: operation time [SCD], TTP: throughput time [SCD], WC: work content [h] [6].

The BOLA analyzes the production workflow from a order and resource oriented view (figure 3d). In a first step all individual workstations are statistically evaluated and ranked according to their mean output rate (= mean performance), their work-in-process level (WIP), as well as the due date reliability and the throughput times of the passing orders (3a, b).

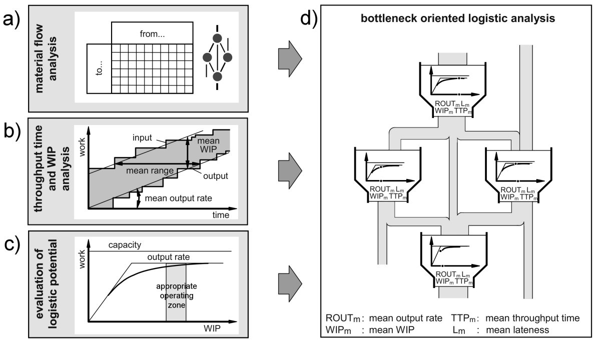


Figure 3: Bottleneck Oriented Logistic Analysis collects statistical information about all workstations (a), analyzes their throughput times and work-in-process-levels (b) and applies the Logistic Operating Curves Theory to balance parameters in respect to the target achievements [6].

That way workstations with high WIP can be identified and represents an initial step towards identification of bottleneck worksystems. Also known as inventory, WIP describes the balance between incoming and outgoing orders at a workstation, in other words: it is the composite work content (WC) of all orders at a workstation at a time. It is notable that the WC of incoming orders is measured by the time needed for processing (refer also to fig. 2b). This is due to varying durations of the operations needed for various product types and varying lot-sizes. WC is calculated in hours:

$$WC = (tp \cdot LS + ts)/60, \qquad (1)$$

whereas tp is the processing time per piece for the specific product type (in minutes), LS is the lot size (number of pieces in order) and ts is the setup time of the workstation (in minutes).

According to the funnel formula [6] high WIP causes long mean throughput times (which can be considered equal to the range in steady state systems) and therefore long delivery times in the workflow. A workstation with high WIP will consequently be identified as a WIP related bottleneck of first order. Too low WIP on the other hand results in a poor performance causing a utilization-related bottleneck. Therefore WIP should be kept at a safe minimum level (which in practice is 2,5 times the calculated ideal minimum WIP [6] level assuming that the worksystem is not outer margin) avoiding the risk of performance losses and at the same time keeping throughput times as short as possible. The appropriate operating zone can be calculated by the Logistic Operating Curves Theory [6], which indicates the influence of the respective parameters on each other (figure 4). Furthermore, in BOLA one aim is to identify the so called order-related throughput time bottleneck work systems. Those work systems are in general characterized by a high number of orders being processed in combination with a long mean throughput time.

We will concentrate in the following as a first validation step of mapping production logistics feedback data to identify WIP bottlenecks as an overload situation of a work systems.

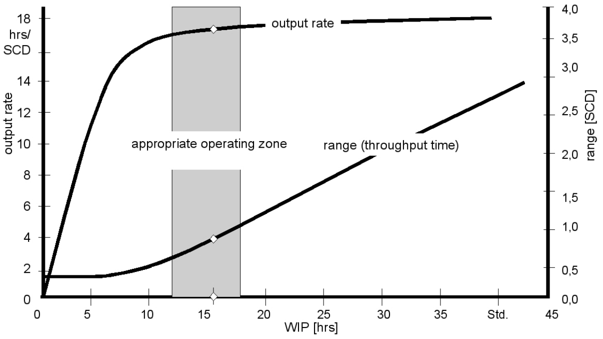


Figure 4: Ideal minimum WIP in relation to output rate and range [6]. A larger WIP will increase the throughput time without increasing the output rate significantly. Whereas a very low WIP would affect the performance significantly.

3. GENERAL CONSIDERATIONS CONCERNING DATA MAPPING IN AUDITORY DISPLAY

Operational feedback data of production processes is time-based as is any approach using auditory display. The immediate mapping of workflow data into sounds representing the length of the operations is self-evident. The content of the operations, i.e. the amount of pieces produced over the length of time can be represented as pitch:

$$f(i) = 1 / (TOP / x), \qquad (2)$$

whereas i is the operation, f is the displayed frequency, TOP is the total operation time, x is the number of pieces produced, respectively the lot size.

The simplest possible scenario of a production chain is a linear arrangement of machines with operations of identical lot size and processing time without any waiting time in between (Figure 5a). In a representation as auditory display each workstation (OPx) produces tones of identical frequency (f). If on the other hand e.g. the second workstation (OP2) needed double the time for an operation, it would sound in half the frequency (f/2) and pause the sound of subsequent machines sequentially. Additionally, a queue will build up in front of OP2, the work-in-process level (WIP) will increase.

Figure 5b provides a solution to this problem by extending the second workstation to a group of two workplaces. Depending on whether the auditory display represents workstations or individual workplaces (where each workplace is represented by its own sound), it will either sound unison with the other workstations or the two workplaces of OP2 sound individually at the lower octave to the other workplaces. In both cases always all the workplaces will sound and there will be no queue.

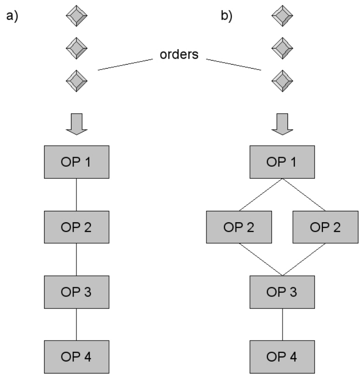


Figure 5: Examples of linear workflow, whereas OPx are sequential workstations.

Due to the large variety of product types and the complexity of routes all products may take in the workflow, “not sounding” of a workstation cannot be taken as an indication of bottlenecks. Pauses of workstations outside the main workflow, which are only used for special product types, happen regularly. The building up of queues in front of workstations is a much better indicator for WIP -related bottlenecks in this matter. It therefore should be most efficient to observe the waiting time (interoperation time) of the individual orders before their processing, as well as the work in process (WIP) at the workstations.

4. DATA MAPPING APPROACHES

4.1. Development of software interface

The prototyping software developed for the auditory display of production data is based on a combination of the Pure-Data graphical programming language [8], which is used for the audio related parts (including additive synthesis and a variable surround setting between 1 and 16 output channels) and Python [9] for data reconditioning. It allows the auditory display of any combination of operations at workstations as well as the workflows of orders. Since sounds in the realm of the human hearing range are crucial for sonifications, several options are provided to adjust the data (figure 6) in this regard.

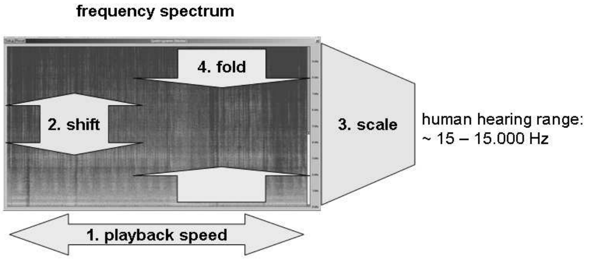


Figure 6. The software interface offers four options to arrange the frequency ranges to the human hearing range: 1st change of playback speed. 2nd shifting the frequency of a data section. 3rd linear and logarithmic scaling of all frequencies to human hearing range. 4th folding frequencies outside the audible range to their closed octave within the hearing range.

4.2. Occurrences of Inconsistencies in the Data

The research on fundamental data-to-sound mappings [7] revealed various inconsistencies of the data set with high time resolution. Accuracy of operational feedback data is a known problem in logistic analyses. Although there have been many improvements in the last years, many recordings of operations are still executed by humans. The data set used by Nyhuis and Wiendahl for the BOLA was reduced to daily precision and does not contain any inconsistencies. Additionally it should be considered that BOLA relies in general on averaging values over an evaluation period of some weeks, and therefore detailed information just will be weighted. On the other hand it is one of the advantages of auditory display to be able to deal with complex data on a high resolute time scale and it is one of the targets of the project to explore, if analysis based on auditory display will lead to similar results but obtained (eventually) in faster time and offering potentials towards complexity driven analysis.

The unexpected sounding results of the software displaying on the high resolute time scale revealed that up to some 60 processes were registered simultaneously at individual workplaces, where only one process – displayed as a monophonic signal - was considered at a time, which is due to the workers at the workplaces use to collect several orders at once in order not to loose time. That required an integration of the conflicting parallel processed orders. Therefore the overlaying sections of the conflicting orders were compounded to so to speak "virtual orders", which represent all pieces in process at a certain time span.

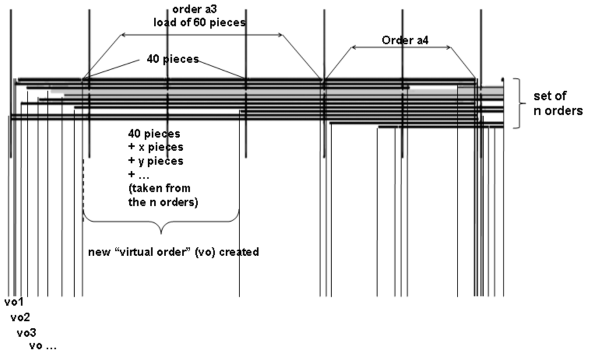


Figure 7: Overlaying orders are compounded in "virtual orders".

This method also allows the display of interoperation times of orders in front of workstations, which naturally overlay, as well as their work in process (WIP).

4.3. Results of directly mapped auditory display

After most of the data inconsistencies concerning the reported schedule and operational sequences had been corrected experiments with sonifications based on the implemented direct mapping strategies, meaning that all workstations would display any of their operations, proved to be far too complex to spot out any irregularities in the production workflow. Also the consideration of single pieces produced appeared unrealistic, since no reliable production, setup and waiting times for the various product types could be derived from the data, further process related explanation on data entrys is difficult to get. Therefore the idea of a high-resolution time scale was put on hold in favor to an investigation of feedback data with daily precision.

5. AUDITORY DISPLAY OF WORK CONTENT

In this setting not the inaccurately measured throughput time is taken into consideration as an indicator for the operation time but the WC, which relies on planned data. If the compound WC at a workstation is larger than its given capacity this may indicate a throughput related bottleneck. For this purpose the WC and the technical waiting time of an order are calculated backwards from the monitored output day taking the daily capacity of the workstation into respect. That means that, if WC is larger than the workstation's capacity, it will be distributed over the according number of days. The adjusted software integrates the WC of all orders at each workstation on a daily basis and allows the following options for display:

1. Absolute display of all operations at discrete workstations, pitch representing the WC in minutes.
2. Absolute display of WC overload at discrete workstations.
3. Relative display of WC overload, amount of overload in minutes represented by pitch.

The unfiltered sonification of the WC of all data sets (option 1) again sounds too complex to give any evidence of bottlenecks (sound example 1).

The auditory display of the relative WC overload (option 3) is more revealing. In a surround setting, where the workstations are panned between -45° and 45° according to their position in the workflow, a workstation around 0° can be spotted building up high frequencies as can be heard in sound example 2, which displays the overload of all workstations and 3, which displays the concerning workstation "multilayer pretreatment" alone (figures 8).

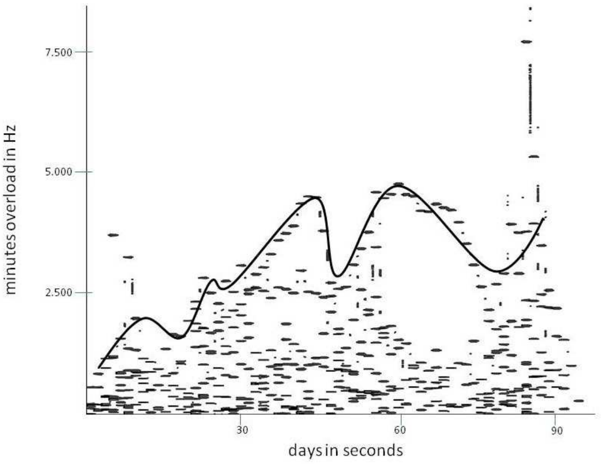


Figure 8: Visual representation with SPEAR [10] of auditory display of performance overload of all workstations over the evaluation period. In the auditory display, the constant peaks can easily be allocated to the workstation "multilayer pretreatment".

Among the bottlenecks spotted by BOLA (figure 1), only the resist coating workstation displays overload worth mentioning in the sonification.

Another interesting, yet difficult to interpret result is the auditory display of the two succeeding workstations named "resist coating" and "resist structuring". As can be heard in sound example 4 (figure 9), which again displays the relative overload, "resist structuring" (45° panning) "joins in" into the permanent overload of "resist coating" (-45° panning) with frequencies close to the ones displayed by "resist coating" synchronously most of the times with slightly higher pitches, i.e. higher overload. The fact that "resist structuring" only displays short sequences and then becomes mute again indicates that the workstation is capable of dealing with the workload over a specific period of time and therefore is not necessarily a bottleneck as defined by BOLA, which treats only longer periods of time. But it certainly can be considered a short term hindering of the workflow, which impact has to be further investigated.

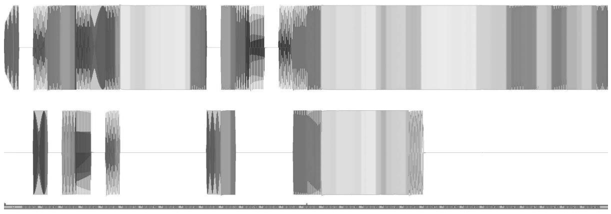


Figure 9: Resist coating (top) and resist structuring workstations visually displayed by Samplitude's comparasonics [11]. The grayscale brightness of the waveform represents the frequency.

6. CONCLUSION AND OUTLOOK

The actual study has not yet achieve the target to identify identically bottlenecks as the BOLA. But nevertheless these first results with auditory displaying the compound WC, which is an equivalent to the work in process (WIP) at the individual workstations raise a number of questions to be answered in the follow-up steps of research: 1st Why do most of the workstations, which are identified as possible bottlenecks by significant WC overload, differ from the ones identified by BOLA? 2nd What is the explanation behind the partial "joining in" into the melodic lines of neighbor workstations? How can their interplay be interpreted and manipulated? 3rd Is daily precision of data sufficient enough to investigate the impact of sequential changes of orders, which are executed e.g. to shorten setup times or to prioritize urgent orders? After these questions have been answered, the Grooving Factory should be capable not only to spot bottlenecks in a workflow immediately but also to precisely analyze the interferences of various scenarios in production planning and control.

7. ACKNOWLEDGMENT

The research of Prof. Dr.-Ing. Katja Windt is supported by the Alfried Krupp Prize for Young University Teachers of the Krupp Foundation. The Grooving Factory Project was initiated and supported by the research group "Rhythm" of The Young Academy at the Berlin-Brandenburg Academy of Sciences and Humanities and the German Academy of Natural Scientists Leopoldina: www.diejungeakademie.de

8. REFERENCES

[1] Windt, K., 2001, Engpassorientierte Fremdvergabe in Produktionsnetzen. Düsseldorf: VDI Verlag GmbH.

[2] Kramer, G., 1994. Auditory Display. Sonification, Audification, and Auditory Interfaces. Reading: Addison-Wesley Publishing Company.

[3] Gaver, W., Smith, R., O'Shea, T., 1991, Effective Sounds in Complex Systems: The Arkola Simulation. Proceedings of CHI, New York: ACM.

[4] Alicke, K., 2004, Musikalische und logistische Reihenfolgeprobleme - eine vergleichende Darstellung, in: IM Die Fachzeitschrift für Information Management & Consulting, Nr. 3, 2004, pp. 73-78. GBI-Genios Deutsche Wirtschaftsdatenbank GmbH

[5] Scholz-Reiter, B. and Tervo, J. T. and Freitag, M J. T. , Freitag, M., 2006. Phase-synchronisation in continuous flow models of production networks. Bremen Institute of Industrial Technology and Applied Work Science in science-direct

[6] Nyhuis, P. and Wiendahl, H.-P., 2009. Fundamentals of Production Logistics: Theory, Tools and Applications. Berlin: Springer-Verlag.

[7] Windt, K., Iber, M. and Klein, J., 2009. Grooving Factory – bottleneck control in production logistics through auditory display. Step 1: strategies of data mapping. (under review)

[8] Puckette, M. S., 2009. Available from: http://crca.ucsd.edu/~msp/software.html [Accessed 6 May 2009]

[9] Python Programming Language, 2009. Available from http://www.python.org [Accessed 6 May 2009]

[10] Klingbeil, Michael, 2009. SPEAR Sinusoidal Partial Editing Analysis and Resynthesis. Available from http://www.klingbeil.com/spear [Accessed 1 Feb 2010]

[11] Magix Samplitude 11 Pro, 2010. Available from http://www.samplitude.de [Accessed 1 Feb 2010]

EFFECTIVE DESIGN OF AUDITORY DISPLAYS: COMPARING VARIOUS OCTAVE RANGES OF PITCH AND PANNING

Paul Ritchey, Lindsey Muse, Harry Nguyen, Ricky Burks and S. Camille Peres

Research on the Interaction between Humans and Machines Lab
University of Houston-Clear Lake
Houston, TX, USA
ritcheyp3517@uhcl.edu

ABSTRACT

There is a large volume of research on designing effective visual displays, however there is little empirical research informing basic design on auditory displays. With the decreasing size of hardware (e.g., hand-held devices) and the increasing amount of software available, auditory displays are viable option for communicating data in places that have limited space for visual displays and for eye-busy environments. Auditory graphs are auditory displays that map quantified data to acoustic dimensions, such as pitch and panning, to represent changes in data. In the present study, we investigate the octave range of pitch that most effectively represents the data in an auditory graph, as well as the effects of utilizing the acoustic dimension panning to give participants added temporal context. Significant results were found that support the use of panning. A significant interaction between the reported maximum temperatures and octave range, as well as a significant main effect was found for the type of statistic participants were asked to report (minimum value, maximum value, and average value), these results are discussed.

1. INTRODUCTION

With the explosion of new technologies in the twenty-first century, many people have to learn how to interpret a large volume of information presented through both visual and auditory displays. The implementation of auditory displays not only enhances the effectiveness of various technologies, but also is crucial for individuals with visual impairments, and for people working in eye-busy environments [1,2,3]. Because of this, it is important to investigate how sound patterns can augment and even replace visual displays in communicating quantitative information [4]. Auditory display designers are in need of basic rules and guidelines to reference when designing an effective display, rather than relying on intuition or replicating what has been used in the past. Sándor et al. [5], mentions that designers have several crucial decisions to make regarding what data to represent, how to represent the data, and what dimensions to use. As Barrass [6] pointed out, many of the designers working on new technologies have little understanding and knowledge about making an effective auditory display, so having some basic rules and guidelines would be beneficial so that they can be referenced when needed and reused to guide other projects.

Since the inception of the Graphical User Interface (GUI), design guidelines for visual displays have become relatively well established [1] and while strides have been made in the research of auditory displays, there remains a lack of empirical research to help guide the design process [4,7,8,9].

Peres [1] identified four categories in which auditory displays are used: (1) presenting information to visually impaired people (2) providing an additional information channel for people whose eyes are busy attending to a different task (3) alerting people to error or emergency states of a system, and (4) providing information via devices with small screens such as PDAs or cell phones that have limited ability to display visual information. For the purpose of this study, we focused on auditory graphs, an auditory display classified as data exploration that uses sound to represent quantitative data [8,10]. An example of an auditory graph is when Flowers [4] mapped temperature ranges for each month to a pitch. Flowers [11] found that weather data makes for a compelling auditory display format partly because of its sequential observation across time. Auditory graphs can grant the visually impaired the same benefits that visual graphs offer to sighted individuals, such as a concise summary of data, and trend analysis [12]. Research has also found that the use of auditory graphs is an opportunity to teach statistical concepts like central tendency, variability and shapes of distributions to both sighted and visually impaired individuals [4]. Further, Peres and Lane [13] used sound dimensions to communicate statistical information usually contained in box plots.

Much of the research on auditory displays has examined the impact of these displays on attention, cognitive load, and discrimination from distracters and less on the structure and guidelines for making an effective display [14,15,16,17]. Although there is some research that has investigated acoustic dimensions in attempt to determine which dimensions are most effective in the interpretation of auditory displays, more is still needed. The acoustic dimensions previously studied include: pitch, loudness, timbre, tempo, and panning [1]. In the research presented here, we chose to focus on two dimensions: pitch and panning. Pitch is the most widely used dimension but its level of effectiveness varies depending on the context of the experiment, as a result, more research is needed to further design guidelines [10,13,18,19]. According to Flowers [20], "mapping pitch height (log frequency) to numeric magnitude affords perception of function shape or data profile changes, even for relatively untrained observers". Walker and Nees [10] found pitch to be a good option for representing temperature data. Until the current study the research conducted on pitch has not tested which pitch ranges result in the most accurate interpretation of the data. Flowers [4] stated that additional research could help determine optimal pitch ranges for auditory graphs.

Panning is a mapping technique that presents sounds in a manner so the listener perceives the sounds spatially. It is conceivable that if this dimension were used redundantly with time to represent information on an auditory graph that people's comprehension of the graph could be improved [10]. However, panning has been studied even less than pitch and thus there is currently no empirical data to support this position. Peres and Lane [13] found pitch to be only slightly more effective than pitch mapped redundantly with panning in the presentation of box plots. Interesting, users preferred the redundant condition (pitch with panning) strongly to pitch or panning alone. However, panning mapped redundantly with other dimensions has not been investigated empirically.

In order to test the effects on performance when pitch and panning are used for auditory displays, we used these two dimensions to build different auditory graphs that display temperature data. Specifically we wanted to explore how octave range and panning impacted people's ability to interpret different statistical elements in the data, e.g., trend, mean, range, etc.

2. METHOD

2.1. Participants

59 participants (43 females and 16 males) were recruited for this study. The mean age was 30.34 (SD= 10.56). Each subject served in a single experiment session. All participants were screened by self-report to ensure normal or corrected-to-normal hearing.

2.2. Stimuli

Twelve auditory graphs were created from two sets of temperature data using Sonification Sandbox 5. Each of the two data sets used consists of the recorded high temperatures in a geographical area for a one month period, or 30 days, for a total of 30 data points. The weather data used for the data sets was taken from the Weather Channel website [21]. All of the auditory graphs last 20 seconds and consist of 30 discrete sounds—each sound representing a single data point in the data set. To control for practice effects, each participant was exposed to two auditory graphs, each constructed using a different data set.

Six "pitch-only" auditory graphs were made from a single data set by mapping the temperature data value to pitch. The chromatic scale was used so each octave consists of 12 pitches. For example, an auditory graph made from the first octave range listed below (C3) would consist of the following 12 pitches: C3, C#3, D3, D#3, E3, F3, F#3, G3, G#3, A3, A#3, B3.

For the current study, different octave "ranges" were used to map the temperature data to sound. The octaves used are listed below. Each octave contains 12-notes and the number in parentheses indicates the frequency of the first note in the octave:

- C2 (65.406 Hz)
- C3 (130.813 Hz)
- C4 (261.626 Hz)
- C5 (523.251 Hz)

Thus, there were three different "octave range" designs: a single octave designs (2 of the designs), two-octave designs (3 of the designs), and one 4-octave design. The composition of each of these is listed below:

Single Octave ranges

- C3
- C4

Two Octave ranges

- C2 and C3
- C3 and C4
- C4 and C5

Four Octave range

- C2, C3, C4 and C5

These octave ranges were chosen because they are appropriately varied for the scope of this experiment as there is little consensus with regard to the best pitch ranges to use for auditory graphs [4]. Additionally, the octave ranges chosen are neither too high nor too low, so as to avoid having two distinct pitches that a participant with normal hearing might not be able to discern as different.

There were two data sets used to create the panning and non-panning auditory graphs. One set was mapped to the pitch-with-panning auditory graphs and the other was mapped to the pitch-only auditory graphs. The pitch-with-panning auditory graphs use the same pitch ranges listed above, but also had panning redundantly representing time. To create a "panning effect" the loudness of each sound was manipulated so that the sounds seemed to move through the listener's skull from left to right. The sound representing the high temperature for the first day of the month, the first data point, seems to emanate from a source at the left ear, the sound representing the middle data point seems to emanate from a source between the participants ears, and the sound representing the last point seems to emanate from a source near the right ear.

2.3. Procedure

Participants were given a brief orientation about the auditory graphs they would hear and the surveys they would complete. The orientation was presented on a HTML document that the participants progressed through with the researcher providing additional instructions.

After participants finished the orientation they were prompted with a “tuning page” that preceded each of the two auditory graphs. The purpose of the tuning page was to inform participants of the specific temperature-pitch relationship that was used in the auditory graph. Specifically, participants listened to a sound clip in which a series of three different tones were played and the temperature values of each tone were concurrently displayed visually on the computer screen. This temperature-pitch relation was the same used for both of the auditory graphs the participant hears. The tuning page also provided information about the measures participants completed after listening to the auditory graph.

Each participant listened to two auditory graphs, one pitch-only auditory graph and one pitch-with-panning auditory graph. Both of these auditory graphs used the same octave range. After listening to the first auditory graph participants gave responses about the statistical properties of the data they heard (minimum, maximum and average) and their subjective ratings of the sounds. After providing information about the first graph, participants would then be presented with the tuning page for the next auditory graph. They would listen to a corresponding auditory graph that was followed by the same tasks and subjective ratings. After the second rating was completed, participants were debriefed. For all participants the order in which participants listened to the auditory graphs (i.e., panning, no panning) was counterbalanced.

2.4. Measures

Participants were given two short measures to complete after listening to each auditory graph. The performance measures were comprised of several questions used to determine the participants’ understanding of the data presented in the auditory graph. Specifically, participants were asked to provide estimates of the minimum and the maximum temperature, as well as the mean temperature. A preference measure was also collected to determine the participants' subjective of how enjoyable, helpful, or distracting they found aspects of the auditory graphs to be.

To measure participants’ performance, error scores were created by taking the difference of participants’ responses on the performance measure and the respective true values of the temperature data sets. An error score closer to zero indicates a more accurate response. A negative error score indicates the participant reported a value lower than the actual value.

3. RESULTS

Error scores were analyzed in a 6 x 2 x 3 (6 octave ranges x 2 panning levels x 3 performance measures) factorial ANOVA. A significant main effect was found for panning, $F(1, 52) = 4.513$, $p = 0.038$. Overall accuracy of participants’ responses on the performance measure was greater for auditory graphs with panning. As shown in Figure 1, error scores for the non-panning auditory graphs had were worse (mean of -5.54) than those for the panning auditory graphs (mean of -3.153).

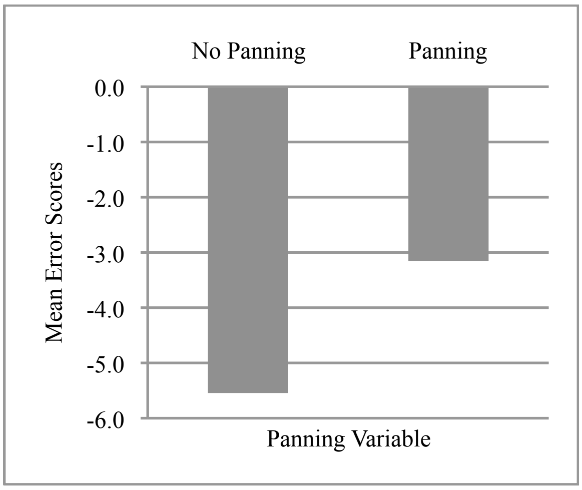


Figure 1: Mean error scores for performance responses for auditory graphs without panning and with panning.

A significant main effect was also found for performance measures, $F(2,104) = 13.207$, $p < 0.001$. Participants were more accurate in reporting the minimum temperature than the maximum or average temperature across all conditions. The mean of the error scores for reported minimum temperatures was 0.997, maximum was -7.445 and average was -6.598. This main effect can be seen in Figure 2.

Figure 2 also illustrates the significant interaction between octave range and performance measures, $F(10, 104) = 2.335$, $p = 0.016$. Participants’ reported maximum temperatures were more accurate when they listened to an auditory graph with octaves ranges of C2-C3, C3-C4, C4-C5, or C2-C5. They were the least accurate in reporting the maximum temperature for the single octave ranges of C3 and C4. No significant main effect was found for octave range and all other interactions were not significant (p s > 0.15).

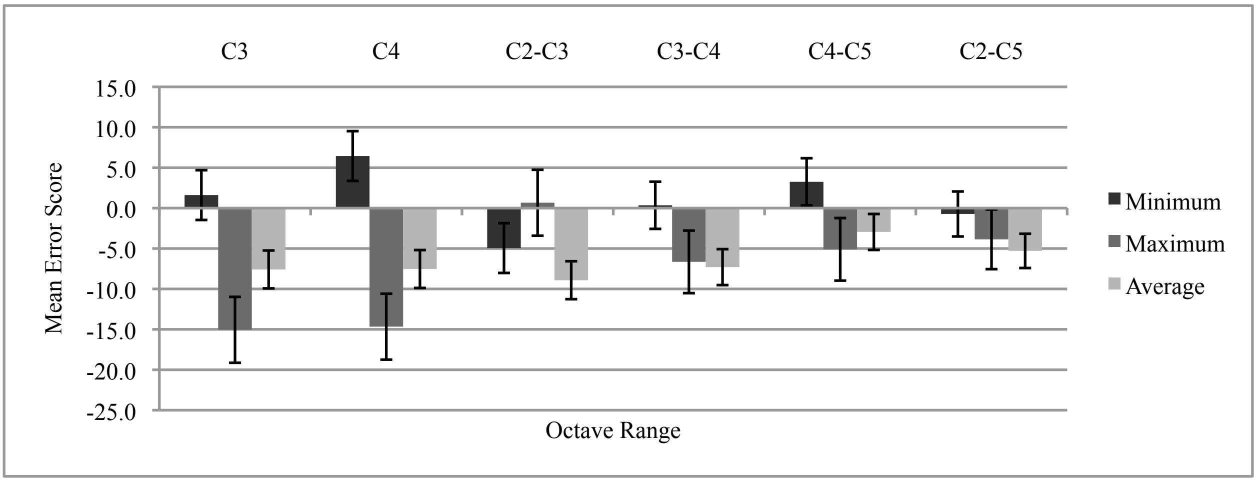


Figure 2: Mean error scores for reported minimum, maximum, and average for each octave range.

4. DISCUSSION

The significant main effect for panning suggests that panning yields additional context to the auditory graph when panning redundantly represented time. Panning possibly added relevant and useful information enabling the participants to more accurately interpret and understand the auditory graphs, irrespective of the octave ranges the auditory graph utilized. Auditory graph designers should consider utilizing panning to assist their users in interpreting the passage of time or anticipating the length of an auditory graph. More in depth research needs to be conducted with a focus on panning, as this research did not focus primarily on different applications of panning.

Across all conditions participants were more accurate when reporting the minimum temperature then either the maximum temperature or the average temperature. This is likely a function of the fact that participants generally underestimated the values they were asked to report (see Figures 1 and 2). Further, their accuracy could actually be due to the design of the “tuning page.” The tuning page was used to introduce participants to the specific temperature-pitch relationship and one of the sample values used was 40. The minimum value in each data set was 33 and so participants were more familiar with a pitch near the minimum than the maximum. Similarly, the average required a much more complicated judgment to determine. It is also possible that low pitches are more distinguishable and understandable than higher pitches in the context of auditory graphs.

Participants were more accurate in reporting the maximum temperature for the auditory graphs that included more than just one octave range (e.g. C2-C3, C3-C4, C4-C5, C2-C5). This suggests that auditory graph utilizing multiple pitch ranges can improve performances in circumstances when users must report the maximum values of the data set.

The results from this study are important and intriguing, however, the effects and application of panning need more attention. Future research should more thoroughly investigate the effects of different octave ranges on interpreting auditory graphs, focusing on larger ranges. While this research suggests that larger pitch ranges may allow for more accurate interpretation of certain statistics,, the lack of a main effect of octave range (e.g. C2-C3, C3-C4, C4-C5 or C2-C5) may indicate that auditory graph designers have more freedom to choose what pitch ranges best fit their specific design.

5. REFERENCES

[1] S. C. Peres, V. Best, D. Brock, B. Shinn-Cunningham, C. Frauenberger, T. Hermann, J. Neuhoff, L. Nickerson, and T. Stockman, "Auditory Displays," in *HCI Beyond the GUI: The Human Factors of Non-traditional Interfaces*, P. Kortum, Ed.: Morgan Kaufman, 2008.

[2] J. H. Flowers, D. C. Buhman, and K. D. Turnage, "Cross-modal equivalence of visual and auditory scatterplots for exploring bivariate data samples," *Human Factors,* vol. 39, pp. 341-351, 1997.

[3] B. N. Walker and D. M. Lane, "Psychophysical scaling of sonification mappings: A comparison of visually impaired and sighted listeners," in *International Conference on Auditory Displays*, Espoo, Finland, 2001.

[4] J. H. Flowers and T. A. Hauer, "Musical versus visual graphs: Cross-modal equivalence in perception

of time series data," *Human Factors,* vol. 37, pp. 553-569, 1995.

[5] A. Sándor, S. C. Peres, and D. M. Lane, "Redundant Sound Dimensions in Auditory Displays: Classical Integrality Increases Performance," in preparation.

[6] S. Barrass, "Sonification design patterns," in *Proc. of the 9th Int. Conf. on Auditory Display (ICAD2003)*, Boston, MA, 2003, pp. 170-175.

[7] J. Anderson, "Creating an empirical framework for sonification design," in *Proc. of the 11th Int. Conf. on Auditory Display (ICAD2005)*, Limerick, Ireland, 2005, pp. 393-397.

[8] S. C. Peres and D. M. Lane, "Auditory Graphs: The effects of redundant dimensions and divided attention," in *Proc. of the 11th Int. Conf. on Auditory Display (ICAD2005),* Limerick, Ireland 2005, pp. 169-174.

[9] P. Sanderson, J. Anderson, and M. Watson, "Extending Ecological Interface Design to Auditory Displays," in *Proc. of the 2000 Annu. Conf. of the Computer-Human Interaction Special Interest Group (CHISIG) of the Ergonom. Soc. of Australia (OzCHI2000)*, Sydney, Australia, 2000, pp. 259-266.

[10] B. N. Walker and M. A. Nees, "An agenda for research and development of multimodal graphs," in *Proc. of the 11th Int. Conf. on Auditory Display (ICAD2005),* Limerick, Ireland, 2005.

[11] J. H. Flowers, L. E. Whitwer, D. C. Grafel, and C. A. Kotan, "Sonification of daily weather records: Issues of perception, attention and memory in design choices," in *Proc. of the 7th Int. Conf. on Auditory Display (ICAD2001)*, Espoo, Finland, 2001, pp. 222-226.

[12] S. M. Kosslyn, "The Psychology of Visual Displays," *Investigative Radiology,* vol. 24, pp. 417-419, 1989.

[13] S. C. Peres and D. M. Lane, "Sonification of statistical graphs," in *Proc. of the 9th Int. Conf. on Auditory Display (ICAD2003)*, Boston, MA 2003, pp. 157-160.

[14] J. Anderson and P. Sanderson, "Sonification Design for Complex Work Domains: Dimensions and Distractors," J. Experimental Psychology: Appl., vol. 15, pp. 183-198, 2009.

[15] J. G. Neuhoff, J. Wayand, and G. Kramer, "Pitch and loudness interact in auditory displays: Can the data get lost in the map?" J. Experimental Psychology: Appl., vol. 8, pp. 17-25, 2002.

[16] M. Watson and P. Sanderson, "Respiratory sonification helps anaesthetists timeshare patient monitoring with other tasks," in Proc. of OZCHI2001 (OzCHI01), Perth, Australia, 2001, pp. 175-180.

[17] M. Watson and P. Sanderson, "Sonification supports eyes-free respiratory monitoring and task time-sharing," Human Factors, vol. 46, pp. 497-517, 2004.

[18] P. Kortum and S. C. Peres, "An exploration of the use of complete songs as auditory progress bars," in Proc of Human Factors and Ergonomics Society 50th Annual Meeting, San Francisco, CA, 2006, pp. 2071-2075.

[19] P. Kortum, S. C. Peres, B. Knott, and R. Bushey, "The effect of auditory progress bars on consumer's estimation of telephone wait time," in Proc. of Human Factors and Ergonomics Society 49th Annual Meeting, Orlando, FL, 2005, pp. 628-632.

[20] J. H. Flowers, "Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions," in Proc. of the 11th Int. Conf. on Auditory Display (ICAD2005), Limerick, Ireland, 2005, pp. 1-4.

[21] http://www.weather.com

DISTRACTING EFFECTS OF AUDITORY WARNINGS ON EXPERIENCED DRIVERS

Johan Fagerlönn

Interactive Institute
Sonic Studio
Acusticum 4, 94128 Piteå, Sweden
johan.fagerlonn@tii.se

ABSTRACT

A range of In-Vehicle Information Systems are currently developed and implemented in trucks to warn drivers about road dangers and vehicle failures. Systems often make use of conventional repetitive auditory warnings to catch attention. In a critical driving situation it might be tempting to use signals that express very high levels of urgency. However, previous studies have shown that more urgent alerts can have a negative impact on the listeners' affective state. A simulator experiment was conducted to examine how urgent warnings could impact the affective state of experienced truck drivers, and their response performance to an unpredictable situation. As predicted, the more urgent warning was rated more annoying and startling. The drivers who received an urgent warning braked significantly harder to the unpredictable event (a bus pulling out in front of the truck). The drivers also tended to brake later after the urgent warning, but no significant effect on response time or time to collision was found. A concluding recommendation for future research is to investigate distracting effects of urgent auditory warnings on less experienced drivers.

1. INTRODUCTION

A number of authors have reported that auditory cues could facilitate drivers in dangerous situations [1-4]. Sound can be perceived at any time, and regardless where the driver has visual focus. Thus, auditory cues may be especially appropriate in urgent situations that require attention.

As the number of In-Vehicle Information Systems (IVIS) increase – so does the number of auditory alerts and warnings. Therefore, it is becoming increasingly important to investigate the potentially negative effects that warning signals can have on drivers. The research presented in this article focuses on how urgent auditory warnings can negatively impact experienced drivers affective state, and ability to detect and respond to new information in the traffic scene.

Appropriate "urgency mapping" between warnings and events could guide drivers attention and help them prioritize better. A body of research has shown that manipulation of acoustical properties can impact the perceived urgency of a warning [5-9]. Edworthy et al. [5], for instance, identified a number of parameters such as pitch, harmonic series, speed and pitch range that had a consistent effect on urgency.

However, the perceived urgency of a sound may not solely depend on acoustical properties. Guillaume et al. [10] showed that the predictions by Edworthy et al. [5] were not completely accurate when applied to real alarms from military aircrafts. Burt et al. [11] reported that even though participants were able to rank "sonic urgency" before an experiment, they were not able to do so after the experiment when sounds had been mapped to situations. In conclusion, it is established that both spectral and temporal aspects of a warning signal can raise urgency. However, perceived urgency may also depend on other associations and learned mappings.

Acoustical parameters that affect rated urgency might speed up reaction time (RT). Haas and Edworthy [8] found that higher pitch, signal level and inter-pulse interval (time elapsed from the end of the offset of one pulse to the beginning of the onset of the next) increased perceived urgency. They also reported that increased level and pitch decreased RT in a simple reaction task. Haas and Casali [7] reported that higher signal level and shorter time between bursts raised rated urgency, and increased signal level decreased RT in a simple reaction task. Jaśkowski et al. [12] also reported that increased signal level resulted in a faster RT. Suied et al. [13] showed that shorter inter-onset interval raised perceived urgency and decreased RT in a simple reaction task. Edworthy et al. [14] found that envelope shape, harmonic structure, pulse-pulse interval, rhythm, average pitch, pitch range and pitch contour) affected RT in a simple reaction task.

Previous studies have shown that parameters that affect urgency could impact perceived annoyance**.** Tan and Lerner [15]**,** for instance, evaluated alerts for a collision warning system and reported that signals perceived as louder were rated more annoying. Wiese and Lee [16] reported that warnings designed to sound urgent tended to speed drivers' accelerator release. But they were also rated more annoying. Wiese and Lee recommended that designers should consider an annoyance trade-off in addition to urgency mapping. Marshall et al. [9] identified a number of parameters (harmonic series, pulse duration, inter-pulse interval, alert onset and offset, burst duty cycle, inter-burst period and sound type) that affected both perceived urgency and annoyance. They concluded that annoyance is an important factor to consider in system design, especially when designing alerts for less critical situations. But the various parameters affected urgency and annoyance differently. Thus, the assumption that parameters that increase urgency increase annoyance in a corresponding way may not be completely valid.

Designing sounds that are not annoying is important for several reasons. Unpleasant alarm tones have been found to be a common reason why operators disable the sound of communicating systems [17]. Also, unpleasant signals have been found to impact both drivers' mental workload and performance. Wiese and Lee [16] found a correlation between rated annoyance of auditory warnings and perceived workload (NASA-TXL) when drivers performed a simulated driving task. Baldwin [3] examined semantic and acoustical properties of verbal warning signals and reported that signals of intermediate urgency decreased crash risk during simulated driving. The high-urgency warning used in the experiment was considered to be very annoying and did not reduce crash risk.

We still know relatively little about how acoustical parameters that affect perceived urgency could impact drivers' ability to take in and process information. Inherent urgency may motivate the driver to focus on some particular area of the road scene or interface. However, urgency represents an increased level of threat, which may require an immediate physiological and psychological reaction (higher arousal). One sign of high arousal levels is increased attentional narrowing [18, 19]. A certain degree of alertness and focus is probably appropriate in an urgent situation. But severe attentional narrowing may not be appropriate in complex and eventful situations that require the driver to chare attention between several ongoing events. Thus, a better understanding of how warning signals can impact drivers attention have important implications for IVIS design.

Based on the previous studies of acoustical properties and annoyance there are reasons to believe that urgent signals can impact drivers affective state. A number of studies have found that characteristics in sound that raise annoyance and urgency also increases perceived arousal. Tajadura et al. [20] investigated alerts from an emotional perspective and found that higher pitch increased perceived arousal. Västfjäll et al. [21] reported that perceived annoyance of aircraft noise correlated with perceived arousal.

The potential effect of arousal on drivers' selective attention was demonstrated by Chapman and Underwood [22]. An experiment was conducted to investigate drivers' visual behavior when watching traffic situations with different levels of danger. More dangerous (arousing) situations "were characterised by a narrowing of visual search, shown by an increase in fixation durations, a decrease in saccade angular distances, and a reduction in the variance of fixation locations".

The present experiment was designed to investigate how an urgent warning can impact the affective state of experienced drivers, but also their ability to detect and respond to less predictable events in the traffic scene. Based on previous research it was predicted that a more urgent warning would be considered more annoying and startling. It was also predicted that a more urgent warning would result in a delayed response to an unpredictable traffic event.

2. METHOD

24 professional truck drivers between the ages of 23 and 70 years (M=43.3, SD=13.1) participated in the experiment. Their truck driving experience ranged from 1 to 46 years (M=21.0, SD=12.9) and their annual driving ranged between 15000 and 150000 km (M=90218, SD=3838). All drivers had normal or corrected-to-normal vision and self-reported normal hearing. They all gave their informed written consent to participate in the study.

2.1. Apparatus

The experiment was conducted in the VTI Driving Simulator III at the Swedish National Road and Transport Research Institute [23]. This high-end simulator has an advanced motion system that enables lateral or longitudinal acceleration forces up to 0.8g. A vibration table is implemented under the vehicle cab to simulate different road conditions. The traffic scene is presented on three main screens covering 120 degrees of the driver's visual field. These projections are accompanied by thee rear mirrors covering the rear view. Taken together, the VTI Driving Simulator III is capable of producing a realistic driving experience in a highly controlled setting.

2.2. Stimuli

Two auditory warning signals were created prior to the experiment. Both signals were designed to warn drivers about vulnerable road users (pedestrians) standing close to the roadside. They started with a 1000 ms verbal message, "pedestrians", presented in Swedish by a female voice. The message was followed by one of two sets of tone bursts that lasted for 1500 ms. Both spectral and temporal parameters of the burst sets were manipulated to make them different in terms of perceived urgency. Pitch and harmonic series have been suggested to affect perceived urgency [5, 6, 8, 9 11]. The low-urgency warning had a fundamental frequency of 179 Hz (G3). The high-urgency warning consisted of a cluster of tones (B4, C5, D5, C6, B6), which formed a disharmonic sound with higher frequency components. The speed of a signal has also been suggested to affect urgency [5, 7-9]. The low-urgency warning contained 2 bursts with a 300 ms inter-pulse interval. The high-urgency sound had 8 bursts with 10 ms inter-pulse intervals. Shorter amplitude onset and offset have been found to increase perceive urgency [5, 9]. Amplitude onset and offset times for the low-urgency warning was 300 ms and 450 ms. Onset and offset times for the high-urgency warning was 25 ms and 210 ms. Haas and Casali [7] and Haas and Edworthy [8] reported that higher loudness increased rated urgency. Warnings were calibrated to approximately 80 dB(A) and 85 dB(A), which prevented them from being masked by other sounds in the environment. The background noise was calibrated to be approximately 64 dB(A) at the drivers' position at 50/km speed. Both warnings were presented in the spatial position of the pedestrians in a 6.0 channel speaker setup (Anthony Gallo Acoustics Inc, CA, USA).

2.3. Evaluation of auditory signals

A study was conducted to test whether the two signals would be perceived differently in terms of perceived urgency and affective reaction. 18 volunteer subjects (16 males and 2 females) participated. Their ages ranged from 20 to 56 years (M=32.4, SD=8.4). The sounds were presented in counterbalanced order in a pair of KOSS UR5 headphones (Koss Corporation, WI, USA). The participants listened to background noise recorded inside a mini van for one minute. After 20 seconds the first warning was triggered. The participants were then required to rate perceived urgency, startling effect and annoyance using rating scales ranging from 1 (not at all) to 7 (very much). The participants also rated their affective reactions using the Self-Assessment Manikin (SAM) [24]. After another 30 seconds the second warning was triggered. Results of the ratings are presented in Table 1. Two-tailed paired t-tests were used to test for significance between distributions. The urgent warning produced significantly higher ratings in all parameters at the 0.01 alpha level.

	High urgency	**Low urgency**	**p-value**
Urgency, 1-7	5.9 (1.2)	4.0 (1.6)	<0.01
Startling, 1-7	4.6 (2.0)	2.6 (1.7)	<0.01
Annoyance, 1-7	5.5 (1.2)	3.0 (1.8)	<0.01
Arousal, 1-5	3.8 (1.1)	2.5 (0.9)	<0.01
Valence, 1-5	3.8 (0.9)	2.6 (0.6)	<0.01

Table 1: Mean ratings for the two auditory warnings. Standard deviations are presented in parentheses.

2.4. Traffic situations

Two critical situations were designed for the experiment. In one situation (bus), illustrated in Figure 1, the driver received a warning about pedestrians standing near the roadside. A bus was parked ahead of the crowd. Just as the truck passed the pedestrians the bus started to pull out and the driver was required to brake immediately to avoid a collision.

In the other traffic situation (car), the truck was heading an intersection with a small crowd of people standing near a cross walk. The driver received a warning about the pedestrians. Just as the truck entered the intersection, a passenger vehicle approached at high speed from the right and the driver were required to brake to avoid a collision.

Pilot trials were conducted with four drivers to identify any issues regarding the structure and timing of the critical events. A problem found was that the drivers tended to stop for the pedestrians. It was therefore decided to move the pedestrians further away from the road. Another issue was regarding the timing of the critical event in the car situation. It was problematic to get the car in a position so that drivers would spot it and take action to avoid a collision. The timing was adjusted in the pilot trials and it was decided to use the situation in the experiment.

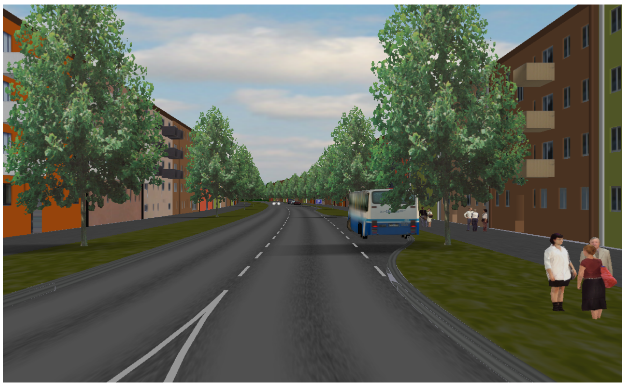


Figure 1: Traffic situation with pedestrians and a parked bus. Drivers received a warning about the pedestrians standing to the right. Moments later the bus pulled out in front of the truck.

2.5. Procedure

The experiment was conducted using a within-subjects design. Critical situations with warning signals were presented in counterbalanced order. At arrival, the drivers were introduced to the VTI Driving Simulator III and the driving task. They were informed that the vehicle was equipped with a system capable to warn them about potential road dangers. Each participant drove one practice scenario that lasted for about 8 minutes, and then the main driving scenario that lasted for 25-30 minutes. In total, each driver passed 18 intersections and 8 buses during the main driving scenario. Each critical event occurred three times – one time directly after a high-urgency warning, once after a low-urgency warning, and once without a warning. Both types of warning also occurred one time without a following critical situation. The drivers were told not to exceed the speed limit of 50 km/h. Brake response time, time to collision (TTC), brake force and subjective ratings of annoyance and startling effect defined the main dependent variables.

Directly after the trial, participants completed a questionnaire containing statements about the critical situations, the driving task and the warning signals. The drivers were required to rate perceived annoyance and startling effect using rating scales ranging from 1 (not at all) to 7 (very much). A loosely structured interview was also conducted to collect complementary driver input. At this point the experimenter revealed the purpose of the experiment and the drivers were allowed to talk freely about any issues experienced during the trials. The experimenter especially paid attention to comments about the auditory signals and how drivers focused their attention in the dangerous situations.

3. RESULTS

Results are based on data from 24 participants. Complete brake response data was collected in the bus situation. Mean time between the drivers received a warning and the bus pulling out was 2807 ms (SD=957). Unfortunately, there was severe loss of data in the car situation. The reason was an issue with timing, which prevented many participants to brake for the car. Thus, all data from that situation was excluded from analysis.

3.1. Affective reactions

Table 2 shows mean values for the ratings of perceived annoyance and starling effect. As predicted, the drivers rated the urgent signal as significantly more annoying and startling. 2 drivers rated the low-urgency warning as being more annoying, and only 1 driver rated the non-urgent warning as being more startling than the high-urgency warning. A two-tailed paired t-test revealed significant differences between the sounds both in terms of rated annoyance (t(23)=2.94, p=0.007) and startling effect (t(23)=3.14, p=0.005).

Annoying (1-7)	**Mean**	**Median**	**SD**
High urgency	4.42	5	1.84
Low urgency	3.42	3	1.56
Startling (1-7)			
High urgency	3.71	4	1.88
Low urgency	2.71	3	1.60

Table 2: Subjective ratings of annoyance and startling effect.

3.2. Brake response

Table 3 shows mean response time, time to collision and brake force. All drivers successfully avoided a collision. Brake force was measured in terms of maximum brake pressure level. Two-tailed paired t-tests failed to show any significant effects between treatments in any of the dependent variables at the 5% alpha level. A moderate correlation was found for the variables brake force and TTC (Spearman's rank order correlation, r=-0.54, p<0.01), and brake force and brake response time (r=0.59, p<0.01).

Response time (ms)	**Mean**	**Median**	**SD**
High urgency	1441	1410	381
Low urgency	1352	1290	284
Time to collision (ms)			
High urgency	2000	1900	490
Low urgency	2088	2100	411
Brake force (bar)			
High urgency	4.5	4.05	2.33
Low urgency	3.76	3.75	1.79

Table 3: Descriptive statistics for the driving parameters.

3.3. Analysis of first situations

Several drivers stated that they radically changed their expectations about threatening situations after the first critical situation. Also, the drivers responded considerably faster in the second situation (M=1233, SD=269) compared to the first situation (M=1559, SD=320). A two-tailed paired t-test showed that the difference was significant (t(23)=3.63, p=0.0013). It was therefore decided to examine the results from the first situations in more detail. In this analysis 12 drivers who received an urgent warning were compared with 12 drivers who received a low-urgency warning. Mean time between drivers receiving a warning and the bus pulling out was 2564 ms (SD=769) in the first situation. Mean brake response time, time to collision and brake force are presented in Table 4.

Response time (ms)	**Mean**	**Median**	**SD**
High urgency	1637	1610	370
Low urgency	1482	1520	251
Time to collision (ms)			
High urgency	1900	1850	381
Low urgency	2133	2050	369
Brake force (bar)			
High urgency	6.06	7.15	2.05
Low urgency	4.23	4.15	1.74

Table 4: Driving parameters in the first situation.

3.3.1. Brake response

Mean brake response time was longer after the high-urgency signal compared to the low-urgency signal. The mean difference between groups was 155 ms. However, an independent samples t-test returned no significant difference between the distributions at the 0.05 alpha level. Mean time to collision was also shorter, but the difference was not significant.

Most drivers who received an urgent warning braked harder than drivers who received a non-urgent warning. 58 % of the drivers who received an urgent warning reached brake pressure levels close to highest possible brake pressure. Normal distribution of data was not assumed. Both a two-tailed Mann-Whitney U-test, and a two-tailed independent samples t-test returned a significant difference in maximum brake pressure between the distributions (U=113, n_1=n_2=12, p<0.05) (t(22)=2.43, p=0.024).

4. DISCUSSION

The purpose of this experiment was to investigate how urgent auditory warning signals may impact experienced drivers affective state and ability to respond to other, more unpredictable events in the road scene.

One could argue that the most important property of a warning is that it will be detected by the driver and contribute to a fast response. Previous studies have shown that more urgent signal could speed response time in a simple reaction tasks [7, 8, 13, 14]. Wiese and Lee [16] investigated the effects of an urgent warning during simulated driving and reported that increased burst density of a collision warning speeded accelerator release.

However, annoying auditory signals could undermine acceptance, and have been suggested to be a common reason why operators turn of system alerts [17]. Wiese and Lee [16] suggested an annoyance trade-off when designing warning signals for in-vehicle use. The results obtained in the present experiment indicate that warning signals presented in a truck cabin could impact affective state differently. As predicted, the high-urgency warning was rated more annoying and startling compared to low-urgency warning. These results were not at all surprising and they are in line with previous findings suggesting that acoustic properties can affect rated urgency, annoyance and arousal [9, 15, 16, 20]. But most previous studies have been conducted with ordinary car drivers or not in a driving context. The present study was conducted in a high-end truck simulator with highly trained truck drivers. On the basis of the result from this and previous studies we suggests that truck manufacturers should not only consider alarm efficiency, but also annoyance potential when designing and implementing auditory warnings.

Mean scores of annoyance were almost identical to the results in the pre-study for the low-urgency warning. But the high-urgency warning was rated considerable less annoying by the professional truck drivers. A two-tailed t-test reviles that the difference is significant (p<0.05). There are several possible explanations to this effect. One could be that professional drivers are used to handle critical driving situations, and simply felt less affected by the urgent sound. Other contributing factors could be that the subjects in the pre-study only listened to the sounds one time, and that the signals were not mapped to any situations. The professional truck drivers listened to them 3 times, and the sounds were mapped to specific traffic situations. Previous findings suggest that perceived urgency of warnings can change considerably when they have been mapped to situations, even though the listener is told to ignore any associations and just focus on the sound [11]. Also, in the interview one driver stated that it was hard to remember the sounds being different. The ratings were performed after completing the 25-30 minutes driving task and the participants may not have been able to provide precise ratings of their affective state at this time. In future studies it may be more appropriate to let drivers rate their affective state directly after the critical situations.

Analysis of response performance in first situations showed that the drivers who received a high-urgency warning braked significantly harder than drivers who received the low-urgency warning. Previous experiments have suggested that increased arousal [12] and the ”stimulus-response compatibility” [25] could lead to more forceful reactions. The moderate correlation found between response time and brake force suggests that drivers compensated for late responses by braking harder.

Drivers who received the high-urgency warning also tended to brake later compared to drivers who received the low-urgency warning. But there were large differences between drivers, and the differences did not reach statistical significance. But even so, there are reasons to consider more studies examining distracting effects of urgent alerts on drivers. Today, car manufacturers are developing and implementing new technology to assist ordinary car drivers in dangerous and eventful situations. Experienced drivers are probably more used with critical and demanding situations than are less experienced drivers. Chapman and Underwood [22] found that novice drivers showed longer fixation durations than experienced drivers in critical traffic situations, indicating that they are less able to share attention appropriately in these situations. Future studies should examine the effects of urgent warning signals on less experienced drivers.

Previous studies have emphasized the use of early warnings instead of late warnings in a driving context. Lee et al. [4] found that early warnings helped distracted drivers more effectively than did late warnings. In a second experiment they showed that early warnings resulted in a safety benefit by reducing the time required for drivers to release the accelerator. Early and more comfortable warnings that inform the driver about important states and ongoing events could be an especially interesting alternative to alarming and annoying signals. If an extremely fast response time is important, it is probably better to consider overtaking systems such as automatic brake systems.

The length of the trials prevented any investigation of long-term effects and habituation of the signals. No considerable effect on brake response behavior was found after the first situation, indicating that response performance for experienced drivers will not be negatively affected by “sonic urgency” when critical situations are expected.

5. CONCLUSIONS

The results of the experiment suggest that acoustical parameters that increase urgency can impact experienced drivers’ affective state in demanding traffic situations. Urgent signals could potentially also impact drivers’ responses to unpredictable events in the traffic scene. These results have implications for system design, especially for systems designed to warn and inform drivers in very complex and eventful situations. Previous authors have suggested developers should consider an annoyance trade-off when implementing auditory warnings in vehicles. The results of this study imply that it may be a good idea to also consider a tradeoff between perceived urgency and contextual complexity. A recommendation for future research is to investigate distracting effects of auditory warning signals on less experienced drivers.

6. ACKNOWLEDGEMENTS

This research was financed by Scania Commercial Vehicles, Vinnova and partners within the Swedish excellence centre “Virtual Prototyping and Assessment by Simulation” (ViP). Special thanks to Håkan Alm at Luleå university of technology, Robert Friberg and Anna Svensson at Scania Commercial Vehicles, and Anders Andersson, Anne Bolling and Anders Genell at the Swedish National Road and Transport Research Institute (VTI).

7. REFERENCES

[1] C. P. Fung, S. H. Chang, J. R. Hwang, C. C. Hsu, W. J. Chou, and K.K. Chang, *The study on the influence of auditory warning systems on driving performance using a driving simulator*, Taiwan, Institute of Transportation, 2007.

[2] C. Ho, and C. Spence, “Assessing the effectiveness of various auditory cues in capturing a driver’s visual attention,” *Journal of experimental psychology: Applied*, vol. 11, no. 3, pp.157-174, 2005.

[3] C. L. Baldwin, “Acoustic and semantic warning parameters impact vehicle crash rates,” in *Proc. of ICAD 2007*, Montreal, Canada, 2007.

[4] J. D. Lee, D. V. McGehee, T. L. Brown, and M. L Reyes, “Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator,” *Human factors*, vol. 44, no. 2, pp. 314-334, 2002.

[5] J. Edworthy, S. Loxley, and I. Dennis, “Improving auditory warning design: Relationships between warning sound parameters and perceived urgency,” *Human factors*, vol. 33, no. 2, pp. 205-231, 1991.

[6] E. J. Hellier, J. Edworthy, and I. Dennis, “Improving auditory warning design: quantifying and predicting the effects of different warning parameters on perceived urgency,” *Human factors*, vol. 35, no. 4, pp. 693-706, 1993.

[7] E. C. Haas, and J. G. Casali, “The perceived urgency and detection time of multi-tone and frequency-modulated warning signals,” in *Proc. of the human factors and ergonomics society 37th annual meeting*, Santa Monica, CA, USA, 1993.

[8] E. C. Haas, and J. Edworthy, “Designing urgency into auditory warnings using pitch, speed and loudness,” *IEE Computing and Control engineering journal*, vol. 7, no. 4, pp. 193-198, 1996.

[9] D. C. Marshall, J. D. Lee, and P. A. Austria, "Alerts for in-vehicle information systems: annoyance, urgency and appropriateness," *Human factors*, vol. 49, no. 1, pp. 145-157, 2007.

[10] A. Guillaume, C. Drake, M. Rivenez, L. Pellieux, and V. Chastres, "Perception of urgency and alarm design," in *Proc. of ICAD 2002*, Kyoto, Japan, 2002.

[11] J. L. Burt, D. S. Bartolome, D. W. Burdette, and J. R. Comstoc JR, "A psychophysiological evaluation of the perceived urgency of auditory warning signals," *Ergonomics*, vol. 38, no. 11, pp. 2327-2340, 1995.

[12] P. J. Jaśkowski, K. Rybarczyk, F. Jaroszyk, and D. Lemański, "The effect of stimulus intensity on force output in simple reaction task in humans," *Acta Neurobiol. Exp*, vol. 55, no. 1, pp. 57-64, 1995.

[13] C. Suied, P. Susini, and S. McAdams, "Evaluating warning sound urgency with RTs," *Journal of experimental psychology: Applied*, vol. 14, no. 3, pp. 201-212, 2008.

[14] J. Edworthy, E. Hellier, and K. Walters, "The relationship between task performance, RT, and perceived urgency in nonverbal auditory warnings," in *Proc. of the IEA 2000/HFES 2000 Congress*, San Diego, CA, USA, 2000.

[15] A. K. Tan, and N. D. Lerner, *Multiple attribute evaluation of auditory warning signals for in-vehicle crash avoidance systems*, National Highway Traffic Safety Administration, Office of Crash Avoidance Research, Washington DC, USA, 1995.

[16] E. E Wiese, and J. D Lee, "Auditory alerts for in-vehicle information systems: The effects of temporal conflict and sound parameters on driver attitudes and performance," *Ergonomics*, vol. 47, no. 9, pp. 965-986, 2004.

[17] F. E. Block JR, L. Nuutinen, and B. Ballast, "Optimization of alarms: a study on alarm limits, alarm sounds, and false alarms, intended to reduce annoyance," *Journal of clinical monitoring and computing*, vol. 15, pp. 75-83, 1999.

[18] C. D. Wickens, and J. D. Hollands, *Engineering psychology and human performance (3 rd ed.)*, Upper Saddle River, USA, Prentice-Hall Inc., 1999.

[19] J. A. Easterbrook, "The effect of emotion on cue utilization and the organization of bahavior," *Psychology Review*, vol. 66, no. 3, pp. 183-201, 1959.

[20] A. Tajadura-Jiménez, A. Väljamäe, N. Kitagawa, and D. Västfjäll, "Affective multimodal displays: Acoustic spectra modulates perception of auditory-tactile signals," in *Proc. of ICAD 2008*, Paris, 2008.

[21] D. Västfjäll, M. Kleiner, and T. Gärling, "Affective reactions to interior aircraft sounds," *Acta Acustica united with Acustica*, vol. 89, no. 4, pp. 693-701, 2003.

[22] P. R. Chapman, and G. Underwood, "Visual search of driving situations: Danger and experience," *Perception*, vol. 27, pp. 951-964, 1998.

[23] http://www.vti.se/simulator

[24] M. M. Bradley, and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49-59, 1994.

[25] R. Ulrich, and S. Mattes, "Does immediate arousal enhance response force in simple RT?," *Q J Exp Psychology*, vol. 49A, no. 4, pp. 972-990, 1996.

COMPREHENSION OF SPEECH PRESENTED AT SYNTHETICALLY ACCELERATED RATES: EVALUATING TRAINING AND PRACTICE EFFECTS

Christina Wasylyshyn, Brian McClimens, and Derek Brock

U.S. Naval Research Laboratory
Washington, DC 20375
christina.wasylyshyn@nrl.navy.mil

ABSTRACT

The ability to monitor multiple sources of concurrent auditory information is an integral component of Navy watchstanding operations. However, this leads to attentionally demanding environments. The present study tested the utility of a potential solution to listening to multiple speech communications in an auditory display environment: presenting speech serially at synthetically accelerated rates. Comprehension performance of short auditory narratives was compared at seven accelerated speech rates. Practice effects and training effects were examined. An optimum acceleration rate for comprehension performance was determined, and training was found to be an effective method when synthetic speech was presented at slow to moderately accelerated rates.

1. INTRODUCTION

It is expected that future Naval forces will be defined by their agility and their capacity for coping with high-stakes, uncertain environments [1]. The individual Naval watchstander might be responsible for the concurrent monitoring of numerous radio communications channels, along with actively monitoring and responding to events on multiple visual displays. Such attentionally demanding environments have motivated various HCI solutions to help warfighters deal with the vast number of information sources needing to be monitored in order to perform their duties successfully.

However, a critical consideration when designing potential solutions in auditory display research is to take into account the limitations in listeners' abilities to attend to multiple competing communications channels. For example, communications performance has been shown to decline significantly as watchstanders were asked to handle additional radio circuits [2]. Likewise, comprehension in multi-talker speech displays decreased when listening to concurrent speakers [3]. Both of these studies [2], [3] examined performance in conditions with up to four concurrent speakers. These results were replicated and extended by comparing comprehension performance in four different conditions: two concurrent speakers, four concurrent speakers, four serial speakers with normal speech rates, and four serial speakers with accelerated speech rates (accelerated 75% faster) [4]. Major findings included better comprehension performance in the accelerated speech condition compared to the concurrent speech conditions (for both two and four concurrent speakers). Participants performed better in the normal speech rate condition than the accelerated speech rate condition, suggesting that while listening to accelerated speech is more difficult than normal speech, there is a significant improvement in one's ability to make sentence comprehension judgments when listening to accelerated speech compared to listening to concurrent speech.

Presenting listeners with messages that are serialized and accelerated, therefore, may be one potential solution to concurrent monitoring of communications channels. However, building an effective system of synthetically accelerated voice communications will require new information paradigms that directly address the strengths and limitations of human operators. This paper reports on a work in progress that examines listeners' abilities to adapt to synthetically accelerated speech in an auditory display environment through the use of practice and training.

Unlike reading, in which the information input rate can be controlled by one's eye movements, comprehension of speech is often dependent on a transient acoustic signal whose information input rate is largely controlled by the talker, not the listener. The information input rate, thus, is determined by the environment, and previous information is often not reviewable. In order to comprehend auditory information effectively, input must be analyzed, segmented, and processed for structure and meaning, all of which must occur even as new auditory information continues to arrive. When auditory input is rapid, listeners will have even less time to carry out these integrative processes, and successful comprehension will require greater effort at accelerated rates of speech.

Accelerated speech is marked by an increased word rate, so that more information can be transmitted per unit of time. Even when the pitch and prosody of the original speech is preserved, loss of information occurs, most notably from the loss of processing time that the listener would typically use to integrate the auditory information [5]. However, with practice, subjects have shown increased recall of information that was presented at an accelerated rate [6], and mere exposure to accelerated speech has been shown to generalize to increased speech comprehension of other accelerated speech, even if the subject is exposed to accelerated speech in a foreign language with similar phonemes [7].

While these studies suggest that there are detectable performance differences in accelerated speech comprehension, it isn't clear to what extent training participants to listen to synthetically accelerated speech will be helpful. The amount of practice and/or exposure to accelerated speech needed to produce benefits in comprehensibility differs across studies, as does the ability to distinguish practice effects from training effects.

The present study tested the utility of listening to synthetically accelerated speech by comparing comprehension performance of information presented at seven accelerated speech rates. Speech rates were blocked so that three short narratives were presented at each rate. This approach allowed for the testing of practice effects across the three narratives for each speech rate. Furthermore, the blocked narratives were either presented at incrementally faster rates (the "training" group) or in a random manner (the "random" group). This allowed us to determine whether presenting accelerated speech in a systematic way from slow to fast speech rates was beneficial to comprehension performance or whether participants were only able to integrate accelerated auditory information up to a certain speed-related processing threshold. The study sought to answer three primary questions:

1. Do practice effects occur, i.e., is there systematic improvement in comprehension performance after listening to multiple auditory excerpts presented at the same accelerated speech rate?

2.. Does comprehension performance vary by presentation method, i.e., can listening to accelerated speech be trained or do participants simply have a natural threshold for listening to and processing accelerated speech content?

3. What is the optimum acceleration rate for comprehension performance, i.e., what is the fastest rate at which speech can be presented so that comprehension performance does not differ from comprehension performance of speech presented at a normal rate?

2. METHOD

2.1 Participants

Twenty NRL employees participated (11 males, 9 females). All participants were native English speakers and claimed to have normal (i.e., non-corrected) hearing. Participants were randomly assigned to the "training" (5 males, 5 females, mean age = 40.9, SD = 11.1) or to the "random" (6 males, 4 females, mean age = 39.6, SD = 10.8) presentation group. There are no significant differences in participant characteristics between groups.

All participants were presented with a baseline listening comprehension task, that is, all participants listened to two narratives at a normal speech rate and completed comprehension questions. The training group (mean = 0.76, SD = 0.11) performed equally well as the random group (mean = 0.78, SD = 0.10), t(18) = -0.34, p = .54, so that there were no differences between the groups for baseline comprehension performance.

2.2 Task and Apparatus

The main battery was composed of brief auditory narratives and comprehension questions [8]. Each narrative described an event in a person's life. These narratives were approximately 300 words in length (range = 298 – 308 words); they were equated for number of ideational propositions and content difficulty. Each narrative was recorded in a female voice at a speaking rate of approximately 180 words/min (normal speaking rates are anywhere between 130-200 words/min).

After listening to each narrative, participants were asked to evaluate statements about ideas represented (or not) in the narrative. These consisted of 24 statements that included both main ideas and specific details about the narrative. Three different types of statements were included for comprehension evaluation:

1. *True* statements represented ideas that were included in the narrative
2. *False* statements represented ideas that were inconsistent with those told in the narrative
3. *Distractor* statements represented ideas that were consistent with the narrative, but were not actually part of it.

The narratives were synthetically accelerated at rates ranging in 15% increments from 50% to 140% faster-than-normal. The "training" presentation group listened to the accelerated narratives at incrementally faster rates from 50% to 140% faster-than-normal. The "random" presentation group listened to the narratives at accelerated speeds presented in a random fashion. For both presentation groups, the narratives were presented in triads at each speed to test for practice effects within speeds. For example, a participant in the "training" group would have heard three narratives at 50% faster-than-normal, followed by three narratives at 65% faster-than-normal, followed by three narratives at 80% faster-than-normal, and so on, up to 140% faster-than-normal.

In order to create the accelerated test battery, the narratives that were first recorded at a normal speaking rate were subjected to a patented NRL speech-rate compression algorithm [9], known as "pitch synchronous segmentation" (PSS). PSS retains the fundamental frequency of speech signals and preserves a high degree of intelligibility. This high degree of intelligibility remains because the PSS method does not try to generate an electric analog of the human speech production mechanism. Instead, PSS represents the speech waveform by individual pitch cycle waveforms. The output speech sounds more natural because it is constructed from raw speech and because pitch interference is absent in the speech representation.

The visual part of the study was displayed on a large flat-panel monitor and the auditory component was rendered binanrally in Sony MDR-600 headphones. Brief auditory examples of what participants heard at each accelerated speech rate are given in the following sound files:

Speed50%	[SPEED50.WAV]
Speed65%	[SPEED65.WAV]
Speed80%	[SPEED80.WAV]
Speed95%	[SPEED95.WAV]
Speed110%	[SPEED110.WAV]
Speed125%	[SPEED125.WAV]
Speed140%	[SPEED140.WAV]

2.2 Procedure

Participants were randomly assigned to either the “training” or the “random’ presentation group. After providing informed consent, participants completed a short practice exercise that resembled the format of the experimental task and a baseline comprehension measure. Immediately, after listening to each narrative, participants were visually presented with 24 statements (8 true, 8 false, and 8 distractor) and asked to evaluate whether or not the statement identified ideas heard in the narrative.

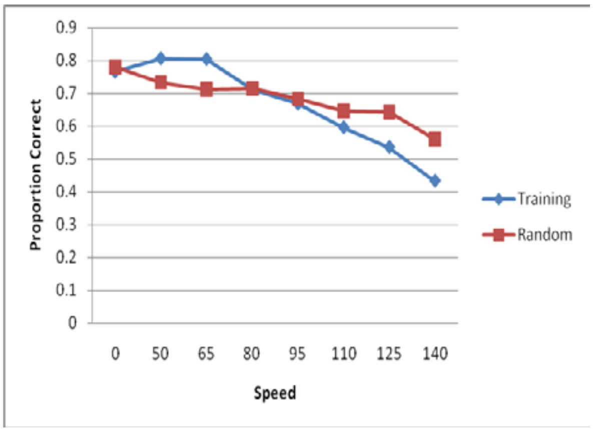


Figure 1: Mean proportion of correctly identified comprehension statements by speed for each presentation method group (training and random).

3. RESULTS

The proportion of correctly identified comprehension statements served as the dependent measure. The dependent measure was submitted to an 8(speed: normal, 50%, 65%, 80%, 95%, 110%, 125%, 140%) x 3 (practice: narrative 1, 2, 3 within each triad) x 2 (presentation method: training, random) mixed ANOVA. Speed and practice were repeated-measures variables, and presentation method was a between-groups variable. The main effect of speed was significant, F(7, 126) = 19.88, p < .0001. Regardless of presentation method and practice, participants correctly identified more comprehension statements when narratives were presented at the slower speeds (i.e., normal, 50%, and 65%). There was no main effect of practice, F(2, 36) = 0.63, p = 0.54. Across presentation method and speed, the mean comprehension scores for narratives presented first in each triad was 0.68, second in each triad was 0.67, and third in each triad was 0.67. There was also no main effect of presentation method, F(1, 18) = 0.13, p = 0.72. Across speed and practice, the average comprehension score in the training presentation group was 0.67 and the average comprehension score in the random presentation group was 0.68.

However, the speed by presentation method interaction was significant, F(7, 126) = 3.01, p = 0.006. Figure 1 displays the mean proportion of correctly identified comprehension statements by speed (across practice) for each presentation method. Note that the values marked Speed 0 indicate each group’s average comprehension score at baseline (i.e., normal speed speech). As can be seen in Figure 1, both the training and the random presentation groups performed equally well at Speed 0. Planned contrasts indicated that the optimum acceleration rate for comprehension performance was 65% faster-than-normal, that is, 65% faster-than-normal was the synthetic speech rate at which comprehension performance did not differ from comprehension performance of speech presented at a normal rate. Planned contrasts also indicated that participants in the training group correctly identified more of the comprehension statements at Speeds 50 and 65 (mean proportion correct = 0.81 and 0.80, respectively) compared to the participants in the random group (mean proportion correct = 0.73 and 0.71 for Speeds 50 and 65, respectively). This suggests that training participants may be an effective method, but only at slower speeds. Figure 1 seems to suggest that the random presentation method is more effective at higher speeds (e.g., Speeds 125 and 140) than the training presentation method, however, there were no significant differences between the presentation groups at the higher speeds. The way in which the narratives were presented to the training group (i.e., at incrementally faster rates) may have induced fatigue over the course of the experimental session, further supporting the notion that training may only be effective at slower speeds.

In summary, participants seemed to adapt quickly to comprehending the synthetically accelerated speech. Training was effective at slower accelerated speech rates, however, systematic training to higher accelerated speech rates led to fatigue. Practice (i.e., performance across the three narratives within each speech rate) did not seem to aid comprehension performance.

4. DISCUSSION

This present study reports results from a work in progress that examines listeners’ abilities to adapt to synthetically accelerated speech in an auditory display environment through the use of practice and training. Previous research conducted at NRL [4] demonstrated that comprehension performance can benefit from accelerated and serialized audio communications channels, compared to comprehension performance when listening to concurrent speech on two and/or four channels. However, participants in this previous study did not perform as well when listening to synthetically accelerated speech rates at 75% faster-than-normal as when listening to normal speech rates [4]. The present study extends those previous results. We tested a larger scale of synthetic speech rates ranging in 15% increments from 50% to 140% faster-than-normal. We found that the optimum acceleration rate for comprehension performance was 65% faster-than-normal. This was the fastest rate at which synthetically accelerated speech could be presented where comprehension performance did not differ from comprehension performance of speech presented at a normal rate.

The main analysis compared participants who listened to the narratives at incrementally faster rates from 50% to 140% (the training group) to those participants who listened to the narratives at speeds presented in a random fashion (the random group). As expected, comprehension performance declined as speech rate increased. At faster synthetic speech rates, participants were not able to integrate the structure and meaning of the narratives as well as they were able to at slower speech rates. The training presentation method was found to be effective for comprehension performance compared to the random presentation method, but only at the slower synthetic speech rates (i.e., 50% and 65% faster-than-normal). What was not expected, however, was how quickly listeners adapted to the synthetically accelerated speech. This can be seen by the lack of practice effects within speeds; on average, participants tended to perform equally across the three narratives of each triad.

That being said, it should also be noted that the highest comprehension accuracies were between 78-81%. Participants were particularly good at distinguishing the true and false statements, but performed significantly worse when presented with the distractor statements. Again, distractor statements represented ideas that were consistent with the narrative, but were not actually part of it. Determining ways to improve listeners' abilities to distinguish between distracting and true information is especially relevant to building effective systems of synthetically accelerated voice communications that can be used in attentionally demanding environments.

The current results may have future applications for coordinating the numerous communications between various disaster relief organizations and municipal services, for managing air traffic control centers, and for organizing communications in Naval combat information centers. Once we know the limits of human operators' abilities to listen to synthetically accelerated speech, we can begin to design auditory display environments that capitalize upon strengths and minimize weaknesses. The present study addresses two critical areas of concern: the trainability of listening to synthetically accelerated speech and the optimum acceleration rate for comprehension performance. Future research seeks to enhance the auditory display environment by presenting information in a way that approximates how listeners more naturally perceive it, that is, by employing auditory cues to specify communications channels that are rendered in a virtual listening space.

5. ACKNOWLEDGMENT

This work was supported by the Office of Naval Research.

6. REFERENCES

[1] V. Clark, "Sea power 21 series – part I: Projecting decisive joint capabilities," *Naval Institute Proceedings Magazine*, 2002.

[2] D. Wallace, C. Schlicting, and U. Goff, *Report on the Communications Research Initiatives in Support of Integrated Command Environment (ICE) Systems*, Naval Surface Warfare Center, Dahlgren Division, TR-02/30, Jan. 2002.

[3] D. S. Brungart, M. A. Ericson, and B. D. Simpson, "Design considerations for improving the effectiveness of multitalker speech displays," *Proceedings of the 2002 International Conference on Auditory Display*, Kyoto, Japan, 2002.

[4] D. Brock, B. McClimens, G. Trafton, M. McCurry, and D. Perzanowski, "Evaluating listeners' attention to and comprehension of spatialized concurrent and serial talkers at normal and a synthetically faster rate of speech," *Proceedings of the 14th International Conference on Auditory Display*, Paris, France, 2008.

[5] A. Wingfield, P. A. Tun, C. K. Koh, and M. J. Rosen, "Regaining lost time: Adult aging and the effect of time restoration on recall of time-compressed speech," *Psychology and Aging*, vol. 14, no 3, pp. 380-389, 1999.

[6] G. T. M. Altmann, and D. Young, "Factors affecting adaptation to time-compressed speech," *Eurospeech 9*, Berlin, Germany.

[7] N. Sebastian-Galles, E. Dupoux, A. Costa, and J. Mehler, "Adaptation to time-compressed speech: Phonological determinants," *Perception & Psychophysics*, vol. 62, pp. 834-842, 2000.

[8] R. A. Dixon, and C. M. de Frias, "The Victoria Longitudinal Study: From characterizing cognitive aging to illustrating changes in memory compensation," *Aging Neuropsychology, and Cognition*, vol. 11, pp. 346-376, 2004.

[9] G. S. Kang, and L. J. Fransen, *Speech Analysis and Synthesis Based on Pitch-Synchronous Segmentation of the Speech Waveform*, Naval Research Laboratory, TR-9743, Nov. 1994.

EXPLORING AMBIENT SONIFICATION OF WATER TOXICITY

Mikael Fernström

Interaction Design Centre,
Department of Computer Science and
Information Systems,
University of Limerick, Ireland
mikael.fernstrom@ul.ie

Sean Taylor

Sculpture & Combined Media
School of Art & Design
Limerick Institute of Technology,
Limerick, Ireland
sean.taylor@lit.ie

ABSTRACT

We explored the possibility of using ambient auditory display in the context of sonification of water toxicity. We looked at the existing work procedures carried out in an aquatic toxicity laboratory and developed a design that could replace or complement existing periodic visual monitoring of samples. The design was further developed as an art-science installation in a public exhibition in the Science Gallery in Dublin, Ireland, where visitors experienced through hearing the life and death of small aquatic crustaceans in real-time.

1. INTRODUCTION

With the proliferation of recent water contaminations in Ireland [1], we set out to explore possibilities to contribute to the public awareness of the underlying problems as well as to engage with environmental scientists handling the day to day monitoring of water quality in Ireland. The project reported in this paper is mainly situated in the artistic domain, but with a direct connection and potential for applications in science using auditory display.

There have been several interesting contributions in the ICAD community that have inspired our work. Cohen's *Out to Lunch* system may be one of the first to create and ambient auditory display [2][3] and Gaver, Smith and O'Shea's ARKola simulation also had ambient aspects [4].

Bly's work on multivariate mappings and auditory display showed clear possibilities for scientific use [5]. Paine's *Reeds* installation had both artistic and scientific dimensions and was situated in a public space [6]. Sturm's sonification of ocean buoys can be listened to both from a musical as well as scientific perspective [7][8]. In previous work we, the authors, have also explored auditory display at the boundary between art and science [9][10][11].

2. BACKGROUND

After an open call for contributions from the Science Gallery in Dublin in Ireland under the exhibition title *Infectious*, we submitted a proposal for collaboration with Enterprise Ireland's Aquatic Toxicity Laboratory (ATL) in Shannon, Co. Clare. Over a couple of months, we visited the laboratory in Shannon and learned about some of their methods for measuring and monitoring the toxicity of water samples. One of the main methods used is to use living organisms, *Daphnia magna* (hereafter called *Daphnia*). They are small, planktonic crustaceans, between 0.2 and 5 mm in size. *Daphnia* are members of the order *Cladocera*, and are one of several small aquatic crustaceans commonly called water fleas because of their saltatory swimming style. They live in various aquatic environments ranging from freshwater lakes to ponds, streams and rivers. These tiny crustaceans are very sensitive to their environment and are also used in laboratory research for analysis of water and soil toxicity. Because *Daphnia* may be used to test the effects of toxins on an ecosystem, this makes *Daphnia* an indicator species, particularly useful because of its short lifespan (typically 1 to 3 months) and parthenogenetically reproductive capabilities (they become mature in about 2 weeks and can then produce offspring every ten days).

To test a sample for toxicity, varying amounts of the sample is mixed with amounts of pure water with a small *Daphnia* population. The population is then observed for some time to check the mortality rate of the *Daphnia*. The toxicity is defined as Lethal Concentration for 50% mortality (LC50).

In addition to the testing, the ATL has to breed and maintain a healthy population of *Daphnia* under pure conditions. The breeding and feeding has to be monitored to assure that there always is sufficient supply of *Daphnia* available.

Currently, the laboratory staff in Shannon monitor samples by periodically visually inspecting beakers and counting the number of live *Daphnia,* see Figure 1.

3. CONCEPT

In collaboration with the laboratory staff at ATL, we proposed to use a simple web camera to monitor movement of *Daphnia* magna in a beaker and to sonify the movement of the living *Daphnia* as an ambient auditory display. The term ambient here refers to the ideas of Mark Weiser and John Seely Brown, as outlined in their paper on Calm Tehnology [12]. The advantage with such a display would be that the staff wouldn't have to do so many visual inspections and counts, and instead hearing in the periphery of their awareness when the *Daphnia* population deteriorated to a level when it would require more frequent and detailed monitoring. The same technique could also be used for monitoring the breeding of *Daphnia*.

We suggested using a simple mapping between the *Daphnia* movement and audio. The field of view of the web camera would be mapped to musical notes with the vertical mapped to pitch and the horizontal to note duration. This was an intuitive choice, largely based on the saltatory movement of *Daphnia* looking like an aquatic ballet. We could then use different timbres to help the staff segregate between different samples in progress at the same time.

This proposal for an exhibit at the Science Gallery in Dublin was approved and we were commissioned to develop this concept into an installation in a public gallery.

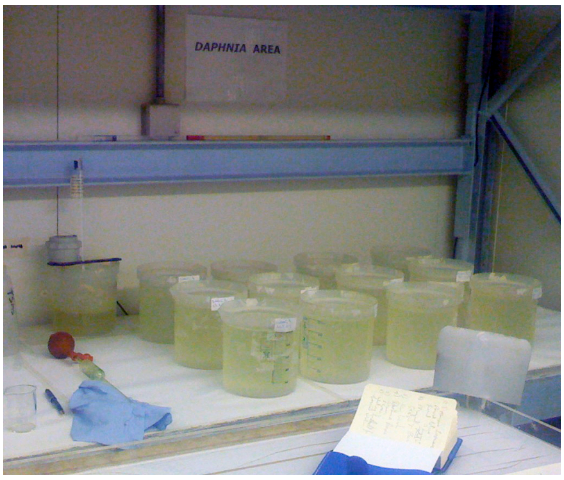


Figure 1: Water toxicity testing at the Aquatic Toxicity Laboratory in Shannon, Co. Clare.

4. DESIGN

For the public exhibition we decided to use four containers with living *Daphnia*. A web camera connected to an Apple iMac computer running *Pure Data* with the *Graphics Environment for Multimedia* (PD/GEM*)* monitored each container, see Figure 2 and Figure 3. We designed and implemented a PD-patch for capturing the video and using blob detection to track the movement of *Daphnia*. The movement was then mapped pitch pitch along the vertical axis and note duration along the horizontal axis. For timbre, we chose a synthetic human singing voice and the four containers mapped to the ranges of bass, tenor, alto and soprano. As each container typically had ten living *Daphnia*, this resulted in a complex choral polyphony. Our metaphor behind this choice was 'the budgie in the coal mine', i.e. alluding to that when the *Daphnia* die the singing stops and when humans notice this, our own end may be nigh unless we take immediate action.

As we could not use real and potentially toxic samples in the exhibition, we reverted to using a substance used by the ATL for calibration purposes, a solution of Lipopolysaccharides (LPS). When LPS was added to a container, a proportion of the *Daphnia* population died. This was directly reflected in the ambient auditory display as the number of notes per minute and range of pitches used were reduced and eventually stopped.

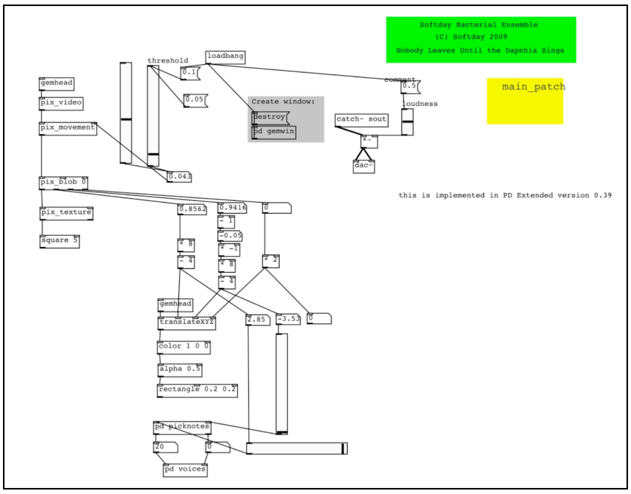


Figure 2: Main PD/GEM patch

4.1. Installation

For the physical installation we were inspired by the look and feel of the ATL in Shannon. We borrowed lab tables, glassware and various props and configured the installation into four stations, each with its own Petri-dish with *Daphnia*, video camera and computer with our specially designed PD/GEM-patches. We gave our exhibit the working title "Nobody leaves 'til the Daphnia Sing", and installed the equipment on the first floor of the Science Gallery in Dublin. See Figure 4.

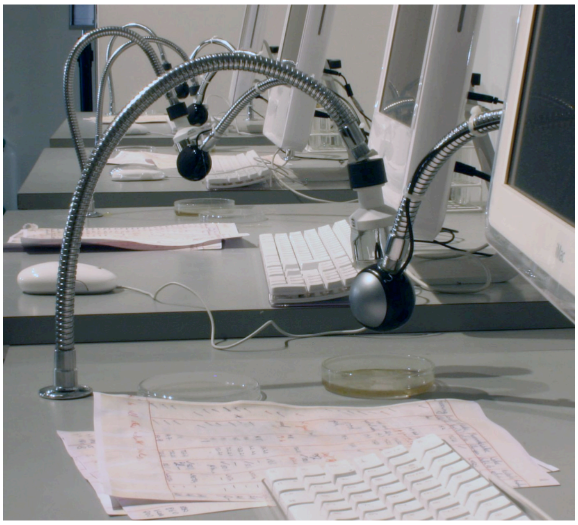


Figure 3: Web camera focused on a Petri-dish with sample of *Daphnia*.

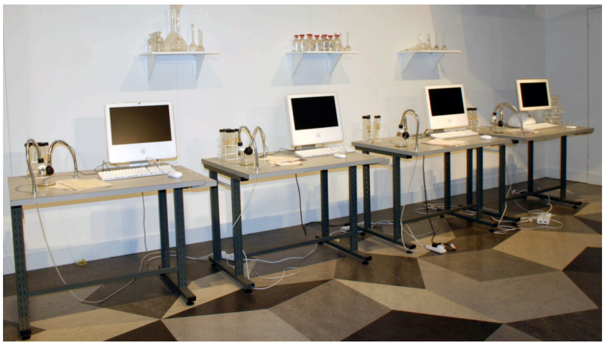


Figure 4: Installation with four stations: bass, tenor, alto and soprano.

4.2. Score

For the opening event in the gallery, we also created a live human performance element. We received a spreadsheet with data from actual water samples around Ireland that covered a period of 18 years. We normalized the data and converted it into MIDI and then further processed into a musical score, see Figure 5. This score was given to our *Softday Bacterial Ensemble*, four young musicians from the B.Sc. Programme in Music, Media and Performance Technology in the Department of Computer Science and Information Systems at the University of Limerick. The score provided a framework for improvisation, as the live performance was to be in conjunction with the sounds coming from the live movements of *Daphnia*. The tonality between the score and our computer vision to sound algorithms was aligned to allow for an interesting and unique musical experience.

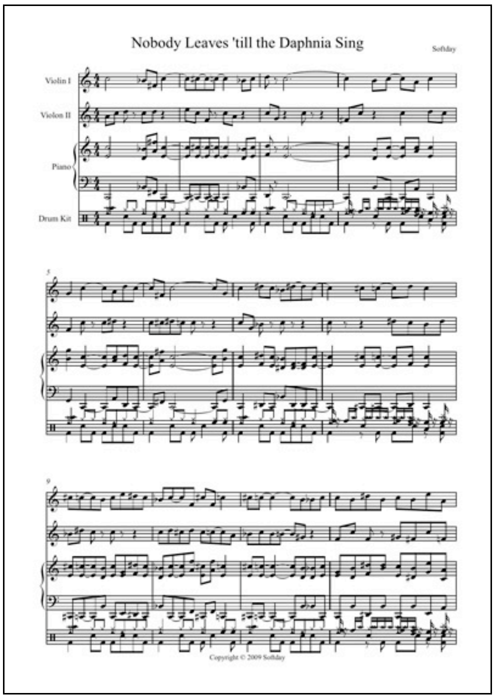


Figure 5: The first page of the musical score.

5. EXHIBITION

On Saturday the 18th of April 2009, we premiered "Nobody leaves 'til the Daphnia sing", as a live performance of a unique multimedia sound art work, as part of the Infectious exhibition at the Science Gallery, Trinity College, Dublin.

The computer generated music composition that the *Softday Bacterial Ensemble* performed was constructed utilizing a variety of *Daphnia* data sources. This composition formed the basis for an improvisation between the human musicians and the ambient auditory display of *Daphnia* populations.

After the opening event, the installation with the ambient auditory display was open to the public until the 17th of July 2009. Over the three months, approximately 45,000 visitors experienced the exhibition.

We didn't carry out any formal scientific evaluation of our exhibit, but we received ample media coverage and communications from visitors via email and phone calls. The intense general interest may also have been due to that the outbreak of A/H1N1 coincided with the exhibition – titled Infectious.

6. DISCUSSION

Based on what exhibition visitors and staff informally told us, and from discussions with staff at ATL in Shannon, it is clear that this kind of ambient auditory display can be used for a peripheral awareness about the health of multiple small populations of *Daphnia*.

Originally we had planned to use a separate sound reinforcement system in the Science Gallery, but during our initial testing on-site we found that the built-in loudspeakers in the Apple iMac computers were fully sufficient for this particular exhibition environment.

While the use of synthesized human voices helped to emphasize the relation between the health of *Daphnia* and humans, if this kind of system would be further developed for used in laboratory contexts similar to ATL in Shannon, other timbres are likely to have to be explored for improving segregation between auditory streams. It is also likely that systems like this cannot be applied in all laboratories, as other kinds of equipment may be using auditory display needing more urgent attention. Still, it is interesting to note that segregation between synthetic voices was quite good, probably due to the physical constraints of a circular container being viewed by a web camera with a rectangular field of vision. This resulted in the probability being higher for mid-register and medium duration notes being generated more often then very low or high-pitched notes. The mappings in the extreme corners of the camera's view could never be triggered, see Figure 6.

It is unfortunately not possible to try this system in the ATL in reality, as their certification and routines are based on existing international conventions regarding the procedures for measuring LC50, i.e. using periodic visual monitoring. To introduce new procedures would require approval of a separate and purely technical and scientific project.

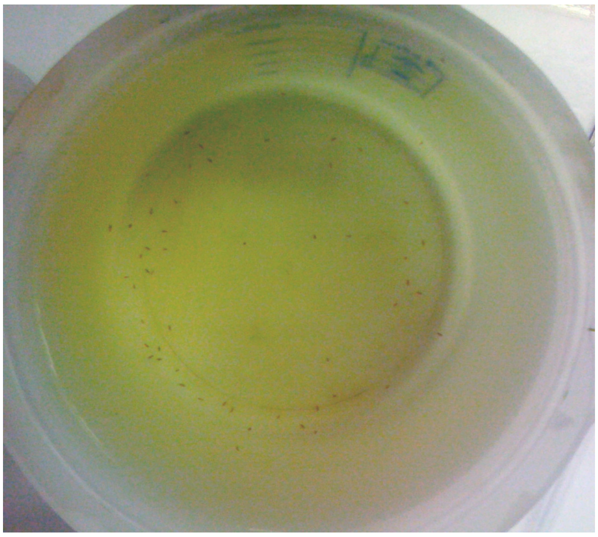


Figure 6: Web camera view of daphnia in container.

7. Technical Details

Our PD/GEM patches can be downloaded from www.idc.ul.ie/mikael/hack/Softday_Bacterial_Ensemble.zip

Video of the performance and more information about the exhibition and be found at www.softday.ie/nlutds/

8. ACKNOWLEDGMENTS

We would like to thank the following people for making this work possible:

- James Clancy, Kathleen O'Rourke and Robert Hernan at Enterprise Ireland's Aquatic Toxicity Laboratory, Shannon, Co. Clare.
- The Science Gallery, Trinity College, Dublin.
- Our musicians, Aoife Caulfield (Violin), Maeve Garvan (Piano), Michael Coen (Bass), Aaron Mulhall (Drums).
- Technical support: Giuseppe Torre, Colm McGettrick, Darragh Pigott.

9. REFERENCES

[1] Lucey J. Water quality in ireland 2007 - 2008, key indicators of the aquatic environment. Environmental Protection Agency, Ireland, 2008.

[2] Cohen J. Monitoring Background Activities. In: Kramer G Auditory Display: Sonification, Audification and Auditory interfaces. Reading, MA, USA: Addison-Wesley Publishing Company; 1994:499-532.

[3] Cohen J. Out to Lunch: Further adventures monitoring background activity. In: Kramer G, Smith S ICAD'94. Santa Fe: ICAD; 1994:15-20.

[4] Gaver WW, Smith R, O'Shea T. Effective sounds in complex systems: the ARKola simulation. In: CHI'91. New Orleans, Louisiana, USA: ACM Press; 1991:85-90.

[5] Bly S. Multivariate Data Mappings. In: Kramer G Auditory Display: Sonification, Audification and Auditory interfaces. Reading, MA, USA: Addison-Wesley Publishing Company; 1994:405-416.

[6] Paine G. Reeds - a responsive sound installation. In: ICAD 2004. ICAD; 2004.

[7] Sturm B. L. Surf music: Sonification of ocean buoy spectral data. In: Proceedings of ICAD 2002. ICAD; 2002:1-6.

[8] Sturm B. L. Ocean buoy spectral data sonification: research update. In: Proceedings of ICAD 2003. ICAD; 2003:164-165.

[9] Fernström M, Griffith N. LiteFoot - Auditory Display of Footwork. In: ICAD. Glasgow: Springer-Verlag; 1998.

[10] Fernström M, Griffith N, Taylor S. BLIAIN LE BAISTEACH - Sonifying a year with rain. In: ICAD. Espoo, Finland: Laboratory of Acoustics and Audio Signal Processing and the Telecommunications Software and Multimedia Laboratory, Helsinki University of Technology,; 2001.

[11] Franco E, Griffith NJ, Fernström M. Issues for Designing a flexible expressive audiovisual system for real-time performance & composition. In: ICAD. Hamamatsu, Japan; 2004.

[12] Weiser M, Seely Brown J. Designing Calm Technology, Xerox Park, December 21 1995. Available at http://www.ubiq.com/hypertext/weiser/calmtech/calmtech.htm

THE ANCHOR MODEL OF MUSICAL CULTURE

Dr. Thomas Fritz
Max Planck Institute for Human Cognitive and
Brain Sciences,
Department of Neurophysics, Junior Research
Group Attention and Awareness
Stephanstrasse 1A, 04103 Leipzig
fritz@cbs.mpg.de

ABSTRACT

In a recent cross-cultural study with participants from an autochthonous African population (Mafa) and Western participants, it was shown that the recognition of several emotional expressions (happy, sad, fearful) in music are likely to be music universals [1]. The Mafa listeners (who were naïve to the Western music) were quite successful at recognizing the emotional expressions in the Western music, although their own music seems not to emphasize, or even comprise this musical feature. Here I propose a model, which is aimed at illustrating how different human musical cultures intersect and "anchor" in a set of musical features that are universally perceived, while also displaying culturally acquired specifics (see Figure 2), that accounts for the Mafa results. It explains also why musical universals cannot simply be determined by specifying the common denominator between the musical features of all cultures, which may actually not exist.

1. INTRODUCTION

It is highly likely that the creation and experience of music is in part furthered by an underlying universal physiology of the human being, because it seems plausible that some form of music has been invented by all human cultures, past and present [2]. In order to identify the physiological mechanisms at work it is useful to determine which musical features are recognized universally. However, in order to investigate musical universals, it is crucial to have a concept of what music may be. This is challenging, because the design features of music are variable and various [3], and the contexts where music is involved differ to a great extent between cultures [4]. Consequently, it is not trivial to agree on what music is. For example many "native" cultures do not even have a term for music at all [1], because it is an integral part of various rituals. If one tried to name a common denominator between what might be considered music in all cultures of the earth, there might be nothing at all besides possibly that it relates to some form of organized sound. However, this does not imply that there are no musical universals.

The investigation of musical universals with Western music stimuli requires participants who are completely naïve to Western music. Even individuals of non-Western cultures who have only listened to Western music rarely, and perhaps without paying explicit attention to the music (e.g. while listening to the radio, or watching a movie) do not qualify as participants because musical knowledge is also acquired implicitly, and thus shaped even through unattended listening experience [5]. However, since the efforts of early pioneers such as Erich M. von Hornbostel at the beginning of the last century, it has rarely been attempted to investigate human individuals who were completely naïve to Western music. Unfortunately, opportunities for intercultural comparisons between individuals exposed to completely incongruent music cultures are becoming increasingly rare, due to globalization. Western music culture mainly spreads with electricity supply (and thus the possibility to operate radios) and Christianization (through Western Christian song).

Evidence from intercultural and developmental studies in humans suggests that relatively basic musical features such as relative pitch, octave generalization, intervals with simple ratios, and tonality are possibly music universals (for a review see [6]). In a recent cross-cultural study with participants from a native African population (Mafa) and Western participants, Fritz et al. [1] showed the intercultural ability to recognize three basic emotions (happy, sad, scary/fearful) expressed in Western music (Figure 1). This indicates that even the supposedly complex musical feature emotional expression can be recognized universally for several emotional expressions in Western music. This is especially interesting for the model proposed here (Figure 2), because the musical expression of a variety of emotions like fearfulness and sadness seems not to be intended by the Mafa, and consequently they seemed to have recognized a putatively universal musical feature, which is not part of their of their own musical repertoire (their music cultural form).

Therefore, the study by Fritz et al. [1] is described here in greater detail. Both participant groups, the investigated Mafa and the Germans were naïve to the music of the respective other culture. The Mafa are one of approximately 250 ethnic groups that make up the population of Cameroon. They are located in the Extreme North, in the Mandara mountain range, where the more remote Mafa settlements do not have electrical supply, and are still inhabited by many individuals who pursue a traditional lifestyle, some of whom have never been exposed to Western music.

There have been previous investigations of the recognition of emotional expressions conveyed by the music of other cultures, but since the participants were not completely naive to Western music, these studies allowed conclusions about cultural specifics rather than music universals [7-9]. The study by Fritz et al. was designed to examine the recognition of three basic emotions as expressed by Western music (happy, sad, scary/fearful), using music pieces that had been used previously to investigate the recognition of these emotions in brain damaged patients [10-11]. Stimuli were computer-generated piano music excerpts with durations between 9-15 seconds, which were specifically designed to express the emotions happy, sad, and scary/fearful according to Western conventions such that they varied with respect to mode, tempo, pitch range, tone density and rhythmic regularity (download examples at http://www.sendspace.com/file/0bl7qa). During the experiment the music stimuli were presented from a CD player and only audible to the participant over headphones to avoid response biases through the experimenter. The participants had to indicate which facial expression from the Ekman archive (happy, sad, scary) [12] fit best with the expression of each music excerpt (forced choice).

The results showed that the Mafa recognized happy, sad and scary/fearful Western musical excerpts above chance (Figure 1), indicating that the expression of these basic emotions in Western music can be recognized universally.

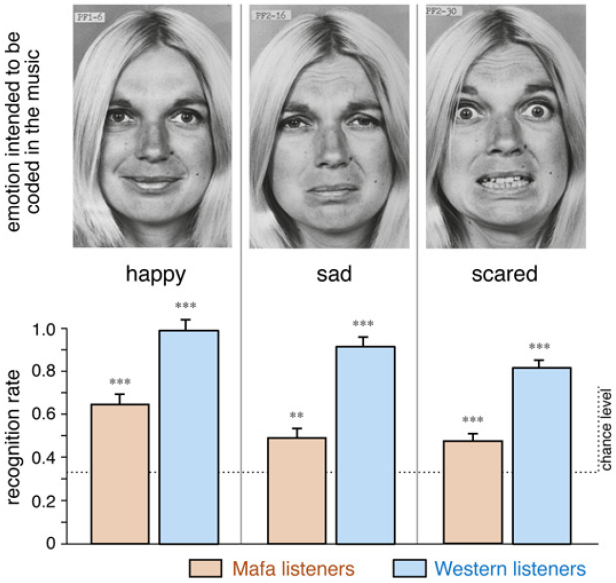


Figure 1. The figure shows the mean performance (M) in percent for the recognition of each emotional expression from the Ekman archive (above) in Western music excerpts, chance level: 1/3 (*** p<0.001, ** p<0.05), standard error (SEM), t-values (df = 20 for the Mafa listeners and df = 19 for the Western listeners), figure from Fritz et al. [1].

The expression of emotions is a basic feature of Western music, and the capacity of music to convey emotional expressions is often regarded as a prerequisite to its appreciation in Western cultures. This is not necessarily the case in non-Western music cultures, many of which do not similarly emphasize emotional expressivity, but may rather appreciate music for qualities such as group coordination in rituals.

Although some of the data presented by Fritz et al. [1] may be interpreted to corroborate the idea of music as a medium to universally mediate emotion, a possible absence of a variety of emotional expressions in Mafa music would rather suggest a different interpretation. If music were in its essence indeed a universal language of emotions, how come Mafa music seems to not express a comparable variety of emotions as occur in Western music? The appropriate answer to this is that although emotional expressions in music are perceived universally, this may not be the principal function of music (as already pointed out by Hanslick in his 1854 essay [13]). Despite the observed universals of emotional expression recognition one should thus be careful to conjure the idea of music as a universal language of emotion, which is partly a legacy of the period of romanticism.

2. THE ANCHOR MODEL OF MUSICAL CULTURE

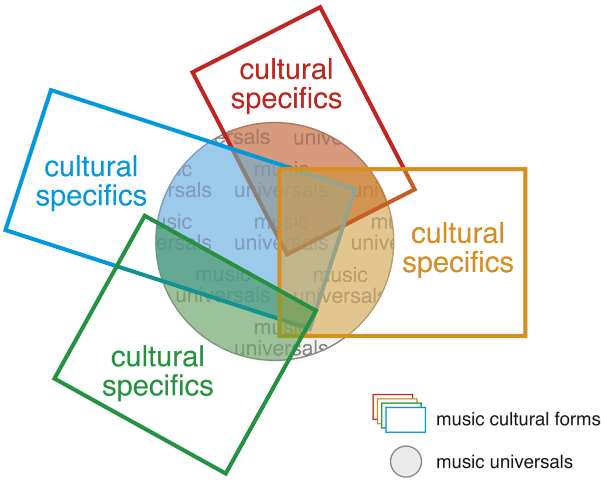


Figure 2. Anchor model of musical culture.

The model (Figure 2) suggests that all music cultures contain both music universals and cultural specifics. The more two cultures share a music cultural influence, the more their musical codes (music cultural forms) overlap. It suggests that despite a universally shared understanding of a partly common code (music universals) in which all music cultures "anchor", no music culture has implemented the whole set of universal musical features in its musical repertoire. Furthermore it shows how the musical repertoires of two cultures can be "anchored" in the set of music universals but do not overlap (the red and green boxes).

The question arises, why the Mafa music does not include a variety of emotional expressions like for example sadness and scaryness/fearfulness if the Mafa were capable of recognizing these expressions in the Western music. The answer may be that the recognition of emotional expression from music is not exclusively a musical capability, but instead a capability that evolved as an adaptation to a different challenge, and was then co-opted for music. While emotional expression may be a sub-category of the musical design feature a-referential expressiveness [3, 14], this does not entail that the capability for emotional expression processing is an exclusively "musical" capability. Like the capability for the production and perception of many other design features of music, emotional expression processing is probably a spin-off of one or several more general-purpose mechanisms. It has even been argued that all the so-called musical capabilities are such spin-offs, and that human music may thus hardly be regarded a special evolutionary adaptation [6].

The universal capability to identify emotional expressions in Western music is presumably at least partly due to the universal capability to recognize nonverbal patterns of emotional expressiveness [15] such as emotional prosody. Emotional prosody has been observed to be mimicked by Western music as a means of emotional expression [16], and other findings indicate that emotional prosody can be recognized universally [17]. This interpretation is consistent with the notion that similar emotion-specific acoustic cues are used to communicate emotion in both speech and music [18-19].

The discussed findings thus demonstrate that music as a means of emotional communicative expression, although probably universal, had to be culturally discovered, and probably transferred from a more general-purpose means of communicative expression. Emotional expression is clearly not a prerequisite for music. Music cultures may have discovered and developed emotional expression in music at some point, but this does not necessarily have to be the case. In Western music, emotional expression is possibly such a prominent feature, because Western music is the result of a very long cultural integration process, a common denominator between the many musical cultures. This probably promotes the cultural transmission of musical features that can universally be understood. In more local cultures such as the traditional Mafa culture, it is not necessary that the music is understood by people from different cultural backrounds, because the musical rituals are passed on to the following generation along with a culturally learned semantic imbuement.

The Anchor Model of Musical Culture (Figure 2) provides a theoretical framework to discuss music cultural intersection, and hopefully, to further our understanding of what musical universals are, and how they relate to musical culture.

REFERENCES

[1] 1. Fritz, T., et al., *Universal recognition of three basic emotions in music.* Current Biology, 2009. **19**(7): p. 573-576.

[2] 2. Nattiez, J.J., *Under What Conditions Can One Speak of the Universality of Music?* The World of Music, 1977. **191**(2): p. 92-105.

[3] 3. Fitch, W.T., *The Evolution of Music in Comparative Perspective.* Annals of the New York Academy of Sciences, 2005. **1060**(1): p. 29-49.

[4] 4. Cook, N., *Music: A very short introduction.* 1998, Oxford: Oxford University Press.

[5] 5. Tillmann, B., J.J. Bharucha, and E. Bigand, *Implicit learning of tonality: A self-organized approach.* Psychological Review, 2000. **107**(4): p. 885-913.

[6] 6. McDermott, J. and M. Hauser, *The origins of music: Innateness, uniqueness, and evolution.* Music Perception, 2005. **23**: p. 29-59.

[7] 7. Gregory, A.H. and N. Varney, *Cross-Cultural Comparisons in the Affective Response to Music.* Psychology of Music, 1996. **24**(1): p. 47-52.

[8] 8. Balkwill, L.L. and W.F. Thompson, *A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues.* Music Perception, 1999. **17**: p. 43-64.

[9] 9. Balkwill, L.L., W.F. Thompson, and R. Matsunaga, *Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners.* Japanese Psychological Research, 2004. **46**(4): p. 337-349.

[10] 10. Gosselin, N., et al., *Amygdala damage impairs emotion recognition from music.* Neuropsychologia, 2006. **45**: p. 236-244.

[11] 11. Gosselin, N., et al., *Impaired recognition of scary music following unilateral temporal lobe excision.* Brain, 2005. **128**(3): p. 628-40.

[12] 12. Ekman, P., *Pictures of facial affect* 1976, Consulting Psychologists Press: Palo Alto, CA.

[13] 13. Hanslick, E., *Vom Musikalisch Schönen*. 1980.

[14] 14. Fitch, W.T., *The biology and evolution of music: A comparative perspective.* Cognition, 2006. **100**(1): p. 173-215.

[15] 15. Eckerdal, P. and B. Merker, *'Music' and the 'action song' in infant development: An interpretation*, in *Communicative musicality. Exploring the basis of human companionship*, S. Malloch and C. Trevarthen, Editors. 2009, Oxford University Press: Oxford. p. 241-262.

[16] 16. Juslin, P.N., *Communicating emotion in music performance: A review and a theoretical framework*, in *Music and emotion: Theory and research*, P.N. Juslin and J.A. Sloboda, Editors. 2001, Oxford University Press: New York. p. 309-337.

[17] 17. Scherer, K.R., *The role of culture in emotion-antecedent appraisal.* Journal of personality and social psychology, 1997. **73**(5): p. 902-922.

[18] 18. Juslin, P.N. and P. Laukka, *Communication of emotions in vocal expression and music performance: Different channels, same code?* Psychological Bulletin, 2003. **129**(5): p. 770-814.

[19] 19. Scherer, K., *Expression of emotion in voice and music.* Journal of Voice, 1995. **9**(3): p. 235-248.

FROM METAPHOR TO MEDIUM: SONIFICATION AS EXTENSION OF OUR BODY

Joachim Gossmann

UC San Diego
Center for Research and Computing in the Arts
9500 Gilman Drive La Jolla, California 92093-0037
jgossmann@ucsd.edu

ABSTRACT

Following Marshal McLuhan's perspective on media as *extensions of man* [14], sonification for the generation of knowledge can be regarded as an extension of our auditory sense toward previously imperceptible properties of our environment. Investigating our own involvement from an ontological perspective allows us to generate conceptual handles for the research, development and use of tools for sonification and the implied extension of our physical body through technology. Based on the nature of our bodies as mediators between the shared exterior and the individual interior, a model of three problematic areas of our extended bodies is presented: the *cognitive*, the *physical* and the *extended*.

When we research, design and develop new applications and methods in sonification, we investigate the models and metaphors used in each of these areas, but it is only when we *use* the developed applications that we actually understand what potentials of perception and exploration we are provided with. It is therefore not sufficient to only build an exterior apparatus: The extended body is each of our own—each researcher and user of sonification develops an individual relationship to all affordances provided.

1. INTRODUCTION

Sonification has been defined as

> The transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication and interpretation [1].

According to Thomas Hermann's summary, research related to the field of sonification is centered on the relationship between data and sound: the mapping of data into schemes of sound synthesis and interaction [9]. We are looking at an apparatus that performs a deterministic transformation of a quantified but imperceptible phenomenon into one which can be directly perceived. The resulting modality, *sound*, is well defined on a technical level: Oscillations of pressure transmitted through solid, liquid or gas, composed of frequencies within the range of human hearing. From this perspective, sonification means the encoding of data into audible vibrations.

The demand of *perceived relations* however requires a quantitative model of perception. Empirical approaches that can provide a handle in this area can be found in the interdisciplinary research field of auditory perception, for example psychoacoustics [16], auditory scene analysis [17], music cognition [18], study of phenomena of *presence* [19]—or the environmental approach to sound perception [12]. But the installment of quantifiable handles on what occurs to us subjectively runs into problems from two directions: the dependence of perception on the subjective involvement of the perceiver on one hand, and the ineffability of perceptual qualities on the other.

From my own experience, I would like to highlight the active role of the listener in the process of perception. In my work as a sound engineer striking perceptual changes seem to occur when my involvement and intention with the sound shifts. Listening from the perspective of the engaging performance of a musician, the balance settings of the mixing console, the loudspeaker projection, recreation of perceptual depth and space, impact, width, pressure etc. all make the sound occur in a noticeably different way. It becomes very evident that what we *listen for* in a sound, the expectation informed by our intentions, can change its perception.

Secondly, our auditory experience can only insufficiently be rendered into words, much less physical or arithmetic expressions. When we attempt to communicate and share our perceptions, we seem to be confronted with a lucent but unapproachable realm of qualia, ephemeral and fleeting impressions connected to inherent or attributed layers of meaning that make up our conscious experience. The description of sound qualities often occurs through synaesthetic and poetic metaphors, comparable to the way we describe the experiential qualities in a sip of wine. Whether working with a composer or as a sound engineer attempting discuss a specific sound with a musician, the language used is often suggestive rather than precise, and harbors a constant danger of sliding into a situation comparable to the one described in Hans Christian Andersen's ”The Emperor's New Clothes”.

Both this quantitative ineffability of auditory perception and the inherent openness of *listening ”for”* cause resistance to the intendend deterministic creation of *perceived relations* shared among different listeners, as these phenomena are hidden from the world shared among different individuals and reside in the area of internal and subjective cognitive processes. Before we return to this specific problem, we would like to insert a more general excursion into the relationships between our perceptions, models of reality and our cognitive approach toward the world we find ourselves in.

2. EXCURSION INTO THE NATURE OF INVESTIGATION: MODELS

When botanists categorize plants, one possible question to ask is for the physical structure of the plant—for example, the structure of its blossoms. This investigation produces so called inflorescence diagrams: Simplified representations assembled from shapes

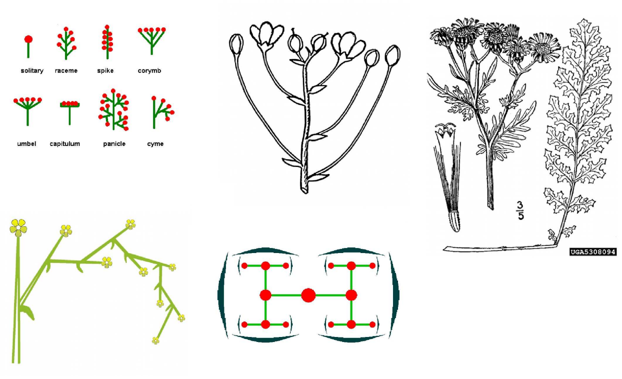


Figure 1: Inflorescence Diagrams (source: Wikimedia)

of clear geometric structure that afford formalization and categorization and can be described, stored, processed and recreated easily.

Interesting about the nature of these diagrams is the degree to which they depart from the actual shape and impression of the original plant. Sometimes, straight sticks and circles representing the flower are enough for a satisfying circumscription. In certain cases, curved structures need to be introduced to depict petals and leaves, or for a correct representation of the structure itself. This requires decisions about the specific geometric shape used: A human observer providing a *best judgement* for the shape used becomes apparent. The more life-like the representation of the plant is required to be to remain accurate, the more unquantified elements appear in the diagram that forego a straightforward modeling of the representation for example by a computer. At the extreme end of this continuum is the creative work of a human artist capturing the essential elements of the plant's structure from subjective point of view—a representation that can no longer easily be recreated or processed. We have traversed a space from simplified representations of complete quantitative transparency to representations of perceptual accuracy that however rely entirely on subjective human perception.

Simplified models are created to remove information from the appearance of the actual plant, allowing us to handle them for a specific purpose—for example to classify them by blossom structure. From this moment on, the plant is no longer regarded as a unique individual, it is handled according to a suiting simplified model chosen as its representation. This process, that we are generally unaware of, allows us to access the world and to use the structures we find for specific purposes according to a matching model of reality that we infer: the metaphors we live by [7]. We are replacing objects we find in our environment with simplified models constructed from structural elements that are implied in the attitude or question of our own approach.

3. MODELLING PERCEPTION

How can we find a model for our perception that we can use to make the sonification apparatus more relevant to us? Through perception, the exterior world becomes present to our thinking and action. How can we describe the way in which perception transgresses the line between what we may describe as our exterior and interior worlds? We tend to view our sensory organs—ears, eyes, nose, et cetera—as parts of our physical body. But even Rene Descartes, the philosopher known for his support of a dualistic separation of body/physical exterior and mind (*res extensa* and *res cogitans*), regarded sensing (sentire) already as a part of thinking (cogitare) [6]. Models that describe our senses not as a passive receptors of stimuli from the exterior, but as a part of our mind and already implicating cognitive activity, are pervasive. Investigating the nature and appearance of perceptual illusions, R.L. Gregory generated the model of a perceptual *hypothesis generator* that receives input from three different directions [5]:

- bottom up: from the sensors in our body that are connected to the physical environment
- top-down: from previous experience and accumulated knowledge about the exterior world
- sideways: from being set for a task

Only one of the streams in this model enters this structure from the exterior world, while the other two information streams are created by cognitive activity of the perceiver. Next to these three inputs, this *hypothesis generator* has two outputs: on the one hand the appearance of conscious perception (qualia) and on the other hand a behavioral action.

More recently, cognitive models were developed in the area of artificial intelligence and robotics research: From the perspective of machines with intelligent behavior and environmental awareness, the relationships between *mind*, *body*, *perception* and *action* have been framed into models of *embodied cognition* that can be traced back to philosopher Maurice Merleau-Ponty [8] and cognitive linguist George Lakoff [7]. Regarding cognition and perception as a distributed process in which the different parts of the body are already actively involved is in elegant correspondence to the distributed sensing and actuator systems exchanging information used in robotic design [3]. This in turn opens the path toward the interpretation of our body as an extensible structure involved in *active perception* as it can be found in the discourse of post-humanism [4]. The act of perception is no longer implying only cognitive activity, but in fact a senso-motoric loop: the performance of physical movements such as involuntary eye motions or scanning across surfaces with the tips of one's fingers. This effectively dissolves the ontological distinctions between *mind*, *body*, *thinking* and *action*.

Before loosing focus in the appreciation of the power of this model to enable the emergence of complex self-regulatory systems that display intelligent behavior, we need to remind ourselves that the focus of this presentation is the question what the respective models afford to us as participants in the design and use of sonification. An analysis of information flow in the cybernetic post-human perspective of our embodied cognition will not provide us with a sufficiently reliable concept of *subjectivity*. We have to insist on an approach toward the world that is based on our *Being* and *Caring* for the world when we are looking for answers to questions such as:

What does a specific model afford us in our perceptual access to the world through audible data relations? How does specific model of cognition and perception allows us to address ourselves with the apparatus we are constructing? What possible actions can we take to enhance our own subjective attitude toward these model mechanisms and the way we use them?

The deconstruction of the subject is the necessary outcome of empirical self-analysis, yet the questions we want to consider imply us as personal designers and users of sonification. We will therefore base our notion of subjectivity on a concept of *personal involvement* and *care* that follows Heidegger's assertion that *the existential purpose of Being is Care* [2]. On this basis, we can regard the sensimotor extension of our body provided by the sonification apparatus as an expansion of our *care for the world* towards previously imperceptible relations in abstract data. This extension is not exclusively *outwards* however: the successful use of new tools requires an extension into the *interior*, into our cognitive approach toward the world. We hope to make this perspective more transparent on a short expedition into the discourse of *embodied interaction* introducing the different areas of the *extended body* that are proposed here as an organizational strategy for the functional elements that sonification engages and the questions that might be implied in them.

4. ONTOLOGICAL APPROACH TO THE COGNITIVE BODY

An example often cited in the context of ”embodied interaction” is Heidegger's description of a shoemaker using a hammer to drive a nail into the heel of a shoe in order to repair it [11, 2]. A well trained craftsman will be so versatile with the tool that it functions like a part of his body, allowing the shoemaker to focus his *care* completely on driving the nail into the heel without thinking about how to handle the hammer. The hammer becomes part of the shoemakers skilled arm, a transparent physical extension to his hand: the established in-order-to has become intuitive and familiar through practice. A good hammer will enable the craftsman to provide exactly the right transformation of the force exerted by the body and gravity to allow a maximum amount of control over the nail. With the terminology of McLuhan, we could say that the material that the hammer consists of is becoming a medium for the activity of the craftsman through its *use as* a hammer [14]. When the hammer breaks however, *using-it-as* a hammer is no longer possible: The *care* is shifted toward fixing it, for which in turn other tools might be applied.

Thus the hammer in this example can be in two different ontological states: it can be a part of the craftman's extended body by which he approaches the exterior world to care for the shoe, or be a part of the exterior world and itself a recipient of care. *Using* something *as* allows us to extend our bodies dynamically by turning models and metaphors into media for our intentions and using them in our improvisatory approach to the world. Paul Dourish cites Suchman, who drew attention to the improvisatory nature of our moment-to-moment actions:

> The sequential nature of action is not a formulaic outcome of abstract planning, but rather is an improvised, ad hoc accomplishment, a moment-by-moment response to immediate needs and the setting in which it takes place [11] .

We are improvising our way through life, using the objects around us as media for our intentions, depending on what we need to get done and what setting it occurs in. So much for the relationship between our physical body and its extension by exterior objects. We would now like to extend this perspective to the models we use in our cognitive access to the world: the metaphors underlying our actions that are not externalized into a specific use of our physical body or the objects we find in our environment, but that we use in our *thinking about the world*.

Science philosopher Paul Feyerabend argues against formalized scientific methods and for an anarchistic use of all available models in our access to knowledge [13]. Feyerabend's *epistemological anarchy* can inspire an extension of the concept of the *use-as* of exterior objects to the metaphors we use in our cognitive access to the world. If we allow the exterior world and our physical body to be used in our daily improvisatory performance, why not care in a similar way for the cognitive and perceptual tools we use to approach the world? This is possible if we have access from both of the ontological states we have described above for the shoemaker's hammer: if we are able to construct suitable metaphorical models and then are also able consciously use and apply these metaphors in the way we handle the world.

These necessarily condensed considerations motivate us to separate three ontological regions or realms of *body* that will allow us to build a structure into which we can organize the problematic areas within the field of sonification:

5. THE THREE REGIONS OF THE EXTENDED BODY

First, our *physical body* is the most obvious interface between our interior and our exterior worlds, between what we perceive as belonging to ourselves and that which belongs to the exterior space that we share with others, or—as Heidegger describes it—between the world and that which is *not* not me [2]. The body is the medium by which we are connected to the exterior in perception and action, and the locus of *sensimotor knowledge* [15].

Secondly, in our daily lives, we use apparatuses that modify or *extend our physical bodies*, providing additional affordances in our approach to the world: exterior objects we use as tools according to a learned or developed scheme of action that extend our interface with the exterior world. These *media* in the sense of McLuhan [14] can provide extension to both the reach and capabilities of sensing as well as behavioral action and ideally provide a successful coupling between extended sensing and extended action.

Thirdly, in addition to this realm of external objects that we use according to a metaphor to extend our body as media for our action and perception, another realm of metaphorical use patterns can be found in our *cognitive approach to the world*: the sensory, abstracting and behavioral capabilities we have made available to us. This third realm of body is differentiated from the projection of metaphors on the physical body and exterior objects: It describes our capability to shift the nature of our cognition and perception, changing our involvement with the world surrounding us: in my practice as a sound engineer, I learned to listen to music *as* an engaging performance, a sound quality, a tight mix, a spectral distribution, a technical transmission, et cetera, all of which correspond to different perspectives on the *audible* sound between which one can change at will.

Perceptual effects caused by shifting cognitive models can be observed in everyday life: an artwork may look quite different before and after we have listened to an art historian provide us with context about its making and historic significance. The way we drive a car might change drastically after we have attend theoretical driving lesson. The carefulness by which we handle a piece of technological equipment might change after we learn how expensive it was. These effects can be attributed to the two cognitive inputs to Gregory's *hypothesis generator* [5]. Some of the cognitive models are obviously not accessible from the ontological perspective of *detachment*: for example, we obviously have difficulties to hear the audible vibrations produced by a person speaking in our native tongue as anything but *language*. However, we can for example choose to listen and *pay attention* to a specific person at a cocktail party, or with some training choose to follow the viola voice in a string quartet recording.

There is a continuum between what is accessible to our conscious choice and will and that which we seem to be simply subjected to—a grey area that is somewhat comparable to the continuum in the ontological status of external objects from being detached from our body to becoming so familiar and integrated that they effectively become transparent parts of it.

The aspects of our cognition and perception which can both be observed and consciously used in our approach to the world constitute what we would like to describe as the third area of the extended body: the *cognitive body*. It contains the analytical tools we have at our disposal to access the world, the focus of our attention, learned schemes, models and thought structures that we can use to voluntarily shape the way the world occurs to us.

We can expand the *congitive body* by learning to *see-things-as* and *consider-things-as*, or on the contrary, buy into the Zen ideal of *not* seeing the world *as* something, and let the world occur to us differently thereby. There is empowerment in keeping our minds flexible in the approach of something unknown, when we want to learn about something unfamiliar. We may venture to ask if Feyerabend's demand for epistemological anarchy should receive more attention in the education of the young.

Before moving back to the topic of sonification we would like to summarize: Cognitive models available to our improvisatory behavior of *thinking* can inform, influence and educate our percpetion, abstraction and thereby our behavioral/physical approach toward the world.

6. THE EXTENDED BODY IN SONIFICATION

Regarding the field of sonification, we can now start to place the different strategies under investigation into the three realms of body and consider them in their dual ontological status as *metaphorical models* that await construction and care and as *media in use*. This may grant us a better overview of elements that play a role in the application and use of sonification and their interrelation. From the perspective of care, we can analyze the questions that are relevant to each area. From the perspective of *use*, the three layers form an interconnected senso-motoric media channel that is ideally transparent to the data relations present. Due to the scope of this presentation, we can only deliver a pointillistic collection of possible considerations. The gentle reader is invited to draw up a corresponding table for their own sonification research and development project.

6.1. The physical body

Our physical body appears (for example) as a biological senso-motoric system: It affords us with sensory reception and behavioral action. We are obviously highly experienced in its use as a versatile medium for whatever it is we may be doing. In most cases, our body in fact becomes transparent for our intentions leaving us unaware of how it is used specifically in the activities we are involved in.

When we regard the physical body in the ontological status of a *recipient of care*, we can analyze its audition-related aspects: how can we describe the affordances of hearing and physical action in the most fitting way for what we plan to be doing? How can we allow our physical body to hear a sound and interact with a sound generating strategy in the most effective way? Maybe with gestures? Many aspects of our bodies are involved in the way sound occurs to us, most obviously the ear itself and its various physiological components, but also our shoulders and head shape the sound by characteristic reflections and diffractions depending on its direction of arrival. These spatial cues can be disabled when the sound reaches the cochlea through bone conduction, as underwater sound, or is generated inside or at the body (such as wind breaking at our outer ear), or otherwise bypasses shoulders, head and pinnae, for example by the use of headphones. Sound as vibration may also be detected through the skin by our tactile sense. Simultaneously, we can use physical movements that trigger, scan and explore a sounding structure. What is the best ergonomic range of motion? Where are we most sensitive to changes, what are the preconditions for the best motor control? Caring for our physical body in the context of sonification implies an investigation of how to best extend the relevant affordances toward the data in a loop of *active perception*: How does the aquisition and use of sensimotor knowledge that is postulated by Noë operate?

Using these auditory affordances of our physiological body in the ontological state of a *medium*, the perspective shifts. We are no longer in contact with theoretical models of how our body is supposed to operate or with measurements and descriptions, but with what we as individuals can actively do to get in contact with the *data relations* we would like to investigate or perceive. We can make experiences, develop usage strategies, train ourselves in them, become better at it. We are becoming involved in an active physical improvisation in order to *hear better*. We can change the positioning of our ears—approaching the object emitting the sound, using our hands to amplify or block the sound. We can also move our body to touch something, scan, trigger, move, organize, etc. external objects to name only very few of the in fact innumerable possibilities of how we can use our bodies in the task of *active listening*. While far from conclusive or complete, these considerations may suffice to support that our physical bodies afford more *action-in-perception* for the designer and the user of sonification than the frequently implied model of listening to a sound in a passive position of sitting, and highlight the important role of the performance based on the approach of the individual listener that often goes unconsidered and is replaced by an implied passive and standardized model human.

6.2. The physical-extended body

The extended body contains all aspects of the external apparatus we are using with our physical body to get in perceptual contact with the data relations. These extensions can be *tools* such as the hammer in Heidegger's example, but they ideally extend the affordances of physical *sensing* as well as those of *action* simultaneously. In the case of sonification, the extension of sensing is implemented with *audible vibrations*, while the mode of *action* is a free choice. On the one hand the creation of an environment that enables *action* of a participant while producing *audible vibrations* implies a suitable physical *display system*. On the other hand the information encoded in the display targets our *cognitive body*: According to McLuhan's analysis that ”the content of a medium is always another medium” [14], what is encoded in these *audible vibrations* connecting the external apparatus to the physical body is always in fact another medium: content that targets our freedom to perceive, the potentials of our cognitive body to focus and explore specific aspects and elements. The *care* for the exterior sensimotor extension can therefore be split into the aspects of *display* and *encoding*.

6.2.1. Display

In the *care* for this area of the extended body, we can find an optimization of the display infrastructure as well as its physical setup to target the senso-motoric capabilities of the body in the most effective way. This can mean for example the design or choice of equipment such as loudspeakers, headphones, converters and amplifiers, video screens, projectors and interaction devices, ergonomic considerations in the setup, the choice and treatment of the room, the removal of unwanted sound sources. What frequency range can our ears pick up? What is involved in making a 3D screen that avoids the sense of nausea? Is the table we have placed the tangible interface elements on too large or too small?

6.2.2. Encoding

Also in the physical-extended body, we find the transformation and modeling of the data relations into audible vibrations, possibly mediated by physical interaction of the participant. The resulting vibrations are designed to target our auditory perception, mediated through the physical display setup and our own physiology.

The strategies of mapping and modeling in the processes of sound generation in the external apparatus as well as the provided affordances of interaction and display are of special interest in the design of the *extended body* and a central concern to design of sonification tools. They can in fact be regarded as a mirror image of our own *cognitive body* of auditory perception: our concepts of what we supposedly can actually *hear and distinguish by listening*, what we consider to be *relevant features of sound* that allow us perceive data relations most clearly. *The sound generation strategy implies the cognitive model we expect to apply when we listen.*

A popular model of auditory perception found in this context is listening for pitch. Other models of auditory cognition are for example derived from the field of music cognition: meter, rhythm, harmony. The strategy of auditory icons employs a model of environmental perception placing relevance on metaphorical references to objects found in the physical environment [12]. In other fields, auditory perception is seen under the model of segregated streams: our capability to distinguish different simultaneously sounding sources of audible vibrations in our environment and to pay selective attention to them [17]. In my own work experimental work I regarded the auditory body as a receiver of information quanta in the form of sine waves accumulating into additive spectra [20]. Examples of strategies that are attributed to the cognitive body of auditory perception include:

- Frequency/Pitch
- Amplitude/Volume
- Tonality/Harmony
- Rhythm, Meter
- Timbre
- FFT/Spectral composition
- Expectation/Form
- Localization
- Stream Segregation
- Intuitive impression of physical process modeled

But are these really the models and metaphors that are closest to the way we get involved with sound? Or do we only use them because we invested so much effort in working our cognitive way through them during our musical training?

For the context of media art, David Rokeby describes the *use* of an interactive computer installation as the development of a belief system about how the installation works [21]. In adopting the apparatus we explore and test as we attempt to extend our sensimotor knowledge into the realm of the physical-extended body. Like the Hammer in Heidegger’s description, we have the ability to become so familiar with this extension that it in fact becomes invisible in our use, but this use of the tool or apparatus is not only dependent on its own making and structure and the strategies that it externalizes, but also on the purpose and context of its use. Our expertise and training grows every time we engage the apparatus for what we are involved in. Do we understand how to use the tool well? Is this tool the right one given our way of thinking about the problem and the context we are using it in? Is it us who do not understand how to use the apparatus properly, or is the apparatus not suitable for what we are trying to achieve? Do I need to change my cognitive approach, or are there problems with the way the display addresses my physical body, or with the encoding strategy used? What affordances do I have with my cognitive and physical bodies in relation to this apparatus? What controls or functions would it need in order to be more suitable to what we are trying to do? These and other considerations highlight the interconnectedness of all three areas of the extended body in use. Once we start to use the tool we have been building, the physical body, it’s technological extension and the cognitive body appear in the shape of a tunnel that we orient and apply toward what it is we want to get done: Perception of data relations through sound, for example.

6.3. The cognitive body

The cognitive body allows us to choose to a certain degree what we *hear sounds as*. How can we train and expand this ability? How can we make the apparatus usable from as many cognitive perspectives as possible and how can we communicate these perspectives to a possible user? Can we expand any of these aspects through education about the sound generation, music theory and analysis, or ear training of technical or musical orientation? What strategies could we be educated in that could later be used in the context of the sound generation strategies implemented in the externalized apparatus? In Thomas Hermann’s approach to Model Based Sonification for example, the quality of the sound in terms of the quantities of pitch, timbre, amplitude that are usual targets for parametric mapping become secondary consideration to the behavior of a sound-producing data-structured physical model [9]. This seems to resonate with Gibson’s concept of *direct perception* [10]. So what is the best attitude a listener should approach these sounds with?

Contributing form my own experience as a sound engineer once more, a music production is most successful if it works on all relevant cognitive perspectives, if there are many different ways of listening and all of them deliver a rich and rewarding experience, none reveal striking flaws: Frequency composition, dynamic range, the creation of spatial depth or width, pressure, presence, transparency and artifacts of data compression are among the more quantifiable aspects. But there are others that can only be accessed through subjective and intuitive criteria that quickly start to preclude quantitative evaluation—the quality of the captured instrumental performance or electronic sound, engagement, musicality, expressivity, etc cetera. Each form of listening implies a different internal attitude of the listener.

What seems of greatest importance is the *freedom of the listener to apply all areas of his extended body to the exploration of the sound in an improvisatorial manner*, shifting between as many different cognitive tools along the way as possible, in order to discover the best position of the extended body that allows the clearest listening perspective on the *relations* to be represented by sound.

7. SUMMARY

The apparatus of sonification is accessible from two different perspectives: its design or enhancement, and its use. Shifting the care toward design and enhancement of the apparatus brings the metaphorical handles of each area to the foreground. In its use on the other hand, the apparatus becomes a medium extending our auditory perception toward the relations found in the data we are investigating. Our improvisational skill in each area of the body involved can now be explored, experienced, trained.

7.1. Organizing the metaphors

The model of the *extended body* with its three regions can be seen as a shelf on which the problems that the design and use of sonification implies can be organized and seen in overview. This can provide us with better access to what is needed for a successful *translation of data relations into perceived relations*.

The *cognitive body* contains considerations about our perceptual approach to sound, how it is influenced by experience and task orientation and what possibilities we have to both expand and dynamically shift between the different cognitive involvements in listening. This will allow us to find better criteria for the models we use for encoding data into vibrations through interaction inside the apparatus.

Focusing on the *physical body* allows us to consider the affordances of its sensimotor capabilities for exploration and active perception.

The *physical-extended body* finally is the locus for considerations about the design of the display system and the implementation of modeling, mapping and interaction used in order to address the physical and cognitive bodies most efficiently.

7.2. Using the medium

Shifting our consideration towards the medium *in-use* we enter our models from the perspective of each of our own bodies, from the perspective of our individual subjectivity. Ideally, the apparatus will become a transparent medium: moving the apparatus with our own motoric skills and sensing the responses, we create a loop of active perception that ideally extends our ability to approach the world.

However carefully we design the apparatus, the *relations* are invisible from the perspective of *care* and *detachment* - they only appear in the use of the implemented apparatus *for sonification*. This use is each of our own responsibility - every person can choose to actively improvise in order to hear the *perceptual relations* better, or approach the environment and their life in any other way they see fit.

From this perspective it becomes possible to shed light on the relationship between *sonification* and *art*, which continues to be an area of much confusion.

8. CONTEXTUAL EXCURSION: EXPRESSION, NARRATIVE AND THE COGNITIVE BODY OF LISTENING

Sonification implies that we as participating listeners are interested in the data underlying the auditory representation: the sound becomes part of a medium the data is observed through.

Often, the word *sonification* is used in contexts in which the sound is related to or generated from non-musical data, but the connection between the *perceived relations* and the data they were created from remains a mere suggestion: The transformation is engaged as an inspirational element of a narrative, such as the suggestion of a specific place or the evocation of an invisible or imaginary structure. The interest is diverted from the investigation of the actual phenomenon that produced the data into a narrative of artistic expression. In the cultural context of audio-art and music, the interest of the listener that the ”*extended body* is oriented towards in order to *hear better*” becomes the expression of an artist or an artistic collaboration, or the inner imagery that is evoked by the sounds but contributed by the listener: due to the different nature of intentional involvement, the origin and structure of data occurs to the participating listener with an essentially different perceptual mode of *aesthetic appreciation*.

Gustav Holst’s 1916 composition ”The Planets” is an orchestral suite in which the listener’s interest is not the retrieval of knowledge about celestial bodies of our solar system, but the sense of being absorbed and entertained by an imaginary dramatic narrative based on an astrological interpretation of characters attributed to each of the planets. In musicology, this genre of music with a suggested *program* that serves as a launch pad for the interior imagery of the audience is called *program music*. Other examples include Beethoven’s ”Pastorale”, Berlioz’ ”Symphonie Fantastique” or Richard Strauß ”Alpine Symphony”.

More contemporary interpretations of program music that are sometimes regarded as *sonification* can for example be found in Alvin Lucier’s piece ”Panorama” of 1993, in which a Trombone traces the outline of an alpine mountain range by sliding along micro-tonal intervals. While it might be possible to indeed re-trace the mountain range from listening to the piece, it can hardly be said that the listener will have any interest in learning more about the alps or the mountains through this mediation of elevation data - the interest of the audience is captured by intricate beating patterns resulting from the microtonal glissandos against the partials of the simultaneously ringing chords of the piano—while harboring a mental image of the alpine skyline. Albeit a very poetic experience generated through a composition method informed by physical properties of sound, the implied involvement of the listener is nevertheless akin to imagining planets spinning in space while listening to the powerful orchestral textures of Gustav Holst. The *data* are not a point of investigation: their origin is taken as a source of poetic imagery that is projected onto the *perceived relations* which constitute, together with the visual impression of the trombonist and the pianist, an *aesthetic experience*.

Neither the composer, not the performers or the audience are sincerely interested in extending their auditory world access towards relations in the underlying data structures *behind* the sounds, which would be the purpose of becoming involved in sonification under the perspective we have laid out in the previous sections of this article.

Without intending to offend my dearest friends and colleagues, the works of *sonification* in which environmental, climatological, demographic, geophysical data are not used to learn about the underlying phenomena but as a narrative reference within a context of an aesthetic strategy can form a long list. The interest in *perceived relations* is shifted towards an interest in *sounding good, (almost) like music* or *fascinating*—in some instances it can even be said that these relations do not matter at all but are in fact only used as a reminder of the context of the data and their origin. If successful, the aesthetic experience achieved can justify this approach - in less successful cases, the participating listeners get stuck between being unable to read patterns and structures from the sound while not being able to enjoy it as *music* either. The audience is then lead into a confusing space of *bad sounding suggestion*.

On the other hand artistic, creative and expressive expertise and sensibility DO have a strong purpose in sonification: In the cognitive models a listener can apply to approach sound, which are largely inaccessible to quantitative description. In this area artistic practice, as *a cognitive body regarding the creation and interpretation of meaning through sound,* is in fact an indispensable aspect of the extended body. It can inform both the strategies implemented in the external apparatus as well as in the possibilities of *how to listen*. The sound-design strategy, the effective mapping of parameters or assembly of rendering models can all benefit greatly from an access to the palpable expressive potentials of sound structures that may otherwise be the subject of a music composition or production, in a similar way in which the accurate representation of a plant may require subjective human judgement. This contribution does not require the composer or artist to work with *data* however. For an access to *new ways of listening*, it may for example be revealing to investigate the work of composers of the classic modern period such as Schaeffer, Stockhausen and Xenakis who approached technological creation and modification of sound with the intention of creating new aesthetic strategies [22, 23, 24]. Sound Art is a necessary activity and experience both to gain deeper access to the potentials of meaning encoded in sound, and to advance the openness of listening ”for”.

9. CONCLUSION

I hope it became transparent that from the perspective of *personal involvement* in the care for and use of sonification, the contribution of musicians, artists, composers et cetera is not so much in the area of creating *aesthetic experiences related to data,* but in the *expansion of cognitive models available to the actively exploring listener*. These conscious strategies to approach the perception of sound are in turn implemented, as mirror images, in the encoding of data occuring in the sonification apparatus: it is the individual listener who has to adopt this extension as his or her own in order to *listen to audible data relations*, it needs to *fit* the listener's cognitive and physical body. This generates questions that we may ask ourselves when we get involved with sonificiation:

As a researcher, developer and creator of sonification tools, the question becomes: How can I enable the listener to take better care for the perception of *audible data relations* that he or she is involved in?

As a listener, the question is: How can I use all tools available to me in order to *hear better*?

What I have attempted in this article is to create a framework of conceptual analysis to support us in extending our auditory sense toward structures in our environment that are otherwise imperceptible.

I firmly believe that we are in the process of scratching the surface of what we are actually able to hear.

10. ACKNOWLEDGEMENTS

I would like to thank Jeff Kaiser for being an inspiring reading partner for much of the literature that formed the basis of this paper, and Volker Straebel for the argument about the relationship between *sonification* and *program music*.

11. REFERENCES

[1] G. Kramer, *Auditory Display: Sonification, Audification, And Auditory Interfaces,* Westview Press, 1994.

[2] M. Heidegger, *Being and Time,* Harper Perennial Modern Classics, 2008.

[3] A . Clark, *Being There: Putting Brain, Body, and World Together Again*, The MIT Press, 1998.

[4] N.K. Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics*, Literature, and Informatics, University Of Chicago Press, 1999.

[5] R.L. Gregory, Knowledge in perception and illusion., in *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 352, Aug. 1997, pp. 1121-1127.

[6] R. Descartes, *Meditations on First Philosophy,* NuVision Publications, LLC, 2007.

[7] G. Lakoff and M. Johnson, *Metaphors We Live By*, University Of Chicago Press, 1980.

[8] M. Merleau-Ponty, *Phenomenology of Perception*, Routledge, 2002.

[9] T. Hermann, Taxonomy and Definitions for Sonification and Auditory Display, *Proceedings of the 14th International Conference on Auditory Display*, Paris, France: 2008.

[10] J.J. Gibson, The Ecological Approach to the Visual Perception of Pictures, *Leonardo,* vol. 11, Summer. 1978, pp. 227-235.

[11] P. Dourish, *Where the Action Is: The Foundations of Embodied Interaction*, The MIT Press, 2001.

[12] W.W. Gaver, What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception, *Ecological Psychology*, vol. 5, 1993, p. 1.

[13] P. Feyerabend, *Against Method*, Verso, 1993.

[14] M. McLuhan and L.H. Lapham, Understanding Media: The Extensions of Man, The MIT Press, 1994.

[15] A. Noë, Action in Perception, The MIT Press, 2006.

[16] H. Fastl and E. Zwicker, *Psychoacoustics*, Springer, 2007.

[17] A.S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound,* The MIT Press, 1994.

[18] M. Leman, *Embodied music cognition and mediation technology*, MIT Press, 2008.

[19] A. Väljamäe et. al., Auditory Presence, Individualized Head-Related Transfer Functions, and Illusory Ego-Motion in Virtual Environments, in *7th International Conference on Presence - Proceedings of Presence 2004*, Valencia, Spain: 2004.

[20] J. Gossmann, Towards an Auditory Representation of Complexity, *Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display*, Limerick, Ireland, July 6-9, 2005.

[21] D. Rokeby, S. Penny, (Ed.), Transforming Mirrors: Subjectivity and Control in Interactive Media, *Critical Issues in Electronic Media*, SUNY Press (USA), 1995.

[22] P. Schaeffer, Trait des objets musicaux, Seuil, 2002.

[23] K. Stockhausen and R. Maconie, Stockhausen on Music, Marion Boyars Publishers Ltd, 2000.

[24] I. Xenakis, Formalized Music: Thought and Mathematics in Composition, Pendragon Pr, 2001.

Listening Tests and Evaluation of Simulated Sound Fields Using VibeStudio Designer

György Wersényi

Széchenyi István University,
Hungary
wersenyi@sze.hu

Hesham Fouad

VRSonic,
USA
hfouad@vrsonic.com

ABSTRACT

This paper presents the results of a user-based evaluation of localization accuracy, distance perception as well as room size perception for headphone and loudspeaker based auditory displays. A total of 50 participants listened to four auditory scenes created with VRSonic's VibeStation application. Each scene was rendered using two methods: loudspeaker panning over a 5.0 loudspeaker array and headphone-based spatial sound reproduction using Head Related Transfer Functions (HRTFs). The four scenes were designed to each test a specific aspect of spatial hearing. Scene 1 tested for localization of fixed sources. Scene 2 was used to examine room size perception. Scene 3 was used to test distance perception and Scene 4 tested for localization of moving sources and listener. The participants responded to questions related to the location of each sound they heard as well as transitions between two room sizes and free field. The results of the current study show that the system setup including hardware and software performs as expected and offers a user-friendly way for virtual audio simulation.

1. INTRODUCTION

Virtual auditory displays deal with simulating real world audio experiences [1-3]. Perceiving an auditory event in the real world entails integrating information about both the event itself and its location with respect to the listener. The ability to perceive the spatial location of a virtual sound source entails recreating monaural and binaural cues, and spectral modifications to the acoustic signal reaching a listener [4]. This can be done either through headphone-based spatial sound reproduction using Head Related Transfer Functions (HRTFs) or through multi-loudspeaker panning techniques [5-7].

Head-related Transfer Functions (HRTFs) describe how an auditory event is heard at the human's eardrum [8]. HRTF measurement is an intrusive and time-consuming process and entails playing sounds from designated locations, while recording the sounds using tiny microphones placed inside listener's ears. Individualized recordings of HRTFs are thought to substantially enhance the human's ability to judge sound locations especially when using headphone-based spatial sound reproduction [4]. Due to the complexity of HRTF recording for individual subjects, different catalogues that store HRTF recordings for multiple subjects have been developed; these include AUDIS [9], CIPIC [10], and LISTEN [11].

This paper provides the results of user-based evaluation of sound fields simulated using headphone-based spatial sound reproduction and loudspeaker panning techniques. The objective of the study was to compare subjects' localization accuracy, distance perception, and space perception with sound fields simulated using both these approaches as well as to test the capability of the software environment.

2. BACKGROUND

As aforementioned, spatial audio technology simulates cues that are naturally present and enable listeners to locate sounds in the real world. More specifically, humans perceive sound location in three dimensions; azimuth, elevation, and distance.

Interaural time and intensity differences (ITD and IID) are used for localizing a sound source's angular position (azimuth). Interaural cues are based on the relative difference between wave fronts at the two ears on the horizontal plane [5, 12]. IID and ITD, do not however provide sufficient information for a listener to disambiguate between source positions in the frontal hemisphere and corresponding positions in the rear hemisphere. This is because IID and ITD values are identical for a given position in one hemisphere and its reflected position in the other ("cone of confusion").

The human pinnae provide spectral modifications to the acoustic signals that aid in both disambiguating front and back sources as well as elevation judgment with respect to the median plane [13]. The spectral modifications resulting from pinnae folds produce a unique set of micro-time delays, resonances, and diffractions that translate into a unique descriptor for each sound source position in the median plane [4]. These spectral modifications are particularly important for modeling the HRTF of a listener.

The intensity of a sound source is the most prominent distance cue in anechoic environments (or with familiar sounds) [14, 15]. The intensity of a sound is inversely proportional to the squared distance from the sound source. In reverberant environments the ratio of reflected to direct sound plays an important role for distance perception [5], this ratio creates perceptual differences in the sound quality that depend on source distance [15].

HRTF-based spatial audio reproduction deals with modeling the acoustic signal modifications resulting from a listener's head, torso, and pinnea reflections. HRTF measurements entail placing tiny microphones inside the listener's ear canal. Then sounds are played from an array of loudspeakers precisely placed at known locations around the listener [16]. When the sounds are played, examining the spectral difference between the known played sound and the sound recorded by the microphones enables the extraction of the modifications that are unique to the listener. These spectral modifications are then stored and can be used to play sounds to a listener. It is important to note that the proper choice of HRTF is crucial to truly simulate sound source positions. For example using a non-good sound localizer HRTF can worsen that of a naturally good sound localizer [17].

HRTF-based sound reproduction is best if individualized HRTFs are used. In one study it was found that localization accuracy using headphones (and individualized HRTFs) resulted in comparable performance to free field listening (i.e., localization blur of about 5-10 degrees), nevertheless the rate of front-back confusions increased from 6% to 11% and elevation judgments became less defined [18]. Using non-individualized HRTFs, ITD and IIDs are synthesized but some spectral information is distorted, which leads to ambiguous elevation judgments and increased front-back errors [19].

The other approach used for sound field simulation is the use of free field loudspeaker arrays with either amplitude panning or wave field synthesis approaches. Loudspeakers strategically placed around a listener can be used to simulate the angular location of a sound source by manipulating the signals being played over loudspeakers. Panning approaches simply scale the amplitude of a sound signal presented over two (2D arrays) or three (3D arrays) loudspeakers to give the impression of a positional source. Most surround sound implementations utilize this approach. The other approach, wave field synthesis, attempts to recreate the incident wave front of a source at the listener using a large number of loudspeakers arranged in a line-array configuration. This approach, while producing good results, requires a very large number of loudspeakers to be effective.

3. METHOD

3.1. Participants

A total of 50 subjects participated in the listening tests, 13 females and 37 males. The minimum age was 18; the maximum age was 50 with a mean value of 29.5 years. Table 1 shows how frequently subjects use headphones.

Daily	Several times a week	Several times a month	Seldom, never	
9	15	18	8	Number of subjects

Table 1. User headphone usage routine.

3.2. Apparatus

The experimental setup consisted of a usual desktop computer equipped with a Creative Audigy sound card and an external TerraTec Aueron 5.1 MK II USB sound card providing 6-channel analog outputs. The loudspeaker display consisted of 5 loudspeakers positioned in a typical surround sound configuration: front-left, center, front-right, surround left and surround right similarly to Fig.2. The JMLAB CC700 was used for center speaker and four Chorus 707 speakers for the rest. All five are 2-way bass-reflex systems with a frequency response of about 60 Hz to 22 kHz. The analog outputs of the TerraTec sound card were connected to the external inputs of a DENON AVR-3805 home theater receiver. The listening room was a nearly empty, large rectangle room with an average reverberation time of 0,8 sec. Subjects were instructed to keep their head still during the listening tests.

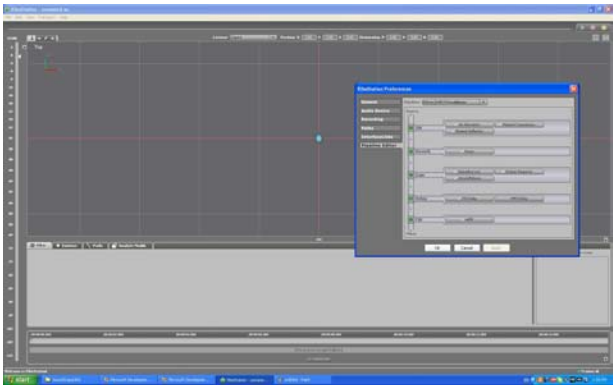


Figure 1. VibeStation application with audio pipeline editor displayed.

Headphone playback was done over a pair of AudioTechnica ATH-D40fs circumaural headphones connected directly to the computer's audio card.

The simulated sound fields were created using VRSonic's VibeStudio Designer software suite [20]. VibeStudio Designer consists of the VibeStation application for spatial audio scene design and the Profiler application for HRTF selection based on a best-fit selection method [12]. VibeStation is capable of rendering scenes over 2, 4, 5 and 7 loudspeakers and over headphones using binaural synthesis with HRTFs. Larger loudspeaker arrays (up to 48 loudspeakers) can also be supported with the addition of a SoundSim Rack external rendering appliance that interfaces with the VibeStudio applicaton. The software allows users to configure the audio rendering pipeline by including and excluding processing stages in the audio pipeline and by selecting rendering algorithms for loudspeaker panning (Fig. 1).

The Profiler application guides the user through a selection process that results in a stored listener profile. Listener profiles specify the user's interaural distance, head tracker offsets and HRTF dataset selection. By default the program provides 7 HRTF datasets from the CIPIC and LISTEN catalogues. These include CIPIC subjects 3, 8, 9, 10, and 11; LISTEN subject 3 and a generalized HRTF dataset. The full CIPIC and LISTEN catalogues can be downloaded resulting in 97 HRTF datasets that can be selected.

Rendered scenes can be recorded for later playback and editing as either a single, multichannel audio file or multiple, single channel audio files. The stereo single file format is well suited for playback over headphones or stereo loudspeaker setups without running the software. Multichannel playback, however, can be realized only while running the software with the appropriate loudspeaker setup.

3.3. Experimental Design

For the listening tests we created four scenes using the VibeStation application. The scenes were rendered over both headphones and loudspeakers at approximately the same loudness. Each participant was presented with both playback methods and the results were compared. For the 5.0 loudspeaker display we selected the Vector Based Amplitude Panning (VBAP) loudspeaker-panning algorithm. For the headphone display we selected the CIPIC "subject 3" HRTF dataset.

Scene 1 used the sound of a ringing telephone. Source locations were positioned 45 degrees around the virtual listener's head (Fig. 2). The playback order was randomized in 6 seconds intervals. The task was to identify the source locations.

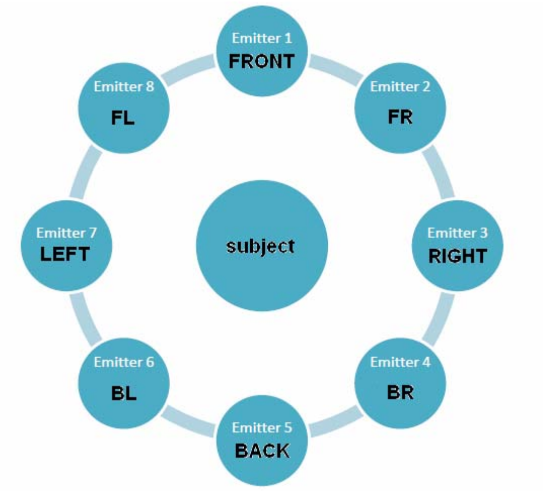


Figure 2. Sound source locations for scene 1. FL, FRONT, FR, BL and BR are also actual loudspeaker positions.

Scene 2 used looped music as the virtual listener moves in the sound field from the free field into a smaller room, then again into the free field and finally into a larger room. The task was to detect the transitions and to estimate room size (which one is small and big). The smaller room was set to 15 x 4.5 x 2 meters whilst the bigger one was set to 20 x 20 x 10 meters, but all other parameters were the same (perfect reflectors material).

Scene 3 used the sound of a honk of a car in front of the listener. The distance first was simulated 40 meters (100%) then it was decreased to 20 meters (50%) and again to 10 meters (25%). The task was to detect that the distance was decreased to the half every time. Finally, we asked the subjects to make a raw estimate in meters.

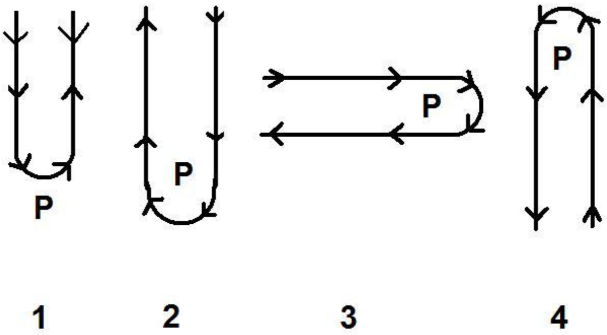


Figure 3. Set of possible trajectories. P indicates the listener's position.

Scene 4 included a trajectory of a flying object. For 5.0 loudspeaker playback we used the sound of a helicopter, for headphone playback we used the sound of an airplane. The task was to select the proper trajectory from a set of four different possibilities as shown in Fig 3.

3.4. Experiment Procedure

Prior to the start of an evaluation session, each participant completed an informed consent and a demographics questionnaire. A detailed explanation of the measurement process was given. Each subject listened first to scene 1 using randomized presentation order of sound sources. This was followed by scenes 2 to 4. After each scene questions were answered referring to that scene. The measurement was about 30 minutes. Measurements with the loudspeaker setup were executed in the university laboratory at a later time. The same 50 subjects participated in this test.

3.5. Evaluation

3.5.1. Headphone Playback

Scene 1

Table 2 summarizes the results of subjects' localization accuracy with the headphone rendering of Scene 1. The diagonal indicates correct answers. There are no left-right reversals but front-back reversals are frequent. Front-back reversals are one of the main problems in virtual and sometimes in real life localization [21, 22] This is also present on the sides where, for example, front-left is confused with back-left. Subjects often described back sources as frontal sources with lower loudness level.

%	Front	FR	Right	RB	Back	BL	Left	FL
Front	**80**	8			43			4
FR	12	**52**	21	20	2			
Right		34	**69**	19				
RB		6	10	**61**	2			
Back	4				**53**			
BL						**61**	14	18
Left						16	**66**	24
FL	4					23	20	**54**

Table 2. Results of Scene 1 with headphone playback. Compare with Table 3.

Scene 2

The recognition of spatial properties was nearly perfect, only 3 subjects failed. Both the transitions as well the room size estimation were easy tasks for the subjects. Only 6 people thought that the first room would be bigger. These decisions were based on the simulated reverberations. Because both rooms were highly reflective environments (metal-like), the differences between transitions were easy to detect. Setting different room sizes or materials to create smaller differences in reflections could result in larger errors.

Scene 3

The first drop (from 40 to 20 meters) in the distance was detected correctly by 75% of the participants, while the second drop (from 20 to 10 meters) was detected correctly by only 62% of the participants. We expected that the estimation of the distance in meters would result in a wide range of numbers. The task was to estimate the middle source position that is simulated at a distance of 20 meters. About 30% could give a relatively good estimation of the distance, 50% estimated the distance as being further than it was (50-100 m) and 20% estimated the distance to be closer than 5 meters. This result is expected as distance perception depends on a variety of cues including familiarity with a sound, the ratio of direct to reverberant energy reaching the listener, and spectral changes to the source. In this scenario the only cue present for detecting distance was spreading loss.

Scene 4

The best performing simulated trajectory was number 2 (Fig. 3), 82% detected it correctly. Subjects were allowed to listen to the sound three times. The mean value for the number of auditions was however only two. We observed that people who seldom or never use headphones needed three auditions. In case of incorrect localization, subjects usually guessed trajectory 3.

In general, younger people (20-27 years of age) and frequent headphone users were better almost in every task. Only in front-back confusions are results independent from gender, age or headphone user routine.

For test with personalized HRTF we had only 10 subjects. Personalization means setting the head diameter and physical properties for a better interaural time difference simulation using the Profiler application. The HRTF used in these conditions was the same as before (subject 3 of the CIPIC database). Seven subjects had the same results with and without personalization. One had worse and two had better results with personalization (decreased rate of front-back confusion). These results are only informal due to the small number of participants.

3.5.2. Loudspeaker playback

Scene 1

%	Front	FR	Right	RB	Back	BL	Left	FL
Front	**86**	8						14
FR	9	**78**	25					
Right		14	**66**	14				
RB			9	**74**	20			
Back				12	**66**	15		
BL					14	**76**	21	
Left						9	**65**	11
FL	5						14	**75**

Table 3. Results of Scene 1 with loudspeaker playback. Compare with Table 2.

Results were overall better for loudspeaker playback compared to headphone listening. It was very helpful that the physical positions of the loudspeakers were identical to the simulated virtual directions in five cases, and sound was only transmitted from the actual loudspeaker (Fig. 2.). In the three cases where the virtual source did not coincide with a loudspeaker position, the virtual sound source was created by two loudspeakers (LEFT, RIGHT, BACK). Front-back confusion disappeared, a symmetrical diagonal can be seen in Table 3. Correct judgments are around 65-86%. In case of localization errors subjects mentioned one of the closest virtual positions. This fact is reflected by the diagonal of Table 3. (e.g if the simulated source was FR, incorrect answers included only FRONT and/or RIGHT).

Scene 2

Surprisingly, in Scene 2 the results for loudspeaker playback were the same as headphone playback: only 3 subjects failed to detect the transitions, and only 5 subjects failed to detect the correct room size. We have to take into account that the listening room play a significant role and different listening rooms could result is different results.

Scene 3

The first drop (from 40 to 20 meters) in the distance was detected correctly by 76%, the second (from 20 to 10 meters) only by 65%. About 18% could give a relatively good estimation. 54% of the rest estimated it too far (50-100 m) and 28% estimated it closer than 5 meters. This is almost the same as by headphone playback.

Scene 4

Subjects performed best with simulated trajectory number 3 (Fig.2.). 74% of the subjects detected it correctly, this is slightly worse than the headphone playback condition. It was helpful that sounds from behind come actually from real loudspeakers behind the listener. Subjects could listen to the sound three times and the mean value for the number of auditions was again two.

In general, younger people (20-25 years of age) are better almost in every task. It is important, how the loudspeakers are positioned and what kind of virtual source directions will be simulated. Front-back confusion is not present, mainly due to the center loudspeaker. It seems to be a good idea to use a center speaker. Listeners suggested that room size detection was easier via headphones, maybe due to the listening room properties during loudspeaker playback.

4. DISCUSSION AND CONCLUSIONS

50 subjects participated in a listening test using headphone playback and loudspeaker setup. Headphone playback included non-individual HRTF synthesis while loudspeaker setup used a 5.0 installation. For both tests four different scenes were rendered to test localization, front-back reversals, distance estimation and room models using VRSonic's VibeStudio Designer. The software environment allows easy access to parameters and controlling the simulation. Results of the listening tests are comparable to former results in the literature.

5. FUTURE WORK

Some considerations about the program and future planning:

- There is no built-in wave editor in VibeStation. Using VibeStation and a wave editor in parallel can sometimes be blocked by the ASIO driver. Other drivers may work parallel.
- The "emitter database" is very small, there are only two built-in wave files. This means, one has to download, record and edit the wave files.
- Adding measured, individual HRTFs to the HRTF database requires that the measured HRTFs be converted into the program's SAF format. There were no tools provided with the program to do this.
- Rooms are very simple, geometrical forms, there is no CAD option and it is a simple reverberation simulation for the room only.

- The distance model could be extended by some low-pass filtering that simulates air absorption. This function is implemented in the current version of the software.

6. ACKNOWLEDGEMENTS

This paper is based upon work supported in part by the Office of Naval Research (ONR). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views or the endorsement of ONR. We also thank the participants involved and Christian Herrmann at the University of Applied Sciences, Leipzig, Germany for programming and leading the listening tests.

7. REFERENCES

[1] G. Kramer, “An introduction to auditory display”. In G. Kramer (Ed.), *Auditory Display* (pp. 1-77). Reading, MA: Addison-Wesley, 1994.

[2] B. Gygi, V. Shafiro, “From signal to substance and back: insights from environmental sound research to auditory display design”. Proc. of ICAD’09. pp. 240-251, 2009.

[3] M. Kleiner, B. I. Dalenbäck, P. Svensson, “Auralization – an overview.” *J. Audio Eng. Soc. vol. 41*, pp. 861-875, 1993.

[4] D.R. Begault, *3D Sound for Virtual Reality and Multimedia*. Academic Press, Inc., Cambridge, MA, 1994.

[5] J. Blauret, *Spatial hearing: The psychophysics of human sound localization*. Translated by J.S. Allen. MIT Press, Cambridge, Mass, 1983.

[6] W. Ahnert, S. Feistel, T. Lentz, C. Moldrzyk, S. Weinzierl, “Head-Tracked Auralization of Acoustical Simulation”. Preprint 6275, 117th AES Convention San Francisco, 2004.

[7] S. Ferguson, D. Cabrera, “Vertical Localization of Sound from Multiway Loudspeakers”. *Journal of the AES, vol. 53(3)*, pp. 163-173, 2005.

[8] E.M. Wenzel, M. Arruda, D.J. Kistler, & F.L. Wightman, “Localization using Non-individualized Head-related Transfer Functions”, *Journal of the Acoustical Society of America, vol. 94*, 1993, pp. 111-123.

[9] https://www.european-acoustics.org/Products/Documenta/Publications/09-de2

[10] http://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm

[11] http://recherche.ircam.fr/equipes/salles/listen/

[12] B.U. Seeber, & H, Fastl, Subjective Selection of Non-Individual Head-Related Transfer Function, Proc. 2003 *International Conference on Auditory Display*, pp. 259-262, Boston University, Boston, MA, July 6-9, 2003.

[13] J. Hebrank, & D. Wright, “Spectral cues used in the localization of sound sources on the median plane”, *Journal of the Acoustical Society of America ,vol.* 56, pp. 1829-1834, 1974.

[14] P. McGregor, A.G. HORN, & M.A. Todd, “Are familiar Sounds ranged more accurately?” *Perceptual and Motor Skills,* vol. 61, 1082, 1985.

[15] J.C. Middlebrooks, & D.M. Green, “ Sound localization by human listeners”, *Annual Review of Psychology, vol.* 42, pp. 135 – 159, 1991.

[16] F.L. Wightman, & D.J. Kistler, “Headphone simulation of free-field listening I: Stimulus synthesis”, *Journal of the Acoustical Society of America, vol.* 85, pp. 858-867, 1989.

[17] E.M. Wenzel, “Localization in Virtual Acoustic Displays, *Presence, vol.* 1, pp. 80-107, 1992.

[18] F.L. Wightman, & D.J. Kistler, “Headphone simulation of free-field listening I: Psychophysical validation”, *Journal of the Acoustical Society of America, vol.* 85, pp. 868-878, 1989.

[19] E.M. Wenzel, M. Arruda, D.J. Kistler, & F.L. Whitman WIGHTMAN, “Localization using Non-individualized Head-related Transfer Functions”, *Journal of the Acoustical Society of America, vol.* 94, pp. 111-123, 1993.

[20] http://www.vrsonic.com/products/vibestudiodesigner.html

[21] D. R. Begault, E. Wenzel, M. Anderson, “Direct Comparison of the Impact of Head Tracking Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source”. *J. Audio Eng. Soc. 49(10),* pp. 904-917, 2001.

[22] P. A. Hill, P. A. Nelson, O. Kirkeby, “Resolution of front-back confusion in virtual acoustic imaging systems”. *J. Acoust. Soc. Am. 108(6)*, pp. 2901-2910, 2000.

REAL-TIME SOUND SYNTHESIS USING AN INEXPENSIVE WIRELESS GAME CONTROLLER

Sharman Jagadeesan

Brain and Work Research Centre
Finnish Institute of Occupational Health, Finland
sharman.jagadeesan@ttl.fi

Matti Gröhn

Brain and Work Research Centre
Finnish Institute of Occupational Health, Finland
matti.grohn@ttl.fi

ABSTRACT

Due to the development of the sensor tehcnology it is possible to manufacture wireless multi-degrees of freedom controllers at reasonable costs. We have tested one manufactured by a small finnish start-up company[1] in controlling real-time sound synthesis parameters. According to our experience, it is very suitable for it.

1. INTRODUCTION

We have explored how suitable an inexpensive wireless game controller is in controlling real-time sound synthesis parameters. In this paper we first describe the controller (or as manufacturer like to say *gaming console*), then we describe both sound synthesis methods used in our system.

2. BLOBO- MOTION CONTROLLED GAMING CONSOLE

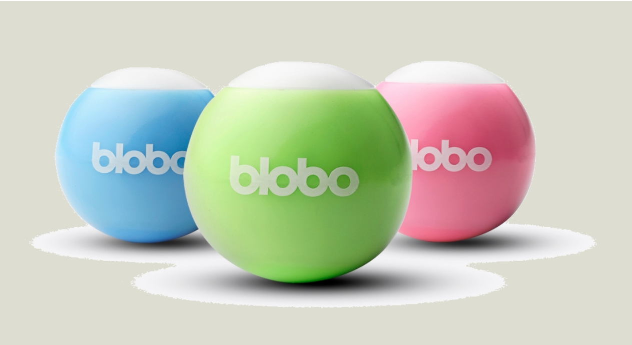


Figure 1: Three Blobo gaming consoles. (Image ©Ball-It)

Blobo[1],[2] is a small sphere (about the size of a golf ball) with built-in multiple sensors and a microcontroller, see figure 1. It has been developed to be used as a game console. The Blobo communicates with a computer via the Bluetooth interface. Accordig to the manufacturer's information the maximum communication speed is 3Mb/s. Internally, the sensors use a 100 MHz frequency. This allows real-time interaction with applications. Magnetometers use 10 bit and accelometers 12 bit accuracy.

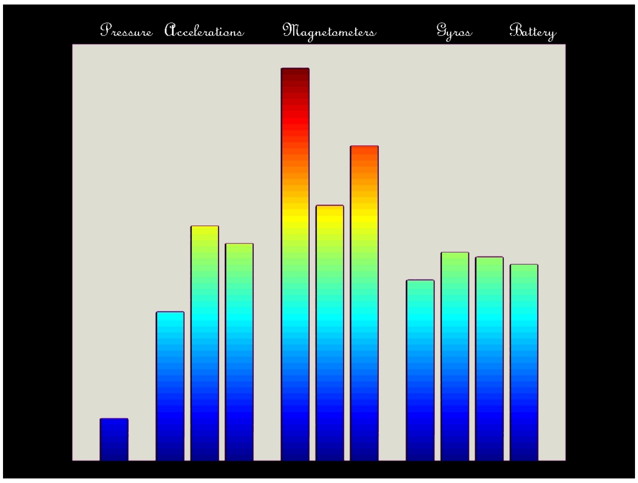


Figure 2: Real-time monitor view of the Blobo -parameters. From the left, Pressure, three accelerometers, three magnetometers, three rotation parameters and the battery status.

The Blobo sends packages to the computer containing various data fields. The data contains control parameters and additional information. The control parameters (see Figure 2) are motion-, rotation-, air pressure- and magnetic field- related. In addition, there is a step counter and a calorie meter. The additional information consists of data such as battery status, id, name and application related data

3. SOUND SYNTHESIS

3.1. Sine wave oscillator

This application is a simple multiple sine wave generator. The Blobo is used for controlling these simultaneous audible sine waves. The number of audible sine waves is determined by the state of the Blobo. Initially only one sine wave is present. It is controlled by one of the three axes. An additional sine wave can be introduced by squeezing the Blobo once. This sine wave will be controlled by an another available axis. A third sine wave can be introduced in the same manner. This wave will be controlled by the third axis. Additionally the Blobo creates a sound when a mode is changed. A state diagram of the application is shown in figure 3.

[1] Ball-It, http://www.ball-it.com

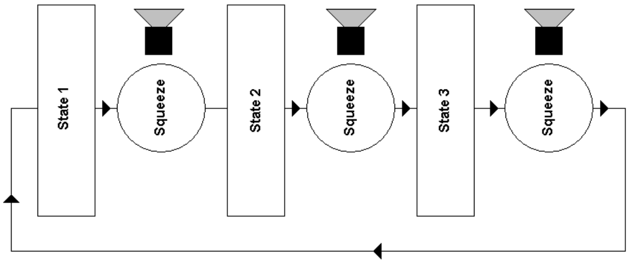


Figure 3: The state diagram of the sine wave controller.

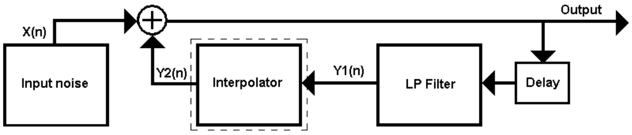


Figure 4: Extended Karplus-Strong block diagram. The interpolator is marked with a dashed rectangle to point out that it is included only in the extended model.

The orientation of the Blobo determines the frequencies of the sine waves. They change either linearly or logarithmically from F_{low} to F_{high} as a function of the controlling axis. The user can define the boundary frequencies as well as the method by which the frequency responds to the orientation.

3.2. Guitar string synthesis controller

The second application implements the famous Karplus-Strong string synthesis [3],[4]. The Karplus-Strong algorithm is relatively easy to implement and is also computationally inexpensive. The block diagram is presented in figure 4.

The basic idea is to generate input noise which separates into the output directly and into the delay line. The delay line comprises of a delay block, a low-pass loop filter and, in the extended model, an interpolator. The delay of the loop branch determines the fundamental frequency of the string vibration and the loop filter determines the decay of the harmonics. This model here, without the interpolator, implements the original Karplus-Strong algorithm [3]. The drawback with this model is that the delay line length is restricted to whole number multiples of the sampling period. To achieve what is called exact tuning, a fractional delay filter must be added to the delay line. This filter, in the simplest form, is a linear interpolation filter. All-pass filters may be used for the same purpose. Another alternative is the Lagrange interpolator.

Here are the signals at different locations of the signal chain.

$$\begin{aligned} x(n) &= rand(length(delay)) \\ y_1(n) &= a_0 x(n) + a_1 x(n-1) \\ y_2(n) &= c y_1(n) + (1-c) y_1(n-1) \end{aligned}$$

where $x(n)$ is a signal from the input noise generator, $y_1(n)$ is a signal after the low-pass filter, the $y_2(n)$ is a signal after the interpolator and a_0 and a_1 are low-pass filter coefficients . Values of the constant c can vary between 0 and 1. The synthesis model can be set to output frequencies either linearly or logarithmically from F_{low} to F_{high} as a function of the rotation angle.

In our application the Blobo is used for controlling the excitation and the fundamental frequency of the synthesis. By squeezing the Blobo a string pluck is emulated and by rotating the Blobo around one of the pre-defined axes controls the fundamental frequency of the synthesis. In this application the loop filter is a two point FIR filter. The fractional delay is implemented using either a linear interpolator or the Lagrange interpolator, chosen by the use.

The synthesis model can be set to output frequencies either linearly or logarithmically, from F_{low} to F_{high} as a function of the rotation angle.

4. CONCLUSIONS

The Blobo has worked well in our experiments as an inexpensive wireless real-time controller. According to our experience, it is hard to use all three rotation axes in controlling the sine wave oscillator. Other more sophisticated sound synthesis methods might be more suitable to be used with this kind of device. It is possible to use simultaneously multiple Blobos. This allows development of new sound synthesis controlling methods or even new instruments.

5. ACKNOWLEDGMENT

We thank Andreas Henelius and Jussi Virkkala for providing valuable help.

6. REFERENCES

[1] http://www.bloboshop.com.

[2] http://www.engadget.com/2009/11/20/ finland-unleashes-blobo-the-squeezable-all-too-cheerful-game-c/.

[3] K. Karplus and A. Strong, “Digital synthesis of plucked string and drum timbres,” *Computer Music Journal (MIT Press)*, vol. 7, no. 2, pp. 43–55, 1983.

[4] D. A. Jaffe and J. O. Smith, “Extensions of the Karplus-Strong plucked string algorithm,” *Computer Music Journal (MIT Press)*, vol. 7, no. 2, pp. 56–69, 1983.

REAL-TIME SOUND SYNTHESIS USING AN INEXPENSIVE WIRELESS GAME CONTROLLER

Sharman Jagadeesan

Brain and Work Research Centre
Finnish Institute of Occupational Health, Finland
sharman.jagadeesan@ttl.fi

Matti Gröhn

Brain and Work Research Centre
Finnish Institute of Occupational Health, Finland
matti.grohn@ttl.fi

ABSTRACT

Due to the development of the sensor tehcnology it is possible to manufacture wireless multi-degrees of freedom controllers at reasonable costs. We have tested one manufactured by a small finnish start-up company[1] in controlling real-time sound synthesis parameters. According to our experience, it is very suitable for it.

1. INTRODUCTION

We have explored how suitable an inexpensive wireless game controller is in controlling real-time sound synthesis parameters. In this paper we first describe the controller (or as manufacturer like to say *gaming console*), then we describe both sound synthesis methods used in our system.

2. BLOBO- MOTION CONTROLLED GAMING CONSOLE

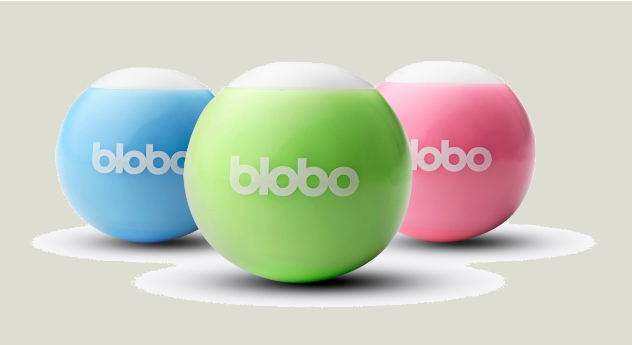


Figure 1: Three Blobo gaming consoles. (Image ©Ball-It)

Blobo[1],[2] is a small sphere (about the size of a golf ball) with built-in multiple sensors and a microcontroller, see figure 1. It has been developed to be used as a game console. The Blobo communicates with a computer via the Bluetooth interface. Accordig to the manufacturer's information the maximum communication speed is 3Mb/s. Internally, the sensors use a 100 MHz frequency. This allows real-time interaction with applications. Magnetometers use 10 bit and accelometers 12 bit accuracy.

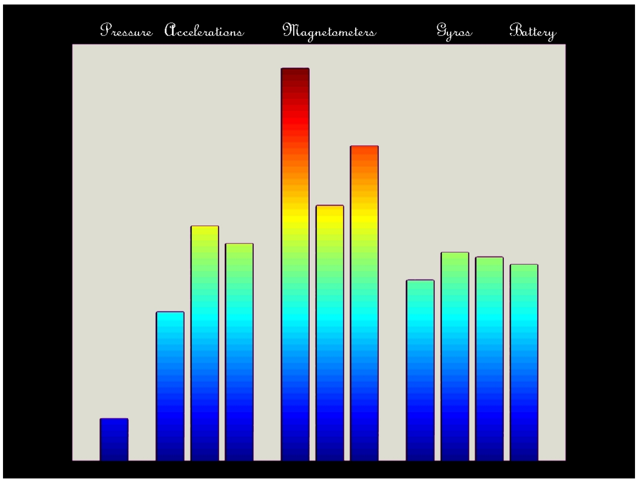


Figure 2: Real-time monitor view of the Blobo -parameters. From the left, Pressure, three accelerometers, three magnetometers, three rotation parameters and the battery status.

The Blobo sends packages to the computer containing various data fields. The data contains control parameters and additional information. The control parameters (see Figure 2) are motion-, rotation-, air pressure- and magnetic field- related. In addition, there is a step counter and a calorie meter. The additional information consists of data such as battery status, id, name and application related data

3. SOUND SYNTHESIS

3.1. Sine wave oscillator

This application is a simple multiple sine wave generator. The Blobo is used for controlling these simultaneous audible sine waves. The number of audible sine waves is determined by the state of the Blobo. Initially only one sine wave is present. It is controlled by one of the three axes. An additional sine wave can be introduced by squeezing the Blobo once. This sine wave will be controlled by an another available axis. A third sine wave can be introduced in the same manner. This wave will be controlled by the third axis. Additionally the Blobo creates a sound when a mode is changed. A state diagram of the application is shown in figure 3.

[1] Ball-It, http://www.ball-it.com

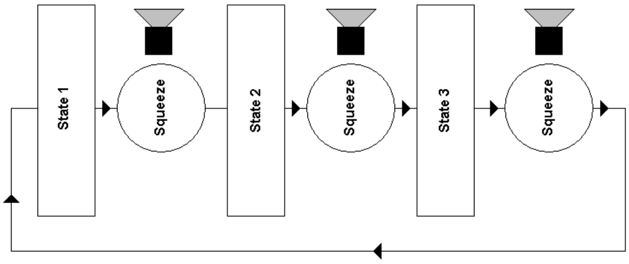


Figure 3: The state diagram of the sine wave controller.

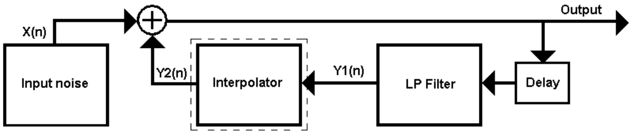


Figure 4: Extended Karplus-Strong block diagram. The interpolator is marked with a dashed rectangle to point out that it is included only in the extended model.

The orientation of the Blobo determines the frequencies of the sine waves. They change either linearly or logarithmically from F_{low} to F_{high} as a function of the controlling axis. The user can define the boundary frequencies as well as the method by which the frequency responds to the orientation.

3.2. Guitar string synthesis controller

The second application implements the famous Karplus-Strong string synthesis [3],[4]. The Karplus-Strong algorithm is relatively easy to implement and is also computationally inexpensive. The block diagram is presented in figure 4.

The basic idea is to generate input noise which separates into the output directly and into the delay line. The delay line comprises of a delay block, a low-pass loop filter and, in the extended model, an interpolator. The delay of the loop branch determines the fundamental frequency of the string vibration and the loop filter determines the decay of the harmonics. This model here, without the interpolator, implements the original Karplus-Strong algorithm [3]. The drawback with this model is that the delay line length is restricted to whole number multiples of the sampling period. To achieve what is called exact tuning, a fractional delay filter must be added to the delay line. This filter, in the simplest form, is a linear interpolation filter. All-pass filters may be used for the same purpose. Another alternative is the Lagrange interpolator.

Here are the signals at different locations of the signal chain.

$$\begin{aligned} x(n) &= rand(length(delay)) \\ y_1(n) &= a_0x(n) + a_1x(n-1) \\ y_2(n) &= cy_1(n) + (1-c)y_1(n-1) \end{aligned}$$

where $x(n)$ is a signal from the input noise generator, $y_1(n)$ is a signal after the low-pass filter, the $y_2(n)$ is a signal after the interpolator and a_0 and a_1 are low-pass filter coefficients . Values of the constant c can vary between 0 and 1. The synthesis model can be set to output frequencies either linearly or logarithmically from F_{low} to F_{high} as a function of the rotation angle.

In our application the Blobo is used for controlling the excitation and the fundamental frequency of the synthesis. By squeezing the Blobo a string pluck is emulated and by rotating the Blobo around one of the pre-defined axes controls the fundamental frequency of the synthesis. In this application the loop filter is a two point FIR filter. The fractional delay is implemented using either a linear interpolator or the Lagrange interpolator, chosen by the use.

The synthesis model can be set to output frequencies either linearly or logarithmically, from F_{low} to F_{high} as a function of the rotation angle.

4. CONCLUSIONS

The Blobo has worked well in our experiments as an inexpensive wireless real-time controller. According to our experience, it is hard to use all three rotation axes in controlling the sine wave oscillator. Other more sophisticated sound synthesis methods might be more suitable to be used with this kind of device. It is possible to use simultaneously multiple Blobos. This allows development of new sound synthesis controlling methods or even new instruments.

5. ACKNOWLEDGMENT

We thank Andreas Henelius and Jussi Virkkala for providing valuable help.

6. REFERENCES

[1] http://www.bloboshop.com.

[2] http://www.engadget.com/2009/11/20/finland-unleashes-blobo-the-squeezable-all-too-cheerful-game-c/.

[3] K. Karplus and A. Strong, "Digital synthesis of plucked string and drum timbres," *Computer Music Journal (MIT Press)*, vol. 7, no. 2, pp. 43–55, 1983.

[4] D. A. Jaffe and J. O. Smith, "Extensions of the Karplus-Strong plucked string algorithm," *Computer Music Journal (MIT Press)*, vol. 7, no. 2, pp. 56–69, 1983.

AN OVERVIEW OF ALGORITHMIC MUSIC COMPOSITION IN THE NOTEWORKS SOFTWARE PLATFORM

Robert Alexander, John Umbaugh, Patrick Turley

Noteworks Music, USA
www.noteworks.net
noteworksmusic@gmail.com

ABSTRACT

Noteworks is music composition software that re-imagines the way music is created, played, and shared. Users create musical compositions by building networks and interacting with them in real time. Noteworks reduces the learning curve for algorithmic-music composition, such that most individuals with a basic knowledge of computer interaction can create original compositions with limited instruction. Dynamic networks have the potential to play back for hours without repeating. This document will provide a brief summary overview of the GUI.

1. INTRODUCTION

Noteworks allows the user to write pieces of music that evolve over time, and features such as stochastic nodes enable the user to think in terms of probabilities and tendencies. Rather than creating one composition at a time, a single network can potentially generate thousands of unique compositions. There may be several ways in which a chord structure could evolve, and theoretically a listener could set a network in motion and allow it to unfold for days without hearing every possible combination. At any point while the composition is playing back, the user can begin writing the MIDI data to a file. They could then import this data into Finale or Sibelius and print out a piece of sheet music, or into a production platform such as Logic Studio.

2. JAVA IMPLEMENTATION

Noteworks is written in Java, and as such can be used on a variety of platforms - including Windows, Mac, and Linux. Noteworks relies heavily on the Java Swing component library for the graphical display, and on the Java MIDI library to render its sound.

The technology behind the sequencing portion of Noteworks is inspired by time-dynamical recursive neural network models, in which individual network nodes can be assigned some sort of musical expression (e.g. a MIDI message, tempo, or some sort of modification), and the arrows represent temporal relationships between nodes.

3. NOTEWORKS GUI OVERVIEW

This section will step through the various elements of the Noteworks interface, explaining each in detail. All musical composition within Noteworks takes place in a centralized canvas, and is facilitated through a set of tools. A video tutorial covering similar material is available at the following URL:

http://robertalexandermusic.com/Noteworks_Demonstration.mov

3.1. Interface Overview

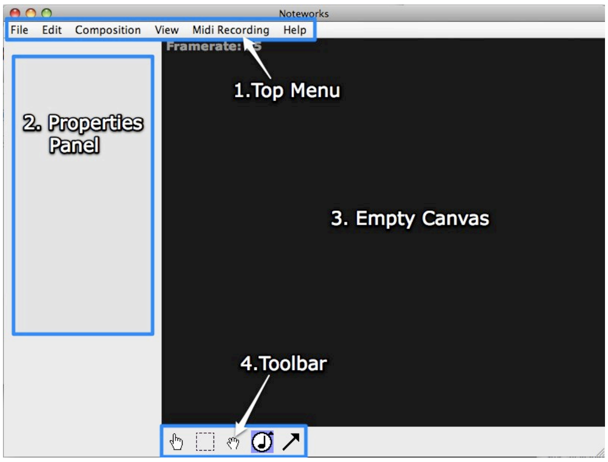


Figure 1: The Noteworks GUI as presented upon initially loading the software.

1) Top Menu: Here the user can save their composition, open a saved composition, edit the instruments used for playback, stop playback completely, export a song as a MIDI file, and select from a number of additional advanced options
2) Properties Panel: Allows the user to edit the settings of a selected node.
3) Empty Canvas: This space is where compositions are created.
4) Toolbar: These tools are used for playback, node selection, moving nodes, creating nodes, and drawing arrows.

3.2. Toolbar Elements

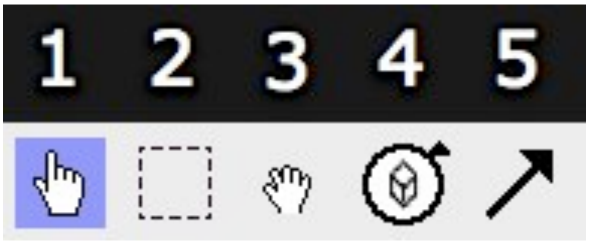


Figure 2: Closer inspection of the Noteworks tool set.

1) Node Firing tool: With this tool selected the cursor turns into a pointing hand. Clicking on a MIDI or Rest node starts the network in motion.
2) Selection Tool: This tool allows for selecting either a single node or multiple nodes. When a single node is selected its attributes will appear in the properties panel. Click and drag around multiple nodes, then use the hand grabber tool to move groups of nodes.
3) Grabber Tool: While in this mode, click and drag on the canvas to adjust your view of the composition. You can also click and drag on nodes to adjust their placement. At any time you can zoom in and out by using the scroll wheel on a two-button mouse or two fingers on a track pad.
4) Node Creation Tool: Click and hold briefly to reveal a sub-menu. You can select from 4 different node types: MIDI (standard instrument), Rest, Chance, and Echo. Click on a node type to select it, and then click anywhere on the canvas to add. See the section on node types for more information.
5) Arrow Drawing Tool – With this tool selected, the user can click and drag from one node to another to create a connection.

3.3. Node Types

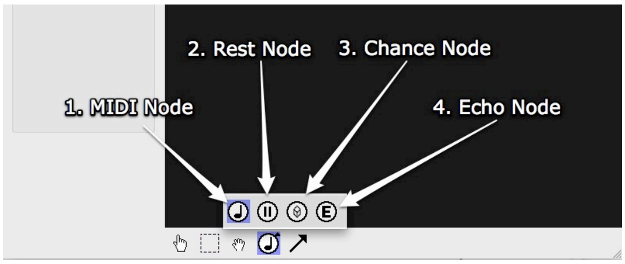


Figure 3: Example of placing a figure with experimental results.

1) MIDI stands for Musical Instrument Digital Interface. These nodes are at the core of every Noteworks composition. Within the MIDI properties panel the user can adjust the note pitch, duration and volume. The instrument can be changed by switching the node to a different channel, or by loading up a new instrument in the edit menu. MIDI nodes will pass an impulse to all outgoing connections.
2) Rest nodes can be used to delay the transfer of an impulse from one node to another. Rest nodes will pass on an impulse to all outgoing connections.
3) A chance node will fire to one of its outlets, which is picked at random.
4) Echo nodes will repeat whatever note is received, with the option of transposing this note to a new pitch and/or adjusting the delay. Echo nodes can be strung together in sequence to create complete melodic phrases. Data can be passed between echo nodes and chance nodes to create stochastic behavior.

4. FUTURE DIRECTION

The addition of sub-networks would enable users to consolidate large sections of a composition into a single, more-manageable object. Future node types would also enable compositions to unfold in a more dynamic manner. Sequential nodes would switch between all outgoing connections in an ordered fashion; this would enable the user to create recurring poly-rhythmic patterns. Interactive nodes would allow input from various external interfaces, such as a keyboard or MIDI controller. Audio file playback could be facilitated through an additional node type, or by sending data to an external sampler or VST instrument.

The object-oriented interface would readily lend itself to touch-screen interaction, future versions of the software could be implemented on the iPhone and iPad platforms. Several interaction modes could include: free composition, level based game play, and social compositions constructed by multiple users in tandem. Early research also indicates strong potential for Noteworks in the K-12 educational market.

AN OVERVIEW OF ALGORITHMIC MUSIC COMPOSITION IN THE NOTEWORKS SOFTWARE PLATFORM

Robert Alexander, John Umbaugh, Patrick Turley

Noteworks Music, USA
www.noteworks.net
noteworksmusic@gmail.com

ABSTRACT

Noteworks is music composition software that re-imagines the way music is created, played, and shared. Users create musical compositions by building networks and interacting with them in real time. Noteworks reduces the learning curve for algorithmic-music composition, such that most individuals with a basic knowledge of computer interaction can create original compositions with limited instruction. Dynamic networks have the potential to play back for hours without repeating. This document will provide a brief summary overview of the GUI.

1. INTRODUCTION

Noteworks allows the user to write pieces of music that evolve over time, and features such as stochastic nodes enable the user to think in terms of probabilities and tendencies. Rather than creating one composition at a time, a single network can potentially generate thousands of unique compositions. There may be several ways in which a chord structure could evolve, and theoretically a listener could set a network in motion and allow it to unfold for days without hearing every possible combination. At any point while the composition is playing back, the user can begin writing the MIDI data to a file. They could then import this data into Finale or Sibelius and print out a piece of sheet music, or into a production platform such as Logic Studio.

2. JAVA IMPLEMENTATION

Noteworks is written in Java, and as such can be used on a variety of platforms - including Windows, Mac, and Linux. Noteworks relies heavily on the Java Swing component library for the graphical display, and on the Java MIDI library to render its sound.

The technology behind the sequencing portion of Noteworks is inspired by time-dynamical recursive neural network models, in which individual network nodes can be assigned some sort of musical expression (e.g. a MIDI message, tempo, or some sort of modification), and the arrows represent temporal relationships between nodes.

3. NOTEWORKS GUI OVERVIEW

This section will step through the various elements of the Noteworks interface, explaining each in detail. All musical composition within Noteworks takes place in a centralized canvas, and is facilitated through a set of tools. A video tutorial covering similar material is available at the following URL:

http://robertalexandermusic.com/Noteworks_Demonstration.mov

3.1. Interface Overview

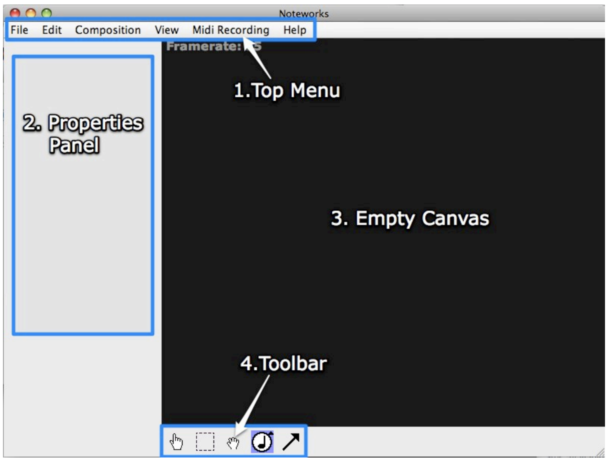


Figure 1: The Noteworks GUI as presented upon initially loading the software.

1) Top Menu: Here the user can save their composition, open a saved composition, edit the instruments used for playback, stop playback completely, export a song as a MIDI file, and select from a number of additional advanced options
2) Properties Panel: Allows the user to edit the settings of a selected node.
3) Empty Canvas: This space is where compositions are created.
4) Toolbar: These tools are used for playback, node selection, moving nodes, creating nodes, and drawing arrows.

3.2. Toolbar Elements

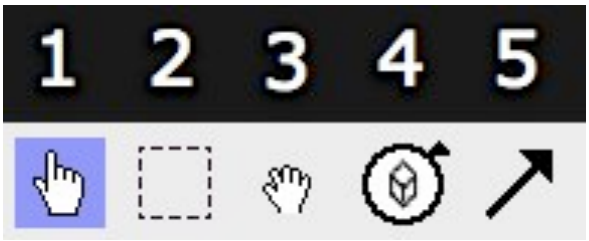


Figure 2: Closer inspection of the Noteworks tool set.

1) Node Firing tool: With this tool selected the cursor turns into a pointing hand. Clicking on a MIDI or Rest node starts the network in motion.
2) Selection Tool: This tool allows for selecting either a single node or multiple nodes. When a single node is selected its attributes will appear in the properties panel. Click and drag around multiple nodes, then use the hand grabber tool to move groups of nodes.
3) Grabber Tool: While in this mode, click and drag on the canvas to adjust your view of the composition. You can also click and drag on nodes to adjust their placement. At any time you can zoom in and out by using the scroll wheel on a two-button mouse or two fingers on a track pad.
4) Node Creation Tool: Click and hold briefly to reveal a sub-menu. You can select from 4 different node types: MIDI (standard instrument), Rest, Chance, and Echo. Click on a node type to select it, and then click anywhere on the canvas to add. See the section on node types for more information.
5) Arrow Drawing Tool – With this tool selected, the user can click and drag from one node to another to create a connection.

3.3. Node Types

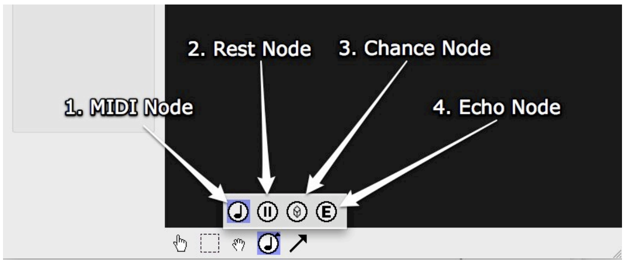


Figure 3: Example of placing a figure with experimental results.

1) MIDI stands for Musical Instrument Digital Interface. These nodes are at the core of every Noteworks composition. Within the MIDI properties panel the user can adjust the note pitch, duration and volume. The instrument can be changed by switching the node to a different channel, or by loading up a new instrument in the edit menu. MIDI nodes will pass an impulse to all outgoing connections.
2) Rest nodes can be used to delay the transfer of an impulse from one node to another. Rest nodes will pass on an impulse to all outgoing connections.
3) A chance node will fire to one of its outlets, which is picked at random.
4) Echo nodes will repeat whatever note is received, with the option of transposing this note to a new pitch and/or adjusting the delay. Echo nodes can be strung together in sequence to create complete melodic phrases. Data can be passed between echo nodes and chance nodes to create stochastic behavior.

4. FUTURE DIRECTION

The addition of sub-networks would enable users to consolidate large sections of a composition into a single, more-manageable object. Future node types would also enable compositions to unfold in a more dynamic manner. Sequential nodes would switch between all outgoing connections in an ordered fashion; this would enable the user to create recurring poly-rhythmic patterns. Interactive nodes would allow input from various external interfaces, such as a keyboard or MIDI controller. Audio file playback could be facilitated through an additional node type, or by sending data to an external sampler or VST instrument.

The object-oriented interface would readily lend itself to touch-screen interaction, future versions of the software could be implemented on the iPhone and iPad platforms. Several interaction modes could include: free composition, level based game play, and social compositions constructed by multiple users in tandem. Early research also indicates strong potential for Noteworks in the K-12 educational market.

Audio-Visual Panoramas and Spherical Audio Analysis using the Audio Camera

Adam E. O’Donovan
Perceptual Interfaces and Reality Laboratory
UMIACS, University of Maryland
adam.o@visisonics.com

Ramani Duraiswami
Perceptual Interfaces and Reality Laboratory
UMIACS, University of Maryland
ramani@umiacs.umd.edu

Abstract

Capturing a scene for later or contemporaneous display needs to capture the complex interactions between the source(s) in the scene and the environment. High order spherical Ambisonics and plane-wave analysis are powerful mathematical tools for such scene analysis. The spherical microphone array (and its embodiment in the Audio Camera) is a useful tool for capture and analysis of scenes. Further information about the environment is available from the visual scene.

We present the audio-visual panoramic camera as a tool that greatly simplifies the task of processing audio visual information by providing one common framework for both modalities. Via the Audio Camera [1], we show that microphone arrays can be viewed as a central projection camera that can effectively image the audible acoustic frequency spectrum. We demonstrate a new device, the audio visual Panoramic camera that is composed of a 64 channel spherical microphone array combined with a 5 element video camera array. The combined sensor is capable of real-time audio visual panoramic image generation using state of the art NVidia Graphics cards. It also provides an order-7 ambisonic description of the scene.

1. Introduction

Nearly every biological creature senses the world with both eyes and ears. This is due to the tremendous amount of complementary information in each of these modalities. The visual system conveys pinpoint geometric information about objects in our environment. The acoustic environment conveys information such as speech and does not suffer as badly as vision from issues such as occlusions. For these reasons and many others it is attractive to investigate utilizing both modalities in problems of scene understanding. Microphone arrays have been an attractive tool for audio processing as they provide geometric information about acoustic sources in an environment as well as the ability to spatially suppress noise. However, it is often difficult to calibrate and utilize both microphone arrays and video cameras to perform multi-modal scene understanding. We take the approach that both microphone arrays and video cameras are central projection devices [1] and therefore can be treated in a common imaging framework. This allows the creation of a pre-calibrated multimodal panoramic sensor, The Panoramic Audio Visual Camera, which significantly simplifies the fusion of both audio and visual information.

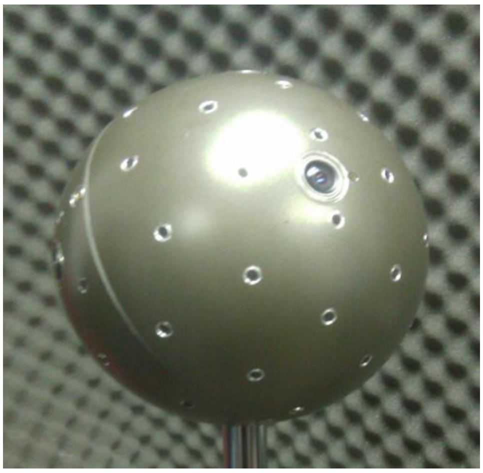


Figure 1: The Audio Visual Panoramic Camera.

2. The Spherical Microphone Array

To generate the acoustic images in the audio visual panoramic camera we utilize the spherical microphone array. There are several benefits to the spherical geometry [2]. The first is that it provides equal spatial resolution in all directions. Additionally, several mathematical simplifications presented in [3] provide efficient algorithms for processing the acoustic information in parallel on commercial graphics cards thus providing real-time capability. The array consists of 64 microphones distributed over the surface of an 8 inch sphere. The microphone signals are then amplified via individual pre-amp circuits and sent to an array of analog to digital converters. The digitized acoustic data is sampled at 44.1 kHz per channel and collected by an FPGA where it is interleaved into a single USB 2.0 Stream. This provides the interface to the PC where the data can be immediately shipped to the graphics processor for real-time processing. Figure 1 shows the Audio Visual Panoramic Camera.

3. Auditory Scene Capture and Playback

The auditory scene can be decomposed into its spherical harmonic components up to order 7. Further, the scene can

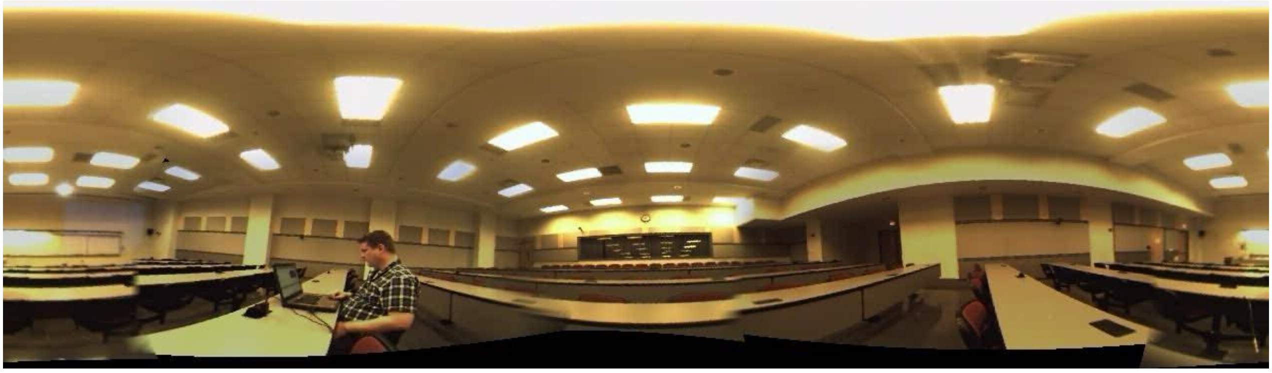


Figure 2: Example of the panoramic video image acquired by our device.

also be decomposed into its filtered plane-wave components [4]. These can be used as inputs to creating ambisonic displays using spherical arrays or mixed with HRTFs as discussed in [5] and recreate the auditory scene over headphones.

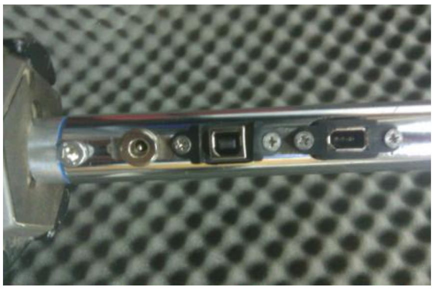


Figure 3. External ports of the Audio Visual Camera.

4. Panorama Stitching

Due to the fact that the spherical microphone array provides an omni-directional acoustic image of the environment it is highly beneficial to have an omni-directional image of the visual environment as well. In order to generate a panoramic image of the scene we utilize a 5 camera array. The placement of the cameras was selected to avoid all microphones in the spherical microphone array via a spatial optimization. Additionally, the placement was selected such that all directions except those present around the mounting handle are seen by at least one camera. Each of the 5 video cameras are 752x480 color Firewire cameras. The frame acquisition is triggered by the internal audio FPGA to provide synchronization of both the audio and video components of the device. The 5 camera image streams are collected via an internal Firewire that allows an interface to the PC consisting of a single Firewire cable. Figure 2 shows an example of the stitch achieved using our 5 camera panoramic video camera.

5. The Panoramic Audio Visual Camera

Given both the omni-directional acoustic and video images we perform a one time joint audio visual calibration to bring both modalities into a single global coordinate system [1]. Because both the audio and visual cameras are collocated and share a common center of projection we can perform acoustic image transfer onto the video stream as described in [3] to provide a final image that represents both acoustic and visual information present in the environment. The extent of the external cable connections of the device consist of a single USB 2.0 port and a single Firewire port as well as external power. Figure 3 shows the interface ports present at the base of the handle.

6. Conclusion

We present a multimodal panoramic audio-visual camera. We demonstrate that we can present both acoustic and visual panoramic video streams in real-time. By combining both the spherical microphone array and an omni-directional camera array we provide a simple means of sensing the world of light and sound using a single common framework. Many applications of the device are possible, including in auditory display.

Acknowledgement: Partial ONR support is gratefully acknowledged.

7. References

[1] A. O’Donovan, R. Duraiswami, J. Neumann, Microphone arrays as Generalized Cameras for Integrated Audio Visual Processing. IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007

[2] J.Meyer and G. Elko, “A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield,” IEEE ICASSP 2002, 2:1781-1784

[3] A. O'Donovan, R.Duraiswami, N. Gumerov, “Real Time Capture of Audio Images and Their Use with Video” Proceedings 2007 IEEE WASPAA.

[4] .D.N. Zotkin, R. Duraiswami, N.A. Gumerov. Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays. IEEE Transactions on Audio, Speech & Language Processing, 18:2-18, 2010.

[5] R. Duraiswami, D.N. Zotkin, Z. Li, E. Grassi, N.A. Gumerov, L. Davis, High Order Spatial Audio Capture and its Binaural Head-Tracked Playback over Headphones with HRTF Cues. 119th AES Convention, 2005

Audio-Visual Panoramas and Spherical Audio Analysis using the Audio Camera

Adam E. O'Donovan
Perceptual Interfaces and Reality Laboratory
UMIACS, University of Maryland
adam.o@visisonics.com

Ramani Duraiswami
Perceptual Interfaces and Reality Laboratory
UMIACS, University of Maryland
ramani@umiacs.umd.edu

Abstract

Capturing a scene for later or contemporaneous display needs to capture the complex interactions between the source(s) in the scene and the environment. High order spherical Ambisonics and plane-wave analysis are powerful mathematical tools for such scene analysis. The spherical microphone array (and its embodiment in the Audio Camera) is a useful tool for capture and analysis of scenes. Further information about the environment is available from the visual scene.

We present the audio-visual panoramic camera as a tool that greatly simplifies the task of processing audio visual information by providing one common framework for both modalities. Via the Audio Camera [1], we show that microphone arrays can be viewed as a central projection camera that can effectively image the audible acoustic frequency spectrum. We demonstrate a new device, the audio visual Panoramic camera that is composed of a 64 channel spherical microphone array combined with a 5 element video camera array. The combined sensor is capable of real-time audio visual panoramic image generation using state of the art NVidia Graphics cards. It also provides an order-7 ambisonic description of the scene.

1. Introduction

Nearly every biological creature senses the world with both eyes and ears. This is due to the tremendous amount of complementary information in each of these modalities. The visual system conveys pinpoint geometric information about objects in our environment. The acoustic environment conveys information such as speech and does not suffer as badly as vision from issues such as occlusions. For these reasons and many others it is attractive to investigate utilizing both modalities in problems of scene understanding. Microphone arrays have been an attractive tool for audio processing as they provide geometric information about acoustic sources in an environment as well as the ability to spatially suppress noise. However, it is often difficult to calibrate and utilize both microphone arrays and video cameras to perform multi-modal scene understanding. We take the approach that both microphone arrays and video cameras are central projection devices [1] and therefore can be treated in a common imaging framework. This allows the creation of a pre-calibrated multimodal panoramic sensor, The Panoramic Audio Visual Camera, which significantly simplifies the fusion of both audio and visual information.

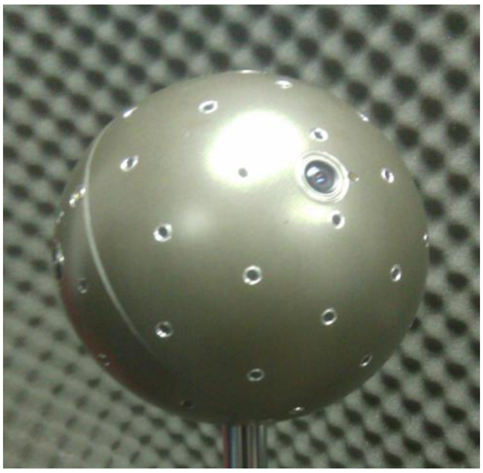


Figure 1: The Audio Visual Panoramic Camera.

2. The Spherical Microphone Array

To generate the acoustic images in the audio visual panoramic camera we utilize the spherical microphone array. There are several benefits to the spherical geometry [2]. The first is that it provides equal spatial resolution in all directions. Additionally, several mathematical simplifications presented in [3] provide efficient algorithms for processing the acoustic information in parallel on commercial graphics cards thus providing real-time capability. The array consists of 64 microphones distributed over the surface of an 8 inch sphere. The microphone signals are then amplified via individual pre-amp circuits and sent to an array of analog to digital converters. The digitized acoustic data is sampled at 44.1 kHz per channel and collected by an FPGA where it is interleaved into a single USB 2.0 Stream. This provides the interface to the PC where the data can be immediately shipped to the graphics processor for real-time processing. Figure 1 shows the Audio Visual Panoramic Camera.

3. Auditory Scene Capture and Playback

The auditory scene can be decomposed into its spherical harmonic components up to order 7. Further, the scene can

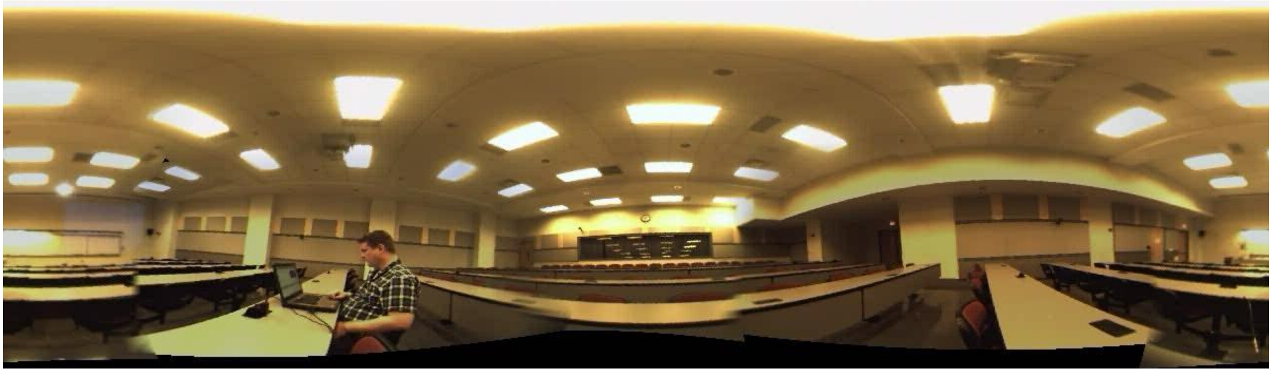


Figure 2: Example of the panoramic video image acquired by our device.

also be decomposed into its filtered plane-wave components [4]. These can be used as inputs to creating ambisonic displays using spherical arrays or mixed with HRTFs as discussed in [5] and recreate the auditory scene over headphones.

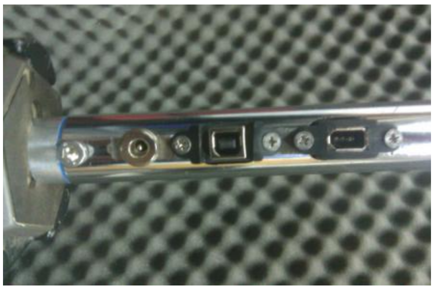


Figure 3. External ports of the Audio Visual Camera.

4. Panorama Stitching

Due to the fact that the spherical microphone array provides an omni-directional acoustic image of the environment it is highly beneficial to have an omni-directional image of the visual environment as well. In order to generate a panoramic image of the scene we utilize a 5 camera array. The placement of the cameras was selected to avoid all microphones in the spherical microphone array via a spatial optimization. Additionally, the placement was selected such that all directions except those present around the mounting handle are seen by at least one camera. Each of the 5 video cameras are 752x480 color Firewire cameras. The frame acquisition is triggered by the internal audio FPGA to provide synchronization of both the audio and video components of the device. The 5 camera image streams are collected via an internal Firewire that allows an interface to the PC consisting of a single Firewire cable. Figure 2 shows an example of the stitch achieved using our 5 camera panoramic video camera.

5. The Panoramic Audio Visual Camera

Given both the omni-directional acoustic and video images we perform a one time joint audio visual calibration to bring both modalities into a single global coordinate system [1]. Because both the audio and visual cameras are collocated and share a common center of projection we can perform acoustic image transfer onto the video stream as described in [3] to provide a final image that represents both acoustic and visual information present in the environment. The extent of the external cable connections of the device consist of a single USB 2.0 port and a single Firewire port as well as external power. Figure 3 shows the interface ports present at the base of the handle.

6. Conclusion

We present a multimodal panoramic audio-visual camera. We demonstrate that we can present both acoustic and visual panoramic video streams in real-time. By combining both the spherical microphone array and an omni-directional camera array we provide a simple means of sensing the world of light and sound using a single common framework. Many applications of the device are possible, including in auditory display.

Acknowledgement: Partial ONR support is gratefully acknowledged.

7. References

[1] A. O’Donovan, R. Duraiswami, J. Neumann, Microphone arrays as Generalized Cameras for Integrated Audio Visual Processing. IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007

[2] J.Meyer and G. Elko, “A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield,” IEEE ICASSP 2002, 2:1781-1784

[3] A. O'Donovan, R.Duraiswami, N. Gumerov, “Real Time Capture of Audio Images and Their Use with Video” Proceedings 2007 IEEE WASPAA.

[4] .D.N. Zotkin, R. Duraiswami, N.A. Gumerov. Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays. IEEE Transactions on Audio, Speech & Language Processing, 18:2-18, 2010.

[5] R. Duraiswami, D.N. Zotkin, Z. Li, E. Grassi, N.A. Gumerov, L. Davis, High Order Spatial Audio Capture and its Binaural Head-Tracked Playback over Headphones with HRTF Cues. 119th AES Convention, 2005

SIGNAL PAINTING

Cooper Baker

University of California San Diego,
Department of Music
9500 Gilman Dr., La Jolla, CA 92093
cb@ucsd.edu

ABSTRACT

Composition submission for the 16th International Conference on Auditory Display.

1. DESCRIPTION

Signal Painting is a time-based work that visualizes relationships between harmonics of synthesized tones. The modulating tones are designed to have specific harmonic content, and they appear to create unfolding interwoven visual patterns when their audio spectrum is seen as a moving color spectrogram.

2. TECHNIQUES

Signal Painting was realized with Cycling ’74’s Max/Msp/Jitter multimedia software. Tone generation and modulation were performed using classical synthesis techniques (figure 1), and a large filterbank of noise was used to sonically print the pixels of a bitmap (figure 2). The resulting audio signal was captured and used to create a moving color spectrogram. The piece was created live, in real time, and recorded directly to a video file.

3. MEDIA FORMAT

The work was rendered as a color 640 x 480 (16:9) stereophonic quicktime .mov file which can be delivered on DVD or sent electronically in any preferred format.

4. PREVIEW URL

http://musicgrad.ucsd.edu/~cbaker/files/baker.m4v

5. STILL IMAGES

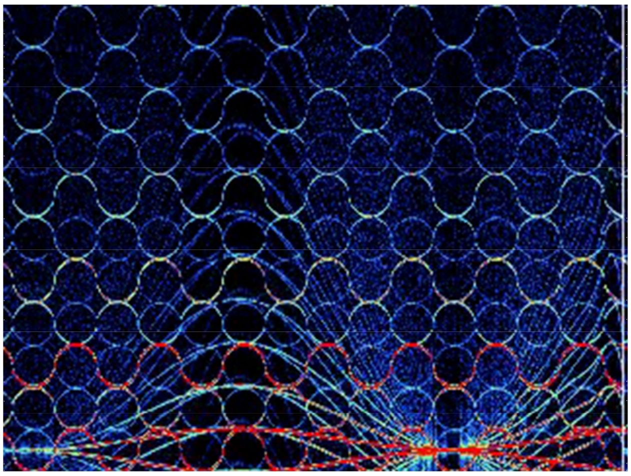


Figure 1: still image from Signal Painting.

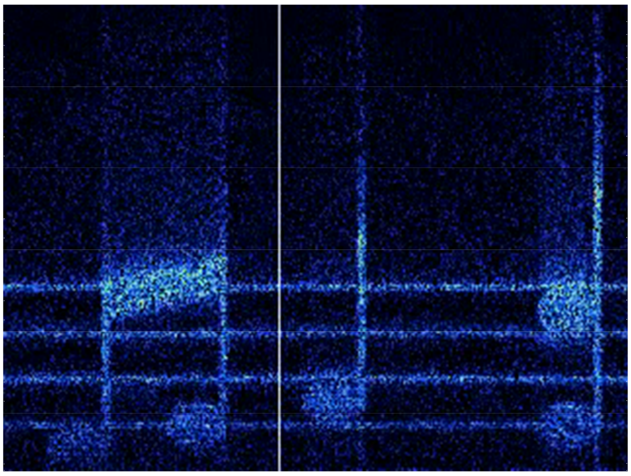


Figure 2: Still image from Signal Painting.

ON *DISPLACEMENTS 1B*: TIME, TYCHISM, SPACE, AND EMBODIMENT

Douglas Boyce

The George Washington University
Department of Music
801 22nd St NW Phillips B-144
Washington DC 20052
dboyce@gwu.edu

ABSTRACT

This paper situates the musical work *Displacements 1b* in musical and philosophical framings operative in those discourses and relevant to the thematics of the conference. Particular attention is paid to C.S. Peirce's concept of tychism and the perception of musical time, and on the impact of spatialization of live, acoustic instruments on meaning in musical works. This document is submitted as bridge between the technical and aesthetic aspects of the conference.

1. INTRODUCTION

> "We necessarily express ourselves by means of words and we usually think in terms of space. That is to say, language requires us to establish between our ideas the same sharp and precise distinctions, the same discontinuity, as between material objects. This assimilation of thought to things is useful in practical life and necessary in most of the sciences. But it may be asked whether the insurmountable difficulties presented by philosophical problems do not arise from placing side by side in space phenomena which do not occupy space, and whether, but merely getting rid of the clumsy symbols round which we are fighting, we might not bring the fight to an end."
>
> - Henri Bergson
> *Time and Free Will*, authors preface

> "Music is based on temporal succession and requires alertness of memory. Consequently music is a chronologic art, as painting is a spatial art."
>
> – Stravinsky
> *Poetics of Music*

*Displacement*s 1b is the first in a series of works, which will use spatialization to interrogate the role of embodiment in musical performance. The term embodiment refers to the manner in which the performer's physical self is the source of the sonic component of music, the spatial and temporal location of musical action and decision-making, and (in many contexts and for many people), the figure or emblem of music itself. The work involves the projection (or diffusion) of the live sound of the performers through an array of speakers, and the movement of performers on the stage or in the performance space. These devices are used to literally alienate the performance from the performer, and in doing so call attention to these basic elements of the grammar of performance.

The work also uses rhythmic devices such as open pulse textures and quasi-improvisatory, aleatoric notation, techniques which are of long standing interest in my music. A dynamic rhythmic character emerges from the juxtaposition of strongly pulse driven metered material with freer, 'open-pulse' material. My notation gives performers choice in the timing of passages, the alignment of events, and the order of phrases. Performers are thus involved in musical decision making generally more associated with composition than performance. Aesthetically and technically, I have found these approaches to generate rich and expressive textures, and result in exciting performance environments in which performers are not merely agents of compositional will, but are equal partners is musicking. I will suggest later that the term 'tychism,' though somewhat antique, is an appropriate term for this dynamic approach, superior to other terms such as aleatory or improvisation.

2. SPATIALIZATION

In *Displacement*s *1b* music (or perhaps musicking itself) is spatialized along two different axes, with the performers themselves moving and secondary sources (or 'ghosts') of the audio signals of the instruments also moving through an emitter array and audio spatialization software. These traces or ghost sources perform alongside the physically present ensemble while at the same moment setting themselves apart, in that they exist without a concrete association with a perceived physical source. This is achieved in real-time via VRSonic's Vibestation™, with an iPhone application triggering the movement of sources along predefined splines.

I have described this ghosting as an alienation of sound and performer. I use the term 'alienation' here carefully and positively, in that is it only through such an alienation from our assumptions about seemingly common place actions like musical performance that we can remind our selves of the basic elements of the grammar of performance which have been effaced or downplayed. The mediation of technology enables this physical dislocation. My goal in the work is to decouple and recouple the sound of the performers with the performers' location, and by doing so to remind the listener of the ontological group effort that is musicking. This dislocation can and perhaps should be thought of as a kind of disembodiment.

The term embodiment refers to the manner in which the performer's physical self is the source of the sonic component of music, the spatial and temporal location of musical action and decision-making, and (in many contexts and for many people), the figure or emblem of music itself. In the tradition of Western art music, there is a tendency to minimize the body, be it through standardized concert attire, the suppression of dance responses, or the penetration of acousmatic projection and the mechanical, electronic and digital sound files into quotidian life. We can lose sight of the body, which is present at the start of the musical chain. Ironically, though technology that spatializes and disembodies music is omnipresent (and which disembodies our experience of music every day), finding technology that can do this with the subtlety and nuance expected of music is rather hard to do.

3. TEMPORAL ORGANIZATION

In this work as in many of my pieces, the performers are afforded a degree of freedom greater than is typical in a classical score. There are extended passages with no shared pulse stream and using Lutoslawskian frame notation and other 'open score' notations. This approach resonates with the above consideration of the performer as a nexus of performative behaviors. My motivation in using these techniques is in part to produce musical interesting textures and events but even more so to produce a kind of music making in which performers and composers are partners, rather than agents bound to a composer's will.

In Western art music, certain kinds of practice are emphasized, in particular a hierarchical notion of accuracy, in which a primary pulse is, literally, the measure of all other temporal relationships in a particular piece. The Stravinsky quote at the opening of this writing foregrounds the power of temporality in music but doesn't interrogate how this chronologic rule is brought into being by music and (especially) by performers. For Stravinsky, the score is a machine, you do what you're told and something good will come of it, especially if you're playing something by Stravinsky.

My music could be described as 'indeterminate,' a rather inelegant term of our art, a negative definition which has the positive attribute of calling attention to the fact that that not all elements of a given work (i.e.: the tradition 'composerly' features like notes, and the rhythmic disposition of the notes) are as fixed as one generally finds in concert music. Some of these features will be left open to the decisions of musicians in the moment of performance. My notation gives performers choice in the timing of passages, the alignment of events, and the order of phrases. Performers are thus involved in musical decision making generally more associated with composition than performance. Some passages in the work exhibit traditional metric and temporal organization, referred to in the score as 'shared pulse.' In open pulse textures, pulse is not shared between the musicians; individuals organize the sequence of events through cuing. The effect is a free flowing, un-metered but rhythmic texture.

In the work, these open textures are at odds with the necessity of a rigid clock time for the live processing. Each ghost source is on an independent clock, thus giving us 4 distinct temporary strata, each managed by human action.

A variety of terms already in use come close to capturing this temporal ebb and flow, but all have drawbacks. "Indeterminate" is a bit inelegant and imprecise; much is still determined by the composer. It has implications of abandonment, rather than mutuality. Similarly, aleatory, used most often to describe the works of Cage, is problematic. The term became known to European composers through lectures by acoustician Werner Meyer-Eppler at Darmstadt International Summer Courses for New Music in the beginning of the 1950s. According to his definition, "a process is said to be aleatoric ... if its course is determined in general but depends on chance in detail" (Meyer-Eppler 1957, 55). This emphasizes a lack of human agency, when the impact of the technics is actually exactly the opposite, a surfeit of agency. 'Open' would seem a more attractive term, though its use in literary theory would emphasize the impact of these structures on the 'meaning' of the work, which is not really the focus of the techniques. Ludism has picked up associations with 'game pieces' and so perhaps misses the extent to which cases. 'Improvisation', though it captures the character of real time decision making, tends to minimized the channeling of decision-making through very precise and detailed techniques of notation is at this point stylistically or genre bound

I find the term 'tychism' as a somewhat antique but highly appropriate word to describe this aspect of musical performance. Tychism is a concept developed by C.S. Peirce to describe the emergence of order from chance events.

> "In an article published in The Monist for January, 1891, I endeavored to show what ideas ought to form the warp of a system of philosophy, and particularly emphasized that of absolute chance. In the number of April, 1892, I argued further in favor of that way of thinking, which it will be convenient to christen tychism (from {tyché}, chance)."
>
> ('The Law of Mind', CP 6.102, 1892)

Peirce's friend and colleague William James perhaps articulated his friend's idea more succinctly, calling tychism "Peirce's suggestion [that] order results from chance-coming." My hope is that these techniques will make overt the dynamics of performance which are always in play, but are sometimes masked behind the edifice of precision and accuracy, and remind us that we are, composers, performers and listeners, all partners in the drawing of order out of chaos, chronos out of aion.

4. SOME CONSEQUENCES

For a work of chamber music like *Displacements 1b*, a consideration of this basic grammar of performance must look at what it is to play together. The version of *Displacement*s performed at the conference underscores the relationship of space in this equation through shifting positions of performers on the stage. Performers pair and related to one another musically, but here these traditional counterpoints and accompaniments are amplified by a choreography of positions, sometimes supporting, sometimes undermining the pairings made by the notes on the page and in the air. The work also underscores the relationship of space in this equation through shifting positions of performers on the stage, and the production of secondary sources for the sound of those instruments, in effect producing 'ghost' versions of the instruments. These traces perform with the physical ensemble, while at the same moment are set apart, in that they exist without a concrete association with a perceived physical source.

The explicit problematization of the unity of performance and performer is at the core; patterns in those embodiments support amplify the formal structure the work articulate in the other, more traditionally musical elements of the work.

Tychism, as a philosophical concept forces to the surface issues of identity, meaning, and job-descriptions. Collectivity; a performance, and even a work, is the product of the actions of many individuals. My point in advocating for the usage of this term is not to imply that my work requires a new language for description, but rather that there is a tychastic element to all musical expressions, and that the development of a language to describe them is of use.

Similarly, the de-situating and re-situation of sonic production in the body of the performers through the use of spatializing technology forces the fact of human agency in forgetting that music is a product of human action and agency. We can forget music is fundamentally humanistic in the abstract and interpersonal in practice. Much in our world makes us behave as if and perhaps feel that music objectified, commodified and generic, when at its best, most nuanced, it is distinct, personal and human.

This interest in the relationship between music with identifiable, locatable sources and those without emerges from personal, compositional and pedagogical observations on the manner in which acousmatic projection and the mechanical, electronic and digital sound files have fully penetrated everyday life. This is in many ways a powerfully positive influence on musical culture, but there is an associated risk of losing sight of music as a product of human action and agency. For me music is fundamentally humanistic in the abstract and interpersonal in practice. In tonight's work, technology and artifice are used to call attention to the human performance of music – performers and the music they make are repeatedly alienated and reunited, a metaphor, perhaps the ways in which our quotidian experience of music (especially in the heavily mediated, post-commidification realm of recordings, radio and the internet) moves again and again from generic and transactional to the distinct and the personal.

A STRATEGY FOR COMPOSING MUSIC USING THE SONIFICATION OF "SNAPSHOT"-TYPE DATA COLLECTIONS "SCHNAPPSCHUSS VON DER ERDE"

David Spondike

Firestone High School Campus for the Visual and Performing Arts,
Akron Public Schools,
333 Rampart Dr., Akron, Ohio 44321
dspondike@aol.com

ABSTRACT

Transforming the sonification of data into a musical experience that is satisfying on scientific as well as aesthetic grounds requires balancing similar and competing objectives and sensibilities. One possible solution, described here, is used to create the digital music composition *Schnappschuss von der Erde* in response to the call for compositions by the International Community for Auditory Display (ICAD) Conference, 2006.

1. INTRODUCTION

The International Community for Auditory Display 2006 Conference provided the basic data resources and objectives for realizing a musical composition for a concert entitled *Global Music - The World by Ear*.

The mapping strategy and compositional process used to create *Schnappschuss von der Erde* are described in the body of this paper. While the possible approaches to transforming sonifications into music are numerous, this investigation will focus on representing data in as clear a manner as possible, while satisfying the aesthetic demands of musical form.

2. DATA COLLECTION

Data was downloaded from the public database at the World Bank website [1]. Data from the most recent available year between 2000 and 2004 is used. The original data set selected by the composer includes 187 countries and 23 categories

3. STRATEGIES

Transforming data that represents a snapshot in time into an art form that unfolds in time, such as music, requires strategies different than say, mapping time series data which would seem more adapted to musical interpretation. In order to satisfy aesthetic demands, interesting, artistic patterns must be a result of the investigation. Scientific interests tend to be more concerned with the level of objectivity in the representation. In order to satisfy the later, the sonic representations of the data are created algorithmically and are not subsequently edited, distorted, or augmented in any way, except as mentioned in the performance notes. The resulting sonifications are then selected for their musical value and are used as is and in their entirety to assemble the musical composition in a collage-like fashion.

My daughter was beading one night while I was working and I offered her this analogy: Imagine a black shoe box that makes beads. You throw bead making stuff into the box and shake it up until you hear the bead start to rattle in the box. You open the box. If the bead is pretty, you keep it. If it is ugly, you discard it. Do this dozens of times until you have the beads that you want. Then string the beads together one after another in a pattern, and you have a beautiful necklace.

3.1. Perceptual background

The most basic description of the structure of a musical sound is its pitch, loudness, timbre, and temporal placement. Each of these descriptors are in themselves complex structures that are not necessarily discrete. For the purposes of this composition, only pitch, loudness, timbre, and temporal placement are used.

Musical events, such as notes, tend to organize themselves into perceptual streams [2]. The fusion of discrete sound events into streams and the segregation of one stream from another falls to the task of the auditory system. In general, the perceptual system fuses and segregates on the basis of similarity or dissimilarity along some particular characteristic or set of characteristics that may either reinforce or compete against each other. Thus the fusion of discrete notes into some sort of pattern will be effected by similarity or dissimilarity of pitch, loudness, timbre, or temporal placement.

3.2. Plotting data

In order to make data useful (to render patterns observable), we often convert it into a visual representation accessible to our perceptual system. Scatter plots are used to look for a correlation between data, with one variable plotted on the horizontal axis and and second on the vertical axis of a Cartesian coordinate plane. A third variable may be used to color the dots. The advantage of using this type of graph as a metaphor rather than a bar graph or a pie chart, is that the data retain their individual identities. They are scattered across two-dimensional space. As a composer, the job is to get the dots to get up off the page and move through time, the fourth dimension.

In most cases, scientists hope for a thin, squashed, oval shape that slants from one corner of the graph to the opposite. This shows a strong correlation and is usually an indicator for further research. An amorphous blob of dots scattered across the page shows no correlation and the relationship is generally disregarded as not being able to provide any useful information.

If one maps time (order of events) along the horizontal axis and loudness along the vertical axis, a strong positive correlation would create increasing loudness over time, or a crescendo. If the the vertical axis maps pitch, then a rising (melodic) line is created. A falling line or decrescendo is created by a strong negative correlation. A zero correlation produces randomness, or simply noise. Or maybe it is not so simple.

3.3. Mining for useful information

Rising and falling melodic lines, crescendo and decrescendo are useful in music composition, but they are not sufficiently complex to sustain aesthetic interest. However, what has been generally disregarded as being unable to provide useful information, the zero-correlations, are exactly the place to look for aesthetically useful material. These data sets are not uniform in their distribution. However, their distribution is not necessarily random (i.e. white noise). There may be subgroups that show some degree of correlation. These correlations may be so variable that the overall result is a zero-correlation. While the aesthetic purpose here is to find patterns that mix predictability with unpredictability, for the scientist, hearing patterns that are not traditionally strong correlations may inspire investigations from a previously untried perspective. One might ask, “Why do these subsets of data form a pattern and not those sets? Why does this subset of data sound so similar to some other subset of data?”

4. METHOD

Data was downloaded and assembled into a spreadsheet. The spreadsheet was the converted to a text file to be used by a commercial music composition language, Symbolic Composer 5.0.1 [3]. SCOM is a LiSP [4] based language that processes lists of data into a MIDI [5] file. Quicktime [6] was then used to render the MIDI files into .WAV files.

4.1. Sonification algorithm

An algorithm was written that allows the composer to select data categories and assign them to control the order of presentation (controlling temporal placement), pitch, or loudness in the resulting sonification. The data, always bound to its country, was then coordinated in ordered lists. In this process, any missing data in any of the three selected categories caused the deletion of that country and its data from the final result.

The ordering algorithm ranks orders the input data and controls the the ordering of all data in constructed lists.

The pitch mapping algorithm rank orders the data before it is mapped to the set of 128 keyboard symbols provided by the SCOM language.

Data that controls loudness is scaled and rounded to integers between 20 and 127. This data is not rank ordered. The integers 0-127 are standard controller values for MIDI messages.

Rank ordering of pitch data distributes the pitches more evenly across the tessitura of the piece and allows for more “step” sized melodic movements. This is important because of the prevalence of stepwise motion throughout the musics of the world [8]. Scaling the loudness data without rank ordering allows for the creation of more sudden changes of loudness. This creates a model for accents (suddenly louder notes) and ghost notes (suddenly softer notes). The pattern of accents creates the perception of rhythm.

Longitude (of the capital city) data for each country is bound with the selected data categories, and is used as values for the main volume controller. This distributes the amplitude between eight (8) separate .WAV files, each of which is used to drive eight individual speakers arranged in a circle around the perimeter of the performance space. Using one speaker to represent 0 degrees longitude and assuming 45-degree spacing between adjacent speakers, a line is calculated from each longitudinal data point to its surrounding speakers. Using a unit circle and allowing 1 to represent full volume, the inverse ratio between the two lines controls the volume data for the two speakers, each assigned their own instrument in the SCOM language. The longitude data places the musical note at a specific location along the perimeter of the room.

Reference to the data sets that control the musical variables will be referenced in the following manner throughout this paper - order : pitch : loudness.

Flexibility exists to invert values or retrograde the group. Inverting values exchanges low values with high values symmetrically around the center. Retrograding the group simply reverses the order. These manipulations preserve symmetry and do not distort the relationships between data. It simply allows the data to be heard from a new perspective.

4.2. Making beads

The process begins with selecting the data categories, then selecting which ordered combination of three (3) categories should be used as input to the sonification algorithm. There are a huge number of possible comparisons that can be made between three variables when each variable can have multiple values. Time and patience preclude examining every possibility.

The strategy above suggests that categories be selected by the following criteria:

1. Select categories of scientific interest.
2. Select categories that create useful musical constructs.

Creating useful musical constructs would include musical streams that ascend and get louder, descend and get quieter, descend and get louder, ascend and get quieter. These can be useful patterns, but the greater point is; what aesthetically interesting patterns lie within the data that await to be discovered?

To create the basic musical tools (i.e. crescendo and decrescendo), data categories were selected where one might intuit a strong correlation, positive or negative, since inverting the relationship is an option. With one category mapped to the order control list, the other can be mapped to either pitch or loudness to create the desired result: a strong positive correlation between order, pitch and loudness should produce a pattern that, on average, gets louder and higher in pitch over time, an ascending melodic line with a crescendo. Of course intuition and reality do not always agree. Some of the actual results are described in the analysis of the composition.

To create streams that do not conform to the shape of these basic tools, categories had to be intentionally selected that would intuitively seem to lack strong correlation.

Since creating the sonifications with the algorithm essentially meant typing in three names and executing the code, it was like a bead making machine. Different categories were simply entered into the execute file to control the musical variables and the files were run on the computer. The results were examined. Results that were aesthetically pleasing were kept. Those that were not were discarded.

Fifty (50) sonifications were retained in the final pool. These were examined and reexamined to the point that they were becoming learned. Descriptions were notated with each name. With the sonifications at hand, assembling them into a musical composition could begin. A total of eighteen different combinations of data ended up in the final composition.

5. COMPOSITION

The sonifications tended coalesce into groups; the successful basic tools, short groups with internal patterns, longer more complex groups, often with embedded subgroups forming internal patterns.

In order to facilitate the linking of one sonification with the next, care had to be taken to consider the beginning and ending characteristics of each.

5.1. Form and analysis

The composition is created in three (3) continuous movements. The shape of the form may be visualized as a horizontal hour glass, in that the more free, more random, and more complex material is used to surround a more restricted internal movement created by a select group of ostinati. The internal structure mixes binary and ternary elements. There is a significant amount of interleaving of material between and within movements. The total duration is 9’ 19”.

5.1.1. First Movement

The first movement is subtitled Jazz Licks. It begins with a short introduction using the grand-piano timbre; a ten (10) second pass that plots longitude against longitude to control order and pitch creates an ascending glissando that starts at zero (0) degrees longitude and traverses counterclockwise the circumference of the room. This is to help orient the listener at the beginning of the piece. The notes are accented by military spending data positively correlated with loudness, i.e. higher spending equals louder notes, to create aesthetic and scientific interest. The controller set, computers : fertility : hiv-rate, with the loudness control inverted, creates a strong descending line with an embedded sequence. The descending line indicates a negative correlation between the number of computers in a nation and the fertility rate. The overall low rate of HIV infection throughout the world provides relatively consistently high loudness values when inverted. Longitude : energy : co2-emissions creates a complex line that spins around the room. A strong correlation between CO2 emissions and energy use per capita produce higher notes that are consistently louder that lower notes.

At this point, the timbre changes to acoustic-guitarnylon. Female-literacy : male-literacy : hiv-rate creates a sparse, ascending accented line. The degree of smoothness of the ascent indicates the strength of the positive correlation between male and female literacy rates. The accented notes indicate countries where HIV rates are high. Computers : fertility : hiv-rate create a complementary descending line. Whereas the descending line was strong when the this combination was used in the piano, here it is quieter, but accented by the high HIV rate countries.

The last three sonifications of this movement use a fretless-bass timbre reminiscent of the late Jaco Pastorius. Computers : gdp-per-capita : co2-emissions creates a line that starts low in pitch. There is a bit of a sequence in the beginning, along with some nice jazz-like accents provided by the CO2 emissions data. Again, the rising line shows a positive correlation. The degree of smoothness of the melodic line (the less it bounces pitch-wise) indicates the strength of correlation. Fertility : Female-literacy : co2-emissions creates a complimentary line by maintaining the same accent structure, but now finding a negative correlation between fertility rate and the level of female literacy. The fertility : infant-mortality : gni-percapita sonification creates an extended line with internal structure. Some relatively extreme loudness variations at the end of the sonification helps to create a perceptual cue akin to a musical cadence to close the first movement. This final phrase will return at the end of the third movement as the final phrase of the piece, thus rounding the form. The duration of the first movement is 2’ 05”.

5.1.2. Second Movement

The middle movement (l’Ostinati) is created from four (4) sonifications, each repeated to become an ostinato. The four controller groups are - military-spending : gdp-per-capita : co2-emissions, military-spending : male-literacy : computers, hiv-rate : male-literacy : female-literacy, and the retrograde of military-spending : male-literacy : computers.

The patterns are repeated and interleaved to attempt to satisfy the aesthetic need for repetition and variation. The movement is bisected by timbre; orchestral-harp for the antecedent and vibraphone for the consequent. The pattern is:

aaaaa bbbb aaa cc bb || bb aa ccc bb cc ddd bd aa cc

The harp timbre returns with the ‘a’ parts near the end of the consequent section. No alterations to the sonifications were made to create rhythms or pitch collections. Since the sonifications here are not rising and falling lines, one should not be listening for correlation in this movement, but rather the focus should be the data points that group together into patterns. It may be that some unexplored variable or relationship in the real world is responsible for creating the pattern that we hear in the musical world. The duration of the second movement is 3’ 12”.

5.1.3. Third Movement

The Percussion Finale begins by borrowing the ostinato idea from the second movement. The perceived 4/4 common time is a result of the sonification. The respite from the odd meter is extended by repeating this pattern. The pattern is created by inverting the loudness control of the military-spending : hiv-rate : computers controller set. Xylophone, woodblock, and kalimba timbres are used to explore the same data set with different tone colors. Note that the placement of the few high pitched notes indicate high HIV-rates and that the inversion of the number of computers controls the loudness (fewer computers results in greater loudness).

Throughout the final, freer section of the third movement, one should focus listening more on patterns than on correlation. The listener can draw his or her own conclusions, but the third movement shows that this method of composition can be quite effective for creating a percussion solo. It is similar in sound to some of the music of Iannis Xenakis, known for his exploration of stochastic composition. This section starts with the woodblock. The first of the two sonifications is the retrograde-inversion of fertility : sanitation : child-mortality. To unravel the comparison, the ordering is from high to low fertility. High sanitation rates produce low notes and vice versa. A strong positive correlation between the ordering controller and the pitch controller creates an ascending line. A strong positive correlation between the ordering controller and the loudness controller creates a decrescendo. A strong positive correlation between the pitch controller and the loudness controller means that the lower notes would be louder. The second sonification played be by the woodblock is the retrograde-inversion of co2-emissions : military-spending : gni-per-capita with the loudness controller also inverted. In this case the loudness relationships are inverted compared to the immediately preceding description.

The second instrument in the free section is the synthetic-tom. This timbre evokes the impression of some of Frank Zappa's digital music. It begins with fertility : sanitation : mortality, similar to the first sonification of the woodblock in this section, but without the retrograde-inversion. The ostinato from the beginning of the third movement is then repeated twice. Next, new material is added by computers: gdp-per-capita : gdp, with the loudness controller inverted. This is followed by the second woodblock phrase in the new timbre. The synthetic-tom finishes with male-literacy : hiv-rate : computers, which creates a crescendo of low notes at the end to lead into the final section.

The last section of the finale is scored for Taiko-drum. It begins with computers : gdp-per-captia : co2-emissions with the loudness controller inverted. This creates a battery of loud low notes to announce the beginning of the final section. Gdp : gni-per-capita : gdp-per-capita might imply a rising crescendo, but is quite more interesting to listen to than one might expect. Ordering energy use per capita (pitch) and CO2 emissions (loudness) by longitude with the loudness inverted spins around the room, but it is not a rising decrescendo as one might expect.

The closing section of the Taiko-drum solo uses the patterns from the second movement; military-spending : gdp-per-capita : co2-emissions, hiv-rate : male-literacy : female-literacy, and military-spending : male-literacy : computers. The final phrase begins with new material provided by computers : drinking-water : co2-emissions. This creates a short crescendo of high notes near the end, signaling the coming of the finale sonification of the piece. The end of the final phrase of the last movement is the same sonification as the end of the first movement, thus rounding the form. The duration of the third movement is 4' 02".

6. PERFORMANCE NOTES

The pitch collection is an equal tempered 21 note division of the octave. This allows for note combinations that may or may not lie within the chromatic system familiar to most Western listeners. The rate of presentation is 7.5 notes per second. This allows for stream formation to be influenced by all of the musical variables. It is also a comfortable listening tempo that enables the perceptual formations of faster moving passages of adjacent notes and slower moving rhythms created by notes separated in time, but grouped by pitch proximity or loudness similarity.

The following adjustments were made to the final audio files for performance. The last note was trimmed. Echo was added to the new final note to enhance the ending. For some reason, certain iterations of sublists were rendered slightly louder or slightly softer than others resulting is slight changes in loudness for some sonifications. Since this does not distort the relationships rendered by the sonifications and actually creates some aesthetic value, these anomalies were retained, albeit with occasional attenuation of -3 dB or -6 dB. The second movement was equalized on a 10-band equalizer with the 40 Hz and 80 Hz bands boosted 6 dB and the 160 Hz band boosted 3 dB to enhance the melodic stream that creates the apparent bass line. A stereo mix of the eight sound files was also created.

7. CONCLUSIONS

Though there are many possible approaches to the sonification of data, there are likely even more strategies for using sonification to create music. One can imagine controlling such musical parameters as spectral envelope, harmonicity, pitch weight, duration, ADSR envelope, sonorous structures (chord-like structures), scales, tonal systems, repetition patterns, contour, granular synthesis controllers, etc. For the scientist, it may be possible to search for correlations and patterns between multiple sets of data, each set mapped to a domain of musical information. Yet, even a simple rendering of data to the most basic music controllers, with little need of constructing higher order musical architectures, is sufficient to produce aesthetically satisfying material. Such approaches can produce results that present data in a relatively clear manner, retaining its usefulness possibly even to the point of creating a new manner of investigating the real world. A pattern is a pattern is a pattern. There is no reason we should be less inclined to investigate a pattern that we hear than we are to investigate a pattern we see on the page. At the same time, the material can be presented in an aesthetically interesting way that may bring some pattern to the attention of others. Taken a step further in the artistic direction, such investigations may provide the successful basis for purely musical work without regard to scientific rigors. What remains to be seen is how the interaction of art and science will play out as inspiration and in result.

8. REFERENCES

[1] The World Bank, "WDI Data Query," http://devdata.worldbankorg/data-query/, Queried April, 2006.

[2] A. Bregman, Auditory Stream Analysis: The Perceptual Organization of Sound. A Bradford Book: The MIT Press, Cambridge, MA and London, England, First paperback edition 1994, Copyright 1990.

[3] J. Podrazik, P. Tolonen, et. al., Symbolic Composer 5.0.1, MRAC Publishing Ltd., 1998-2003.

[4] G. Byers, S. Hain, A. Hartley, B. St. Clair, Mac Common Lisp, Digitool Inc., 1995-2003.

[5] Complete MIDI 1.0 Detailed Specification v96.1 (second edition), MIDI Manufacturers Association, 2001.

[6] QuicktimeTM 7.0.4, Apple Computer Inc., 1991-2004.

[7] P. von Hippel, “Redefining Pitch Proximity: Tessitura and Mobility as Constraints on Melodic intervals,” Music Perception 17(3), 2002, pp. 315-327.

ROAD, RIVER AND RAIL: A MULTI-INTERACTIVE COMPOSITION

Jorge García del Valle Méndez

Lauensteiner Str. 14
D-01277 Dresden (Germany)
+ 49 351 8033857
jorgegadelvalle@gmail.com

ABSTRACT

road, river and rail is not only a musical composition in the usual sense. It is also a sort of multi-interaction on different levels between real and virtual life. These interaction-levels comprise, not only a performance-interaction, but also an interactive creation by the composition as well as by the interpretation.

The concepts of *reality* and *virtuality* are here applied to the two sides of the composition: the acoustic and the electronic one.

The source of the work, which comes from the reality (bell sounds of Korean Buddhist temples), will be analyzed through a Fast Fourier Transform (FFT) in order to provide the structural materials as well as the form organization of the composition.

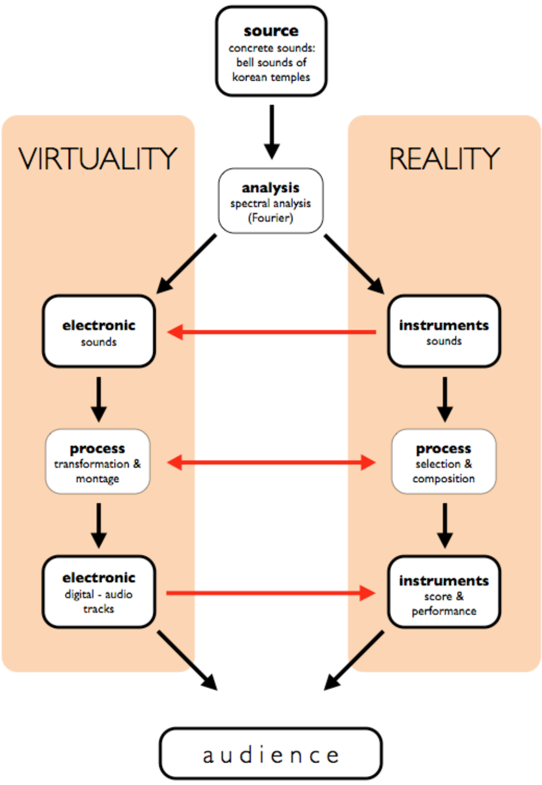
The structural materials (the frequencies of the sound's partials) steer the control of the recorded sounds of the instruments and so will be formed the electronic, the virtual part of the composition. The instruments interact with the raw materials and with the electronic in a structural level.

Both parts of the work (the real one and the virtual one) will be processed with various systems whose handlings also interact in reciprocal form.

The final product of the processes will be the digital audio-tracks (virtual) and the score (real). By the performance, the electronic will interact with the musicians (score), modifying the interpretation of the work in a controlled manner.

The audience will be the observer of the last step on a chain of interaction processes, a sort of feedback without the perception of the dry signal. They will be confronted at the same time with the virtual part of the work (electronic) as well as the real one (score-musicians). The listener will also go through a sort of live-interaction, which will be processed as a perception in a subjective way.

The last step of the interaction-chain will be also controlled through the spatialization of the electronic: the audience and the musicians will perceive the electronic in a dynamic movement through the room, which will also influence the subjective recognition of the audio data and so the final impression (audience) and the interpretation (musicians).



Protolith: Composition for marimba and spatialized electronics

Paul A. Oehlers

Audio Technology Program,
American University,
4400 Massachusetts Ave. NW, Washington, DC 20016 USA
oehlers@american.edu

ABSTRACT

Written with form as the resultant of a process that creates sections of similar and contrasting elements, *Protolith* was written in celebration of the ICAD 2010 conference specifically for percussionist Nobue Matsuoka-Motley. The concert at the conference marks the world premiere of the piece.

1. PROGRAM NOTES

Protolith attempts to derive formal structure by creating sections of music with unified global parameters (spatialization, rhythm, tempo, and meter) and juxtaposing them with elements of contrasting types (decreasing tempo vs. continuous tempo, unmetered vs. metered, close vs. far). The sections of similar and juxtaposed elements form the basis of the piece. The overall unifying parameter of the piece is timbre. Protolith refers to the lithography of a metahorphic rock. Metamorphic rocks can be derived from any other rock. They therefore have a wide variety of protoliths.

Protolith was written using various software synthesizers, resonating filters, convolution processes, and sounds and effects created with electronic and recorded sound, assembled in Pro Tools, and spatialized with VRSonic's Vibe Studio software. Sections were assembled independent from each other and combined to form the global structure of the piece.

2. VIBE STUDIO

Protolith employs VRSonic's Vibe Studio in order to simulate realistic 3D environemtns though which the sounds of the piece travel. “VibeStation is a first of its kind virtual sonic environment design and runtime application. It provides a comprehensive editing suite for creating and incorporating spatial audio content into exhibition, post-production, lecture, immersive theater, or simulation systems. Based on SoundScape3D technology, VibeStation provides, for the first time, an integrated design environment for creating virtual sonic environments.

As a dynamic runtime environment, VibeStation is a powerful real-time audio simulation tool that allows users to build and apply realistic audio environments to networked simulations or research environments with no additional programming. Connect directly to your simulation using InterfaceLink™ technology and see how immersive audio adds a true sense of realism.” [1]

3. ABOUT THE PERFORMER

Nobue Matsuoka began studying marimba at the age of ten with her aunt, Kayoko Kito in Nagoya, Japan. She came to the United States in 1989 and studied percussion at Loyola University in New Orleans with Jim Atwood of the Louisiana Philharmonic Orchestra. She won the Aspen Music Festival Percussion Competition in 1994, became the national winner of the 1995 Music Teacher National Association Young Artist Competition in Percussion and graduated from Loyola with honors. In 1998, she received a master’s degree in percussion performance from Southern Methodist University where she studied with Douglas Howard of the Dallas Symphony Orchestra.

As an active orchestral percussionist, Nobue's professional career includes performances with the Louisiana Philharmonic Orchestra, the New Orleans Opera, the Dallas Symphony Orchestra and the Nagoya Philharmonic Orchestra in Japan. She was a semi-finalist for the Buffalo Philharmonic and the Houston Symphony and a finalist for the Nagoya Philharmonic Orchestra. In 2003, the Gambit Weekly of New Orleans, honored her performance "Sticks and Strings II" with the Tribute to the Classical Arts Award for Best Chamber Performance.

She has worked for Google, Inc. as a Japanese Quality Rater, a Reference/Technical Services Librarian at Notre Dame Seminary and a Public Services Assistant/ILL specialist at Loyola University in New Orleans. Nobue recently moved from New Orleans, LA to become the Music/Performing Arts librarian at American University in Washington, DC. [2]

4. REFERENCES

[1] www.vrsonic.com

[2] http://www.musicacademyonline.com/performers/nobue.php

RHYTHMIC GAIT SIGNATURES FROM VIDEO WITHOUT MOTION CAPTURE

Jeffrey E. Boyd

Department of Computer Science
University of Calgary
boyd@cpsc.ucalgary.ca

Akil Sadikali

Department of Computer Science
University of Calgary
akil.sadikali@gmail.com

ABSTRACT

The goal of gait biometrics is usually to identify individual people from a distance, often without their knowledge. As such, gait biometrics provide a source of data that ties a visible pattern of motion to an individual. We describe our work to convert one particular biometric gait signature into a rhythmic sound pattern that is unique for different individuals. We begin with a camera viewing a person walking on a treadmill, then extract a phase configuration that describes the timing pattern of motions in the gait. The timing pattern is then converted to a rhythmic percussion pattern that allows one to hear differences and similarities across a population of gaits. We can also hear phase patterns in a gait independent of the actual frequency of the gait. Our approach avoids the inconvenience and cost of traditional motion capture methods. We demonstrate our system with the sonification of 25 gaits from the CMU Motion of Body database.

1. INTRODUCTION

Gait is ubiquitous: it is important for personal mobility, and we frequently observe the gaits of the people around us. Our observations of gait are usually visual but are occasionally audio. We often feel that we can identify a friend from afar by viewing their gait. Familiar colleagues produce sounds through their footsteps in the corridor that we recognize even when we cannot see them. This paper presents some of our work aimed at finding connections between gait and sound. While our motivation is largely based on curiosity, the conversion of human motion to sound has potential applications in athletics and therapy.

An obvious approach to gait sonification is to start with a motion capture system to acquire temporal signals corresponding to joint trajectories of a person as they move. Motion capture is a well-developed technology, offering accurate joint trajectories in real time. However, motion capture has some disadvantages. Motion capture is expensive (at least with respect to the apparatus we propose in this paper). Video- and marker-based systems require that all motion be performed within the field of view of a set of cameras. Motion that covers large distances requires many cameras resulting in increased costs. It takes time to attach the markers to a subject. An alternative to video and markers is to attach sensors to the body (even more time-consuming than markers), but this can interfere with the motion of a subject and even be dangerous for some athletic activities.

Past interest in gait biometrics suggests methods of analyzing gait without conventional motion capture [1]. This is because the use of markers or sensors on a subject's body would not be practical for biometrics. Furthermore, biometrics by necessity find variations in gait that can identify individuals. Therefore, if a biometric gait recognition can produce data unique to an individual gait, it should also be possible to sonify that data to produce a sound that uniquely corresponds to that gait.

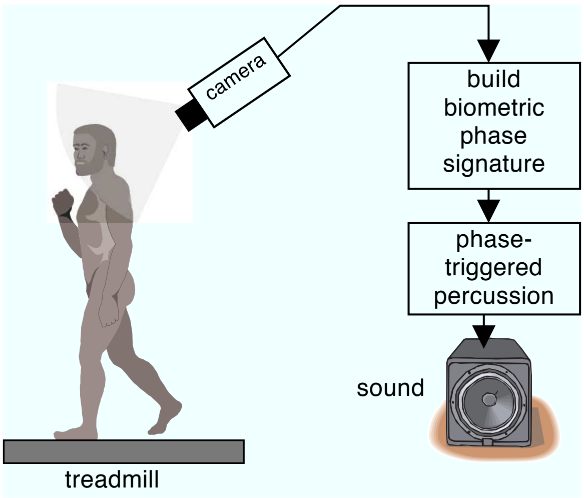


Figure 1: Schematic of phase-locked gait sonification system. The system captures video images of a subject walking on a treadmill and then builds a biometric signature based on the phase configuration of the gait. The system then sonifies the gait signature to allow a clinician and the subject to hear the phase relationships in the gait.

Figure 1 shows the gait sonification system that we propose. A camera views a person walking on a treadmill, but there are no markers or sensors placed on the body. The system builds a biometric signature from which it extracts phase data that describe the relative timing of motions within the gait [2]. As phase signals pass thresholds, the system triggers percussion events to produce a rhythmic portrait off the gait. We demonstrate the system with gaits found in a database intended for testing biometric systems. It is possible to hear the differences and similarities among gaits.

2. BACKGROUND

2.1. Sound and Motion

Many have investigated relationships between human motion and sound. Effenberg [3] and Effenberg and Melzer [4] describe methods for sonification of human motion. They measure the motion of subjects using a variety of methods including a motion capture and pressure-sensitive plates. They display properties of the gait such as force, velocity, and acceleration of body parts by coding data values to pitch. Higher pitches indicate faster velocities or higher accelerations. Effenberg concludes that augmenting the visual display of the motion with sonified data allowed observers to better estimate some motion parameters. Schaffert et al. [5] examine the use of sonification in training elite athletes. Vogt et al. [6] describe a sound feedback system in which a subject "triggers and controls sound parameters" with movement. Some studies describe the use of musical rhythms to influence gait [7, 8]. In these studies, subjects try to match external rhythms to their gait pattern while an observer records their success. The observations focused only on the heal strike and ignored the remainder of the gait. The analysis and observations relied on manual interpretation of the data.

2.2. Gait Biometrics

Biometric systems extract features from people such as finger prints, patterns in the iris and voice properties, to form a numerical *signature* that is unique to an individual. The signature can therefore be used to recognize individuals or verify their identity. While the primary goal of biometric systems is recognition and verification, the extraction of unique physical features has applications in other areas. Since our goal is to produce sounds that relate to how individual people walk, *gait biometrics* offer methods that can extract from a gait precisely the information we need.

Recent interest in biometrics that can be collected covertly resulted in the publication of a plethora of methods for gait biometrics [1]. This desire for covert acquisition means that gait biometrics do not use markers or sensors placed on the body, as is normally required by a motion capture system. The absence of markers and sensors frees a gait analysis system to be used with more versatility, and at lower cost.

A critical property for the perception of gait, and indeed for producing a gait, is phase locking [9]. This means that the various body parts that are moving periodically in the gait are moving at the same frequency and with a fixed phase difference. For example, the left and right legs operate in opposing phases, the right arm swings in phase with the left leg, and the full extension of the shin (knee lock) normally happens slightly after the forward extension of the thigh. Subtle variations in these phase relationships can provide clues to identity, a fact that is exploited by Boyd [2]. Boyd uses an array of phase locked loops to determine the phase of pixel-intensity oscillations in a sequence of video images of a gait. Given that the pixels are alternately covered and uncovered by body parts as they move through the gait cycle, the phase of pixel intensities is directly related to the phase of motion of the corresponding body parts. The *phase configuration* of a gait acts as a biometric signature for recognition, and can also recognize variations in gait across individuals such as walking on an incline versus on a level surface, and walking fast versus walking slow.

2.3. Gait Databases

The biometrics community has provided several publicly distributed gait databases suitable for testing a variety of gait analysis methods. Among the best known gait databases are the University of California, San Diego [10], Carnegie Mellon University *motion of body* (MoBo) [11], University of Southampton [12], and University of South Florida [13] databases.

We demonstrate our system with the MoBo database. It contains samples for 25 subjects. The subjects walk on a treadmill and are recorded by multiple cameras from different viewing angles. Samples for each subject show walking slowly, walking fast, walking on an incline, and walking while carrying a ball. Each sequence contains images covering 10 seconds of time, sampled at 30 frames per second.

3. GAIT BIOMETRIC SONIFICATION SYSTEM

Our gait sonification system (Figure 1) consists of three parts: video gait capture, computation of the biometric signature, and the sonification of that signature. This section describes these components.

3.1. Video Gait Capture

Our testing was based on the subset of the Mobo database that shows the *fast walk* from the side. The side view best reveals the leg, arm, and body motion in the gait. We restricted ourselves to the *fast walk* samples only so that we would have a consistent way to compare the gaits of the 25 individuals.

While it is possible to compute the biometric signatures from figures against an arbitrary background, an initial figure-to-background segmentation forces the amplitudes of pixel oscillations to be uniform. The MoBo database provides all sequences with segmented figures. If one wishes to move beyond the MoBo samples, chroma-keying or any of a number of background subtraction methods published in the computer vision literature can perform this task.

3.2. Biometric Signature

The biometric signature is the *phase configuration* proposed by Boyd [2], and summarized in the following.

3.2.1. Video Phase Locked Loops

A phase-locked loop (PLL), shown in Figure 2(a), is a control system that synchronizes the oscillations in a *voltage-controlled oscillator* (VCO), u_2, to an incoming oscillating signal u_1. The VCO has a center frequency, i.e. the frequency at which it oscillates when the input is zero. A change in the input to the oscillator changes the frequency of its oscillations. Note that the term *voltage* reveals the PLL's origins in electrical engineering. For our purposes, the oscillator is controlled by a numerical input value.

To understand the role of the feedback in the PLL, suppose that u_1 is a sinusoid at the center frequency of the VCO. The PLL reaches a steady state where u_1 and u_2 have identical frequency and phase. The phase difference, u_d, computed by the *phase detector*, is zero. Ignoring the *loop filter* for the moment, the zero phase difference feeds back to the VCO which continues to oscillate at the center frequency and stays in phase with u_1. Now suppose that u_1 increases in frequency. This will cause the phase difference to increase, and the frequency of VCO to increase until it matches the frequency of u_1. In the new steady state, u_2 has the same frequency as u_1, and the phase difference, u_d is constant. Thus, for a sinusoidal input, the PLL will reach a steady state where the VCO matches the frequency of, and is *phase-locked* with the input. The role of the *loop* filter is to remove high-frequency output from the phase detector that is not related to the phase difference. For our purposes, the PLL is a mechanism to measure and track oscillations in images.

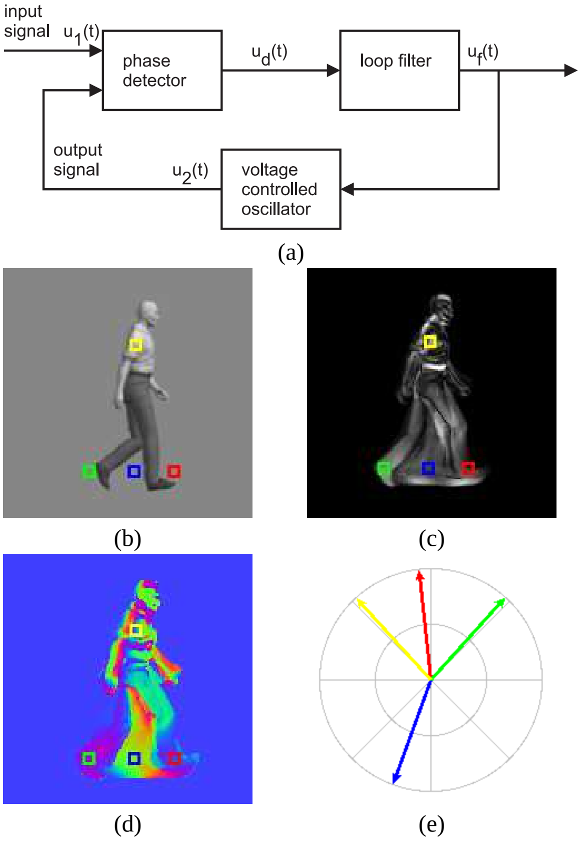


Figure 2: Video Phase Lock Loops: (a) the feedback loop used to lock on to the oscillations in a single pixel, (b) sample raw input image with four points selected, (c) the magnitude and (d) phase of the corresponding oscillations in the video sequence, and (e) the *phasors* (phase vectors) corresponding to the four selected points. As the gait proceeds in time, the phase vectors rotate counter clockwise.

A video phase-locked loop (VPLL) is simply an array of independent PLLs, one for each pixel in a video sequence. Each component PLL locks onto the the oscillations at its position in the image. Since the gaits are themselves phase locked, the component motions of the gait oscillate with the same frequency. Therefore the PLLs in a VPLL all lock to the same frequency, i.e., the fundamental frequency of the gait, and the relative phases of the PLL oscillators are the relative phases of oscillations in the gait image. The array of phase measurements for a video sequence is a *phase configuration* that can be used as a biometric signature.

Figure 2(b)-(e) illustrates the VPLL in operation. Figure 2(b) shows a single frame from a video sequence of a person walking. A VPLL locks onto the oscillations in each pixel to produce two images: a magnitude image (showing the magnitude of the oscillations, Figure 2(c)), and a phase image (showing the relative phases of the oscillations, Figure 2(d)). We can use these images as a whole, or examine the phases at select positions. Figure 2(e) shows *phasors* (phase vectors) for the points delineated in Figure 2(b)-(d), plotted on a unit circle. The phasors rotate with the gait making one rotation per stride in the gait. It is the relative phases that are useful as a biometric, and that we want to hear in our sonification of the gait.

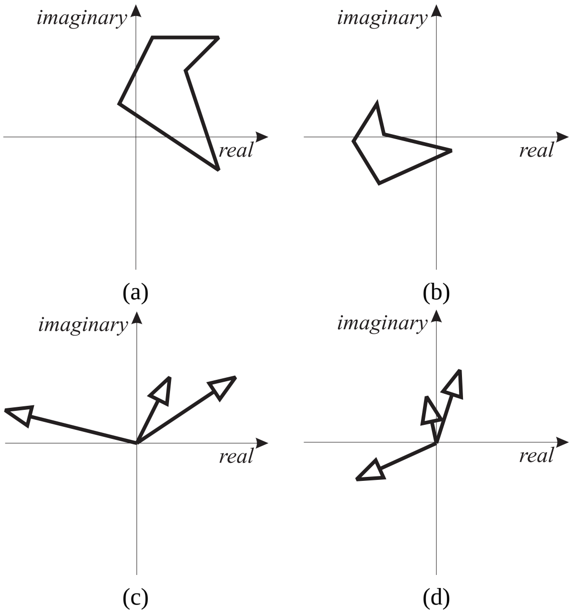


Figure 3: Procrustes analysis applied to shape and phase configurations. In the conventional application, a shape is represented by a vector of complex vertices. The shape in (a) is the same as the shape (b) because one is a translated, scaled, and rotated version of the other. A phasor configuration is also a vector of complex numbers. The configuration in (c) is the same as that in (d) because one is a rotated and scaled version of the other. Rotation is always about the origin so translation is omitted.

3.2.2. *Directional Statistics*

The VPLL operates an array of PLLs independently. In practice, variations in the position of the walker on the treadmill over time, sporadic errors in background subtractions, and spurious changes in the gait (e.g., as the subject raises an arm) all affect the phase measurements at some pixels. The PLL tracks these faithfully, but when transformed into a sound, we hear the anomalies more than we hear the gait. To prevent this, we *stabalize* the phase configurations by averaging over time. Given that the phases are directions that vary over time, we use *directional statistics*.

Procrustes shape analysis is a method in directional statistics [14] that can summarize (by finding means) and compare (using distance measures) shapes. We can represent a phase or timing pattern in a gait as a set of directions, which is mathematically equivalent to a shape, making Procrustes analysis a useful tool for analyzing the phasor patterns that emerge from gaits.

The following is a summary based on Mardia and Jupp [14]. Describe a shape in two dimensions using a vector of k complex numbers, $\mathbf{z} = [z_1, z_2, \ldots, z_k]^T$, called a configuration. Two configurations, $\mathbf{z}_1$ and $\mathbf{z}_2$, represent the same shape if by a combination of translation, scaling, and rotation, their configurations are equal, i.e.,

$$\begin{aligned} \mathbf{z}_1 &= \alpha \mathbf{1}_k + \beta \mathbf{z}_2, \quad \alpha, \beta \in \mathcal{C} \\ \beta &= |\beta| e^{i \angle \beta}, \end{aligned}$$

as shown in Figure 3(a) and (b). That is, $\alpha\mathbf{1}_k$ translates $\mathbf{z}_2$, and $|\beta|$ and $\angle\beta$ scale and rotate $\mathbf{z}_2$. It is convenient to center shapes by defining the centered configuration $\mathbf{u} = [u_1, u_2, \ldots, u_k]^T$, $u_i = z_i - \bar{z}$, and $\bar{z} = \sum_{i=1}^{k} z_i/k$. We can find the mean of a set of n shapes by finding the μ that minimizes the objective function

$$\min_{\alpha_j, \beta_j} \sum_{j=1}^{n} \|\mu - \alpha_j \mathbf{1}_k - \beta_j \mathbf{u}_j\|^2. \quad (1)$$

To find μ we compute the matrix

$$\mathbf{S}_u = \sum_{j=1}^{n} (\mathbf{u}_j \mathbf{u}_j^*)/(\mathbf{u}_j^* \mathbf{u}_j). \quad (2)$$

The *Procrustes mean shape*, $\hat{\mu}$, is the dominant eigenvector of $\mathbf{S}_u$, i.e., the eigenvector that corresponds to the greatest eigenvalue of $\mathbf{S}_u$.

Although Procrustes shape analysis is intended for treating two-dimensional shapes, it is easily adapted to handling vectors of phasors [2]. A vector of complex phasors, or a phasor configuration, is equivalent to a shape configuration, as illustrated in Figure 3(c) and (d). Shapes are invariant through translation, scaling, and rotation. Translational invariance is achieved by using the centered configuration $\mathbf{u}$. When using phasors the issue of translation becomes irrelevant. All phasors rotate about the origin, $0 + 0i$, at the entrained frequency, and the configurations, $\mathbf{z}$, are already centered, i.e., $\mathbf{z} = \mathbf{u}$.

3.2.3. Mean Configuration as Biometric Signature

In the examples reported later in this paper, we produce a biometric signature that is a mean phase configuration for each subject by the following steps.

1. Allow the VPLL time to lock for the first 100 frames of a sequence.
2. Align the VPLL output for the next 40 frames (1.3*s* at 30*fps*) so that the oscillating figures have a stationary center.
3. Crop and resample the oscillating region to 21 by 21 pixels. The lower resolution makes the computation of the eigenvectors of $\mathbf{S}_u$ tractable.
4. Compute the Procrustes mean over the 40 21-by-21 configurations by computing the eigenvalues and eigenvectors of $\mathbf{S}_u$.

The end result is a biometric signature that is 441-element complex vector representing the relative phases of pixel oscillations in the observed gait.

3.3. Sound Generation

Figure 4 describes the process by which we convert the gait biometric into percussive sound. First, we expand the biometric signature in time to form a periodic sequence using

$$\theta_i(t) = \left(\phi_i + 2\pi\frac{t}{20}\right) \bmod 2\pi, \quad (3)$$

where ϕ_i is the phase of the i^{th} element of the biometric signature, and $\theta_i(t)$ is the value of the corresponding element expanded at time $t = 0, 1, \ldots, 19$.

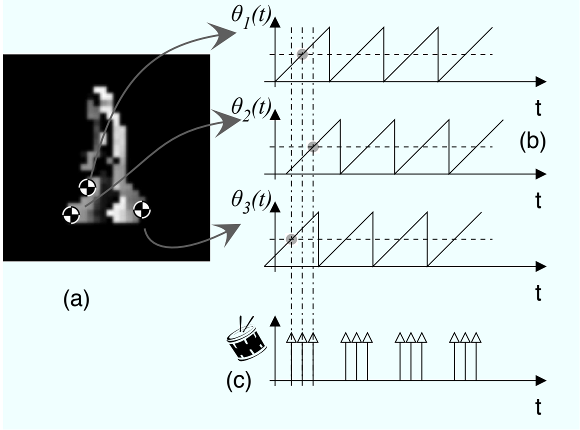


Figure 4: Methods for sonification of the gait signature. (a) The procrustes mean configuration is expanded in time and sampled at selected points. Pixel values correspond to a phase in the range $[0 \ldots 2\pi)$. (b) The temporal signal at the selected points is a phase ramp in time. (c) As each phase signal crosses a phase threhsold, the system triggers a percussion event. The resulting sound is a rhythmic pattern synchronized to the gait and correlated with the individual gait.

When sonifying the biometric signature, we play $\theta(t)$ at 30*fps*. Note that Equation (3) forces the expanded sequence to have a period of 20 samples. Viewed at 30 *fps*, this corresponds to a stride period of 0.67*s*, a value within the range typical of human gaits. Consequently, all sonified sequences have the same gait frequency and the similarities or differences we hear are due only to the phase information, not the frequency of the individual gaits. Figure 4(a) shows a single frame from an expanded sequence.

Then next step is to extract the phase at selected positions in the gait. We have no rules about which points to select, but in order to compare gaits, we must be consistent across our set of subjects. We opted for three positions within the biometric signature:

1. the forward extent of the knee motion ($\theta_1(t)$),
2. the forward extent of the foot motion ($\theta_2(t)$), and
3. the rearmost extent of the foot motion ($\theta_3(t)$).

Figure 4(a) shows these positions. Equation (3) forces the temporal signal at the sample points to form a ramp with a period of 20 frames, and relative phases determined by the biometric signature, ϕ, as shown in Figure 4(b).

The last step is to trigger percussion sound events as the $\theta_1(t)$, $\theta_2(t)$, and $\theta_3(t)$ cross a reference phase, Figure 4(c). In this case, the reference phase is π (or 0.5 normalized to the circumference of the unit circle).

4. IMPLEMENTATION AND TESTING WITH MOBO

We implemented the biometric signature computations and the temporal expansion (Equation (3) in Octave (*http://www.gnu.org/software/octave/*). The expanded sequences were then stored as movie clips, each 20 frames in duration, showing a single cycle of the expanded biometric signature.

A Pure Data (PD, *http://puredata.info/*) *patch* does the final steps of the sonification. *Gem* extensions to PD read the 20-frame clips, playing them cyclically. Custom extensions written in both *C* and *Python* sample the video and implement the phase triggers. The PD patch produces a synthetic drum sound in response to the phase triggers.

We manually selected three positions as described above for each of the 25 sequences from the MoBo database corresponding to a side view of the fast walk. The system allows us to switch subjects (recalling the manually selected positions), experiment with different sample position and sounds, and to record resulting sounds. All of this can be run on a current generation laptop computer with suitable headphones or speakers.

5. DISCUSSION

While experimenting with our system, we identified roughly seven groups of distinct rhythmic audio pattern. The grouping is subjective and we do not know whether or not other observers would make the same groupings. It is, however, reasonable to say that there are distinctive patterns, but that the patterns aren't specific enough to easily resolve all 25 subjects. This is consistent with Boyd's observation that as a biometric, the phase information adds modest but measurable improvements to the recognition of individuals [2].

If we adjust the playback rate of the sequences so that they match the frequency of the original gait, we get a much different impression of the variations in gait. As suggested by Kuo [15], body mass and dimensions affect gait frequency. Therefore, when we extract phase only and ignore the frequency information, we remove an important part of what disambiguates gaits. It is perhaps a strength of our approach that we can separate these aspects of the gait.

Computing a gait signature off-line, then playing back for sonification has limited practical use. For real value, we must compute and sonify in real-time to give immediate feedback to the person walking or perhaps to a clinical observer. While we have experimented with this, it is inherently more difficult to do for the following reasons.

1. A person's position on a treadmill tends to change gradually as they drift forward and backward over time. For real-time sonification, this tracking has to be done reliably in real-time.
2. A mechanism is necessary to find the selected sample positions reliably and continuously in light of the tracking problem mentioned above. Furthermore, variety in body and gait dimensions also confounds automatic selection of sample positions.
3. The stability of the sonified rhythm is improved with Procrustes averaging. As implemented, this is slow and we have reduced the spatial resolution of our data to compensate. Methods for efficient on-line computation of the eigenvalues and eigenvectors would ameliorate this.

Traditional motion capture methods may solve some of these problems, but they do so with high-cost equipment and a loss of convenience.

6. REFERENCES

[1] J. E. Boyd and J. J. Little, "Silhouette-based gait recognition," in *Encyclopedia of Biometrics*, S. Z. Li, Ed. Springer, 2009, pp. 646–652.

[2] J. E. Boyd, "Synchronization of oscillations for machine perception of gaits," *Computer Vision and Image Understanding*, vol. 96, pp. 35–59, 2004.

[3] A. O. Effenberg, "Movement sonification: Effects on perception and action," *IEEE Multimedia*, vol. 12, no. 2, pp. 53–59, 2005.

[4] A. Effenberg and J. Melzer, "Motionlab sonify: a framework for the sonification of human motion data," in *Ninth International Conference on Information Visualisation*, 2005, pp. 17–23.

[5] N. Schaffert, K. Mattes, and A. O. Effenberg, "A sound design for the purposes of movement optimization in elite sport (using the example of rowing," in *International Conference on Auditory Display*, Copenhagen, Denmark, May 2009.

[6] k. Vogt, D. Pirró, I. Kobenz, R. Höldrich, and G. Eckel, "Physiosonic-movement sonification as auditory feedback," in *International Conference on Auditory Display*, Copenhagen, Denmark, May 2009.

[7] J. Hamburg and A. A. Clair, "The effects of a movement with music program on measures of balance and gait speed in healthy older adults," *Journal of Music Therapy*, pp. 212–226, 2003.

[8] M. J. Staum, "Music and rhythmic stimuli in the rehabilitation of gait disorders," *Journal of music therapy*, vol. 20, no. 2, pp. 69–87, 1983.

[9] B. I. Bertenthal and J. Pinto, "Complementary processes in the perception and production of human movements," in *A Dynamic Systems Approach to Development: Applications*, L. B. Smith and E. Thelen, Eds. Cambridge, MA: MIT Press, 1993, pp. 209–239.

[10] J. J. Little and J. E. Boyd, "Recognizing people by their gait: the shape of motion," *Videre*, vol. 1, no. 2, pp. 1–32, 1998.

[11] R. Gross and J. Shi, "The cmu motion of body (mobo) database," Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-01-18, June 2001.

[12] J. Shutler, M. Grant, M. Nixon, and J. Carter, "On a large sequence-based human gait database," in *Fourth Int. Conf. Recent Advances in Soft Computing*, Nottingham, UK, 2002, pp. 66–71.

[13] S. Sarkar, J. Phillips, Z. Liu, I. Robledo, P. Prother, and K. W. Bowyer, "The humanid gait challenge problem: data sets, performance, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, February 2005.

[14] K. V. Mardia and P. E. Jupp, *Directional statistics*. Chichester: Wiley, 2000.

[15] A. D. Kuo, "A simple model of bipedal walking predicts the preferred speed-step length relationship," *Journal of Biomechanical Engineering*, vol. 123, pp. 264–269, June 2001.

SONIFICATION STRATEGIES FOR EXAMINATION OF BIOLOGICAL CELLS

Alistair D N Edwards[1]*, Andy Hunt*[2]*, Geneviève Hines*[3]*, Vanessa Jackson*[4]*, Alyte Podvoiskis*[2]*, Richard Roseblade*[2]*, Jon Stammers*[2]

[1]Department of Computer Science
[2]Department of Electronics
[3]Department of Biology
University of York
Heslington, York, UK YO10 5DD
alistair@cs.york.ac.uk

[4]Cytology Department
Leeds Teaching Hospitals NHS Trust
Britannia House
Morley,
Leeds UK LS27 0DQ

ABSTRACT

Cervical cancer is one of the most preventable forms of the disease thanks to the fact that pre-cancerous changes can be detected in cervical cells. These cells are examined visually under microscopes, but the objective of this project was to ascertain whether their examination could be improved if the visual inspection were accompanied by an auditory representation. A number of different sound mappings were tested. This paper also traces the way the sound experiments evolved in parallel with the underlying research on cell image analysis. The main conclusion is that in this kind of application, the important parameters to sonify are the ‘badness’ of the cell and the reliability of that rating, and some likely sound mappings to convey this information have been identified.

1. BACKGROUND

Cervical cancer is a slow onset disease whose precursor signs can be detected by inspecting visually, under magnification, samples of cervical cells. The UK National Health Service (NHS) cervical screening program organizes in England the collection and inspection of about 4 million samples each year [1]. It is a highly successful program which saves an estimated 4,500 lives each year in England [2].

The work described in this paper is part of a project which aims to produce an auditory representation of the visual information contained in the sample slides, as a means of increasing the number of clues on which the cytologist (medical person working on cell analysis) bases his/her decision on the normality of the sample. The ultimate aim is to improve the accuracy of screening, thereby to reduce the number of errors (false negatives and false positives) and hence to improve efficiency, reduce stress and in some cases to save lives.

In order to achieve this, a mapping from the existing (visual) data to sounds had to be devised. This paper describes a number of approaches that were tested. It represents work-in-progress. There is not, as yet, an optimum sonification tool, but it is felt that lessons have been learned along the way that will be of use to other researchers. The work illustrates some of the problems of making decisions in the vast space of sounds as well as some of the practicalities of developing sonifications in parallel with research on the phenomena to be sonified.

2. REVIEW

The practice of medicine can be very much a multi-modal skill. Traditionally doctors have relied on touch, smell and hearing as part of the diagnostic process and many are skeptical of the modern trends towards purely visual and numerical approaches.

The stethoscope is an example of the medical use of sound. It is not a sonification, as such, since it directly presents existing sounds (there is no data transformation involved) but nevertheless it demonstrates the power of sound in this context.

Experiments have been carried out on the use of sonification in medical applications. An excellent summary of these was presented in a tutorial by Hermann and Baier at ICAD 2006 [3].

As suggested above, modern medicine relies to a great extent on visual representations of data including the kinds of line graphs generated by machines such as electrocardiographs and electroencephalograms (ECG and EEG) for heart and brain monitoring. Physicians learn to recognize patterns in these traces which are indicative of particular conditions. A number of researchers have investigated the power of sonified alternatives, in which the doctor may *hear* the crucial patterns, including ECGs [4] and a number of different attempts to sonify EEGs [5-9]. Electromyography (EMG) is a similar technique for evaluating and recording the activation signal of muscles, and these have also been sonified [10].

Sonification has been applied to the identification of diseased tissue in magnetic resonance imaging (MRI) images [11]. Another experiment was relevant in that it was concerned with the identification of malignancy [7]. This uses a vocal encoding. Grayscale images are reduced to a vector of three values per pixel, ‘the first denoting the probability that the pixel belongs to an abnormal nucleus, the second being the probability that the pixel belongs to a normal nucleus, and the third being the probability that the pixel does not belong to nucleic tissue.’ (*ibid.*) These values are used to control parameters of vocal tract models in generating vowel sounds.

A previous attempt to sonify cells was carried out by Nattkemper and colleagues [12]. They investigated multi-channel fluorescence images of cells in a blood sample, whereby the intensity values identify the presence of a molecule via immunofluorescence. Their sonification was also vocal-related, based on the mapping of data vectors to diphones, thereby generating ‘artificial words’. Testing was carried out with non-biologist participants under three conditions: *visual*, *auditory* and *combined*. Participants had to match sample cells to a reference cell and classify them as either *identical* or *different*. These were then scored as either *correct, false positive* or *false negative*.

Results showed no difference between the three conditions. However, this result should not be discouraging. As long as the combined results are no worse than the (conventional) visual method, there is scope for improvement. In particular Nattkemper et al. were working with non-experts. Furthermore, they were being tested under artificial conditions. In a more realistic environment, where technicians are examining samples for hours at a time, the use of multiple channels might prove to make a difference.

3. INTRODUCTION TO THE PROJECT

It is important to clarify the aims of this project. The idea is to support the human cytologist in making decisions about the cells under review; it is not to provide automated classification of the cells. There are viable approaches to automated screening of cervical cells (e.g. [13]). In practice these can be used to screen out clearly normal samples, but when it comes to making more difficult discriminations, human operators are still required.

By the same token, sonification in this context cannot refer to the generation of an alarm when an abnormal cell is encountered. To be able to do that would amount to automated screening.

Rather, the idea is to present the cytologist with additional information which is either not present in the visual image, or is hard to discern within it. Additional information can come from sources such as:

- the direct computation of certain cell statistics (size of cell, size of nucleus, etc.), which the cytologist needs to estimate using his/her experience;
- the microscope magnification power used to produce the audio, which could be higher than that used while screening;
- the use of image enhancement methods, for instance contrast enhancement, on particularly dark regions of the slides.

The auditory field is envisaged as a *complement* to the visual field and matching the cytologist’s screening pace.

The project involved a number of different aspects. Much effort was expended on processing the visual images in order to extract the information to be displayed in the auditory form. It was also necessary to find a suitable auditory mapping to display that information and it is this latter aspect which is presented in this paper.

A number of different approaches were investigated. These reflect development of the ideas, but also the fact that the objectives changed as the parallel research on the cell analysis changed. That is to say that ideas developed as to *what* was to be conveyed in the sounds. This paper thus represents a review of the development of the sonification strategies. It is hoped that the reader will learn about some alternative approaches to sonification, the tools that were used to create them and our lessons from coping with the shifting sands of research. Further details of this research can be found in [14].

4. BACKGROUND

Cervical cancer takes time to develop. There is usually a period when some of the cells lining the cervix develop abnormal changes but are not yet cancerous; these can give rise to cervical cancer later on. Doctors can pick up these changes through screening, and a simple treatment can prevent cancer developing.

Women who get cervical cancer have had past infections with a high-risk strain of HPV (Human Papilloma Virus, or wart virus), but the vast majority of women infected with these viruses do not go on to develop cervical cancer.

A vaccine to prevent HPV infection has now been licensed for use within the European Union. This vaccine prevents against the strains of HPV that are most likely to cause cervical cancer. However, it is not complete protection against all strains. Also, as it takes between 10 and 20 years for a cervical cancer to develop after HPV infection, it will still be important for women to carry on with cervical cancer screening.

Nowadays, cervical cancer amounts to 10% of all cancer cases diagnosed in women worldwide, with around 2,880 new cases diagnosed in the UK every year[1].

Thus cervical cancer represents one of the most preventable forms of the disease and regardless of the development of vaccination, screening is going to continue to play a vital part.

Women take part in the test by making a visit to their general practitioner’s surgery or to a family planning clinic, where a doctor or a nurse sweeps around the cervix with an implement to collect a sample of surface cells. The sample is then either smeared and fixated onto a glass slide (smear method) or preserved in a fluid (Liquid Based Cytology method) and sent to a laboratory. Women should receive the test result within 6 weeks from the date of the test[2].

At the laboratory, the samples are stained with the Papanicolaou (‘Pap’) stain. As a result of the staining process, the cells and their major components (cytoplasm, nucleus) are made visible. The sample on the slide is protected by a glass cover strip. All slides are labelled and matched to a patient database. The staining process is described in some detail in [15].

Across the UK, the preparation method used for smears is the Liquid Based Cytology (LBC) method – which gives better quality slides. The term ‘smear’ is frequently given a general meaning that includes both smears and LBC slides.

The slides go through a strict screening process, whose aims are 1) to detect any abnormal cell changes, 2) to assess the type and severity of abnormal cell change when it is observed, and 3) to report the presence of a number of infectious agents, when detected.

[1] http://info.cancerresearchuk.org/cancerstats/types/cervix/index.htm

[2] http://cancerscreening.org.uk/cervical/index.html

The number of cells per slide varies, depending on a number of factors, but it is usually of the order of 40,000 to 10,000. See Figure 1.

Two screening modes are used: the *full screen* where every cell in the slide must be inspected, and the *rapid screen*, used in quality control reviews, where only a reduced number of fields of views are inspected. Full screenings should be processed at a rate of 8-12 slides per hour and a recommended rapid screen takes about 60 seconds [16].

In a full screen, the slide is scanned methodically, in a vertical or horizontal fashion and using overlapping fields of view. The screening of a slide is usually done at a lower magnification (x10 or x20), switching to x40 if anything of interest is present on the field of view. Also, although with the LBC technique the cells are mostly arranged on the slide in a monolayer, the cells themselves have a thickness that can be explored by adjusting the lens's focus. The outline of a normal cell's nucleus should be regular and unchanging on the whole thickness of the cell. Cell clumps are also often inspected at various focus depths.

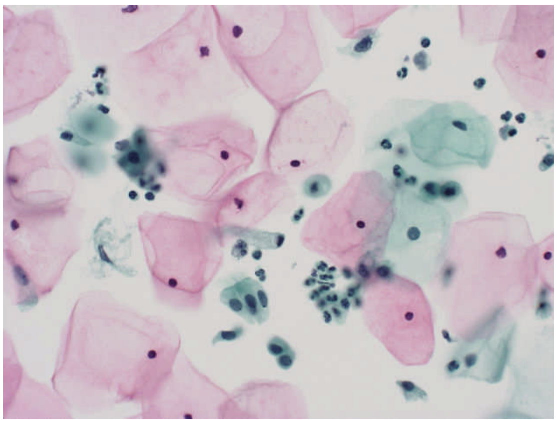


Figure 1. *An LBC slide at x40 magnification. This slide contains no abnormal cells.*

Cytologists work under a strictly controlled regime with regard to the number of hours they can work and the breaks that they must take. Despite all the care taken, errors do occur. False negatives and false positives are both to be avoided as much as possible. A false negative is clearly dangerous as it implies a woman who is likely to develop cancer believing that she is healthy. False positives cause patients unnecessary stress and over-treatment.

The objective of this project is to provide the cytologists with additional support in their task. The hope is that information encoded in sounds will help them to analyze features of cells that are hard to detect visually or even not present in the visual rendering.

5. APPROACHES TO SONIFICATION

Data represent the lowest level of information. In digital technology, data is represented (and can be measured) in *bits* and can be easily manipulated and transformed. At a higher level, data can be transformed and combined to represent *information*. This can be achieved through technology, but it is also something that people are good at. In other words, coherent data, represented appropriately can reveal *patterns*. Many branches of information technology are concerned with this kind of processing: either automatically identifying the patterns in the data, or transforming the data so that the patterns become more apparent to the human observer – or combinations of both of these. This is the objective of sonification: to transform data into an (auditory) form to facilitate pattern recognition and hence extraction of information by human users.

In the case of this project, data are available from the scanning (in visible light) of microscopic cells. Those data are conventionally presented as visual pictures (visualizations), and skilled operators learn to extract the relevant information from that representation (i.e. to recognize abnormal cells). Yet, there is no reason why the same data should not be represented in an auditory form. There are a number of potential benefits:

- information which is contained in the data but which is not apparent in the visual representation may be detected in the auditory one;
- presenting the same data on different channels simultaneously may help the user's interpretation;
- multimodal presentation may also (positively) affect other, higher-level human factors, such as concentration, attention and (alleviation of) boredom.

With these objectives in mind and given the data that were available from cell samples, appropriate and effective sound mappings had to be found. A number of different approaches were tried and they are described in the following sections.

5.1. Color mapping

Since smear slides are colored with chemical stains, an overview of the status of cells is aided by the fact that cell nuclei are colored purple, and that other colors tend to attach to certain cell attributes. Typical signs of abnormal cells include:

- enlarged cell nuclei
- irregular nuclear outlines
- uneven distribution of chromatin (nuclear material)
- generally dark staining of the nuclei.

Thus an algorithm was created which deduced the average HSV (Hue, Saturation and Value, a measure of Brightness) of a section of the slide containing several cells [17]. Using the software toolkit *Pure Data*[1] the user was allowed to move the mouse freely around the image, and sound was continually synthesized, mapping luminance and hue onto a frequency scale, and saturation onto the sound's amplitude.

The synthesis method was very simple, so that the focus could be on the effectiveness of the interaction. Frequency modulation of two sine waves was used, and a series of experiments was carried out to ensure that the more intensely dark-stained a cell was, the higher the carrier frequency, the more extreme the modulation, and the louder the overall amplitude. This has the effect of making darker areas give rise to loud, high frequency sounds which were (on purpose) rather unpleasant. This allowed the user to freely move around the

[1] http://puredata.info/

image and easily hone in on areas which were more densely and darkly stained.

In experiments, test participants were asked to identify a cell field as 'normal', 'slightly abnormal' or 'abnormal', simply by listening to the sounds produced as they moved around an image (invisible to them) of a field of cells. Our researcher Podvoiskis concluded:

> *Results from both experiments showed subjects were able to identify and classify images based on a sound representation only. These results were proven to be statistically significant.* [17]

It is interesting to note that the test participants at this stage were not trained cytologists, but music technology students, yet they were able to identify correctly the more grossly abnormal cells by sound alone.

However, these very positive effects were only apparent when grossly abnormal cells were present in suitably large clusters, and could be picked up by a user moving a mouse to 'focus in' on such denser areas. Subsequent study showed that the majority of cells which need to be identified by cytologists are usually much more borderline, and this method was not able to distinguish these. In addition, the synthesis method was very simple and would not stand up to long-term listening.

The technique of mapping colors of a cell-field to sound is still worthy of further investigation, particularly if the spatial position of each contributing cell could be portrayed in sound.

5.2. Scanning images for texture

Next, we undertook a series of experiments [18] working with *CSound*[1], to generate sounds which represented the internal structure of individual cells. One of the major indicators of abnormal cells is an irregular distribution of chromatin inside the cell nucleus.

This work explored the use of granular synthesis to create sounds whose perceived 'grittiness' portrayed the severity of the distribution of the chromatin, and was thus an indicator of abnormality. The mapping used looked at the gradient of pixel darkness to show where the dark spots were placed within the cell's nucleus. The horizontal spacing of these spots was portrayed using stereo panning; the vertical was represented by a frequency scale. The user is not allowed to freely scan the image with a mouse, but instead the computer performs an auto- scan left to right across the cell and then repeated down the cell.

The segmentation and modification of the image prior to sonification (using custom-defined image processing algorithms in MATLAB) became an important part of the work (Figure 2), but one which was time-consuming.

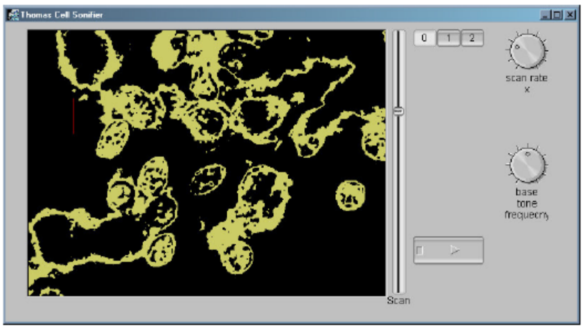


Figure 2. *Interface to the MATLAB/CSound sonification tool, allowing basic control of the audio scan carried out on visually processed cell images.*

Test participants reported that the granular sounds were highly irritating and would not be put up with for long periods.

Later phases of the work explored the use of filtered noise sounds as a 'softer and smoother' portrayal of the chromatin, and later still some more-musical notes based on piano synthesis. Some promising results were obtained by using the scanning technique to directly sonify the pixels as binary values once they had passed through the thresholding algorithm.

One of the main limitations of this method is the long time (not available to pressured cytologists) taken to:

- visually identify a cluster of cells
- zoom in to the correct resolution
- modify the image's coloration to achieve best contrast
- listen to the scanning of the nuclear data from left to right and then downwards.

However, the main problem with this approach is that, whatever the sound quality, it would inevitably be perceived as in some sense an 'average' of the cells in view, whereas what the cytologist is generally looking for is the one cell (or small number of them) which is abnormal, that is *not* average.

The following studies were then carried out to discover if it were possible to clearly portray the state of multiple cells surrounding the current position by using sound spatialization.

5.3. Sound Spatialization

We undertook an investigation into whether all the cells surrounding the user's current position could be rendered in a sonic space around the listener [17].

The software used was *Scilab*[2], an open-source computation package similar to MATLAB. Data was spatialized using Head-Related Transfer Functions (HRTFs). The image being 'viewed' was split into 9 segments surrounding the current 'centre-point'. The software produces a radar-type sweep around the image, and generates sound in the corresponding positions for a listener wearing headphones.

At this point in the research it was decided to produce a 'badness' rating for each cell undergoing examination, by pre-processing the cell data, mapping to a number from 1 to 10, where 1 is 'normal' and 10 is 'highly abnormal'.

[1] http://www.csounds.com/

[2] http://www.scilab.org/

We experimented with a variety of sonification methods to portray the ‘badness’ of each cell surrounding the listener. These included:

a) Additive synthesis, where increasingly discordant overtones are added as the badness number increases. This was found to produce mostly unpleasant sounds.

b) Sampled audio files, where sounds are used to represent a natural landscape (based on [7]). Cows gently mooing were mapped onto ‘not bad’, dogs barking were in the middle and a person screaming represented the severely abnormal cells. (Table 2).

User tests found that the sampled audio portrayal was much easier to listen to and locate. However, the apparently arbitrary choice of animal sounds came across as quite bizarre to some, and not an obvious linear mapping of ‘badness’. Future work in this area should attempt to dispense with the disorientating ‘sweep’ and to play all of the sounds together in one surround-sound field, which is much more analogous to how multiple sounds reach our ears from the real world. Based on these sounds, a questionnaire was devised where the participants were the screening cytologists of the Leeds NHS Trust. It covered questions about:

- The individual’s music preferences and listening mode (headphone, iPod, speakers, live music etc.);
- their attitude to the research (bearing in mind these are visual analysts being asked to consider audio input);

Sample Name	Original Length (ms)	Max Level (dB)	‘Badness’ range	Notes on Design Method
1.wav	14	-18.32	0-99	Created from a slice of human speech. Very short and quiet.
2.wav	80	-23.54	100-199	Unedited recording of a ‘popping’ sound made with lips.
3.wav	57	-11.24	200-299	Edited recording of a bubble popping in boiling water.
4.wav	53	-25.79	300-399	Synthesized ‘pop’ sound – high in treble content. Short reverb used.
5.wav	379	-3.69	400-499	Edited recording of noises made with the mouth. EQ applied.
6.wav	154	-0.15	500-599	Synthesized ‘pop’ combined with recording of mouth noises. EQ
7.wav	354	-5.93	600-699	Recording of another type of mouth ‘pop’, with effects.
8.wav	315	-0.01	700-799	Synthesized ‘pop’ combined with recording of mouth noises. EQ
9.wav	424	-0.19	800-899	Synthesized ‘pop’ combined with recording of mouth noises. Reverb.
10.wav	649	-0.01	900-999	Boiling water recording with heavy editing. Huge amounts of EQ and reverb used.

Table 1. *Sounds used in the sound preferences experiment.*

‘Badness’ range	Sound
0-99	cow mooing
100-199	frog croaking
200-299	horse whinnying
300-399	bird tweeting
400-499	cat meowing
500-599	seagull crying
600-699	man shouting
700-799	dog barking
800-899	monkey howling
900-999	woman screaming

Table 2. Mappings from ‘badness’ values to sounds.

- how they would prefer to interact with a sound-generating system;
- their thoughts about what different types of cell should ‘sound like’.

The questionnaire concluded with a practical session:

- The playback of several of the sounds, and the request to rate them as ‘good’ to ‘bad’, and ‘like’ to ‘hate’.
- Several cell images, with the subject being asked to select from a choice of 3 sounds which best represented that cell. (Figure 3).

Results showed that cytologists, on the whole, would like to hear an ear-catching, alarm-type sound when an abnormal cell is present, but that a quiet sound should be present the whole time, to ‘show that the system is still working’. They did not want to hear sounds which were directly related to real-world sounds (such as some of the examples water-type sounds) and many were not convinced how ‘musical’ sounds might be perceived.

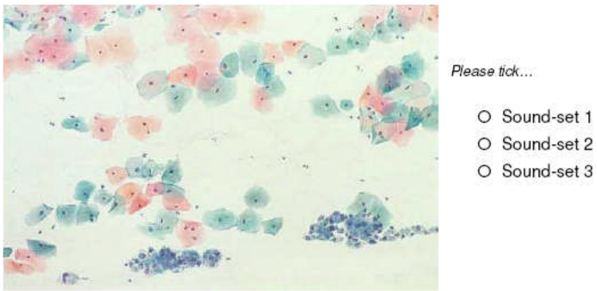


Figure 3. *Sample selection screen. Participants indicated which sound they felt best represented the cells in view.*

5.4. Subjective sound selection

It had become evident in the image analysis research that it would be possible to calculate two quantities for cells: 1) the apparent degree of abnormality and 2) the confidence of that rating. It also became evident that the distinction was not between ‘good’ and ‘bad’ cells, but rather between *normal* and *bad*. That is to say that most of the cells a cytologist will see are normal. The message to be communicated (aurally) to the cytologist for the majority of cells should be calm and neutral.

For cells which might be abnormal there should be an alerting sound (but not an alarm – see the earlier discussion) and the sound should be more insistent if the probability of abnormality is greater.

None of the previous experiments specifically provided guidance on the choice of such sounds. It was therefore decided to embark on a different kind of experiment to help with identifying suitable kinds of sounds. Some of the sounds generated in the earlier experiments were to be included, though, for comparison.

It was important to test the perception of sounds by as wide a population as possible in order to identify ones which would be likely to have the highest acceptability to any users. We would want to include specialists (cytologists) in the testing, but not to be exclusive to them. It was therefore necessary to ask people to map sounds to qualities that would be meaningful to them – and not cell images which would convey meaning only to cytologists. It was therefore decided that the mapping should be to 'Smiley faces', as in Figure 4.

In order to capture data from as wide a population as possible, the test was mounted on the Web[1]. Visitors started on a briefing page and gave their assent to taking part. They would then hear a set of 42 sounds, one at a time (and only once each). They would then select which of the Smileys they thought best matched the sound. They also had the option of selecting *Don't use this sound*, in which case they were invited to explain their opinion. This was in order to ensure that sounds which are (generally) aesthetically unacceptable could be identified. At the end of the sounds the participants filled in a short background questionnaire.

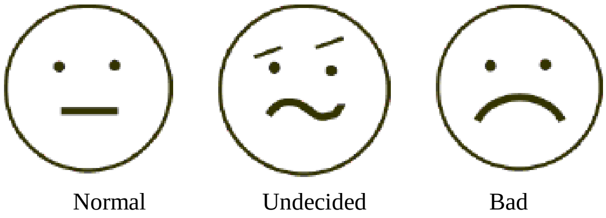


Figure 4. *The three Smileys used in the experiment.* Normal *represents most cells, which are not cause for concern;* bad *would be a cell which is almost certainly abnormal and* undecided represents *the (common) case in which the cell may be abnormal, but the probability that it is so is not high.*

The sounds used varied greatly. Some came from the previous experiments, others were everyday sampled sounds and still others were based on everyday sounds but processed in some way. We started with no preconceptions. That is to say that we had no intent as to which sounds would be mapped to which image. The aim was to find out about the *kinds* of sounds which mapped well to the categories. Later we would investigate how to create a set of sounds which would then convey the required categories – and the spaces between them. That is to say that it is not anticipated that all cells will be classified into one of the three classes; there will be a large space between (for instance) *Normal* and *Undecided*.

This experiment is continuing and it is too soon to draw any conclusions. It is perhaps not surprising that initial results suggest high levels of subjectivity in responses. This reinforces the observation that sound aesthetics are vital and subjective. It might imply that different sound sets should be provided from which individuals can select.

6. DISCUSSION

Pattern matching is a fundamental skill, not the least in medical investigations. Many researchers have remarked on the power of human hearing to detect patterns in sounds and hence have tried to apply sonification as an alternative or an addition to visual pattern recognition in medical data. That is the approach applied in this project.

The richness of the sound space gives much scope for the use of sounds – but it also poses a dilemma for the designer in making choices as to what kinds of sounds to use and how to map the relevant parameters onto them. This is a common problem, articulated in most publications on sonification. Within this project was also apparent another problem (which is probably common in other similar projects) – that the underlying application represents a moving target as the research on it develops.

The work on extracting data from the cell images and classifying it was proceeding in parallel with the development of sounds, and the ideas as to what was important about the cells changed.

The initial assumption was that all cells in the visual field should be sonified in parallel. That was dropped because it became apparent that any such sonification would effectively present an 'average' of the cells, whereas it is the one or two non-average (abnormal) cells which are important. Thus, it was decided to concentrate on the one 'most interesting' cell in the current field of view.[2]

It was realized early in the project that cytologists and others expected that sonification would amount to the playing of an alarm sound on the detection of an abnormal cell but this was technically infeasible. However, it was less apparent as to what the sounds should represent. We gained greater insight into what information could be extracted from the images and as to the nature of the cytologists' task. Thus, it became apparent that the vast majority of cells encountered are normal and no cause for concern. Then there are others which might be abnormal (or 'bad') and so we looked at the assignment of scales of 'badness' and their representation in sound. It was realized that the scale was not (as might be conventionally expected) from 'bad' to 'good', but from 'bad' to 'normal'; there are no cells which are 'better' than normal ones.

Subsequently we came to a further realization, which was that cells cannot be mapped onto a one-dimensional 'badness'

[1] http://www-users.cs.york.ac.uk/~alistair/sonify

[2] The idea of sonifying a field of cells has not been abandoned all together, though, and may be revived in future work. If it is possible to separate out different dimensions, then the 'averaging' effect may not occur. For instance, drawing on this work, it might be that the stain colours are 'heard', along with chromatin textures, but these are displayed spatially for all cells but only emphasizing the worst cases by filtering what we play according to badness.

scale. This is because the information available is not unambiguous. In other words a cell may be classified as 'bad' with different degrees of confidence. A cell which has a high probability of being bad is more significant than a) one which is classed as bad, but only with a low probability and b) one which is classed as quite bad but with a high probability. These subtleties must be captured in sound.

Cells are arranged on the two dimensions of a microscope slide. It can be assumed that the one in the center of the field of view already has the cytologist's attention, but if one off center is of interest, how should the cytologist's attention be directed to that one? Sound spatialization is the obvious mechanism. Experiments with this were positive, although it was found that the 'radar sweep' was inappropriate.

Anyone working in the use of sounds is aware of the importance of aesthetics, of subjective reactions to sounds and we have managed to find some of the preferences of cytologists in this application.

This paper has set out to frankly present this story in the hope that it will be of benefit to future researchers who find themselves working in a similarly shifting environment.

7. CONCLUSIONS

Sonification potentially has a number of applications in medicine. Whereas natural sounds have long been a part of doctors' diagnostic tools, derived sounds have still to make a significant mark in medical applications. This paper has presented one more investigation of the possibility of doing this in one particular application. The work has demonstrated a number of the real-world constraints on this kind of research.

A number of lessons have been learned including:

- Sonification must support the operator in the classification of cells; and is not a form of automatic recognition, generating alarm sounds.
- Selection of the right kinds of sounds is imperative.
- The means by which the user or cytologist interacts with the sonification interface is also very important.
- Spatialization of sounds can be helpful in locating the cells of interest.
- In this kind of application, the important parameters to sonify are the 'badness' of the cell and the reliability of that rating.

The objective of the project is to produce sample sonifications which will be tested. It is hoped that these will demonstrate that screening with sounds is at least as accurate as conventional, purely-visual screening. Work is continuing to that end.

8. REFERENCES

[1] *Cervical Screening Programme England 2005-6*. 2006, The Information Centre (http://www.ic.nhs.uk/webfiles/publications/cervicscreng2006/Cervical%20bulletin%202005-06.pdf).

[2] *Cervical screening: The facts*. 2007, NHS Cancer Screening Programmes.

[3] Hermann, T., et al., *Vocal sonification of pathologic EEG features*, in *Proceedings of the 12th International Conference on Auditory Display (ICAD2006)*. 2006: London.

[4] Ballora, M., et al., Heart rate sonification: A new approach to medical diagnosis. *Leonardo*, 2004. **37**(1): p. 41-46

[5] Jovanov, E., et al., EEG analysis in a telemedical virtual world. *Future Generation Computer Systems*, 1999. **15**: p. 255-263.

[6] Hermann, T., et al. Sonifications for EEG data analysis. in *Proceedings of the 8th International Conference on Auditory Display (ICAD2002)*. 2002. Kyoto, Japan.

[7] Mauney, B.S. and B.N. Walker. Creating functional and livable soundscapes for peripheral monitoring of dynamic data. in *Proceedings of the 10th International Conference on Auditory Display (ICAD04)*. 2004. Sydney.

[8] Hermann, T. and G. Baier. *The Sonification of Human EEG and other Biomedical Data*. 2006 [cited 2010 12 January]; Available from: http://www.sonification.de/projects/eegson-icad2006/index.shtml.

[9] Baier, G. and T. Hermann, *Event-based Sonification for EEG monitoring and analysis*, in *Proceedings of the 12th International Conference on Auditory Display (ICAD2006)*. 2006: London.

[10] Pauletto, S. and A. Hunt, *The sonification of EMG data*, in *Proceedings of the 12th International Conference on Auditory Display (ICAD2006)*. 2006: London.

[11] Martins, A., et al. Auditory display and sonification of textured image. in *Proceedings of the 3rd International Conference on Auditory Display (ICAD 96)*. 1996.

[12] Nattkemper, T.W., et al., *Look & listen: Sonification and visualization of multiparameter micrographs*, in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC2003)*. 2003: Cancun, Mexico.

[13] Parker, E.M., J.A. Foti, and D.C. Wilbur, FocalPoint slide classification algorithms show robust performance in classification of high-grade lesions on SurePath liquid-based cervical cytology slides. *Diagnostic Cytopathology*, 2004. **30**(2): p. 107-110.

[14] Edwards, A.D.N., G. Hines, and A. Hunt. Segmentation of biological cell images for sonification. in *International Congress on Image and Signal Processing (CISP 2008)*. 2008. Sanya, China: IEEE Computer Society.

[15] Roseblade, R., *Sound and interface design for improving detection of cervical cancer*, in *Department of Electronics*. 2006, University of York.

[16] *Laboratory Organisation: A Guide for Laboratories Participating in the NHS Cervical Screening Programme*. 2003. NHSCSP Publication No 14 (http://www.cancerscreening.nhs.uk/cervical/publications/nhscsp14.pdf).

[17] Stammers, J., *Developing synthesis techniques for the sonification of precancerous cells*, in *Department of Electronics*. 2006, University of York.

[18] Butterfield, T., *Improving the detection of cancer by using sonification to supplement visual displays*, in *Department of Electronics*. 2005, University of York.

Acknowledgements

Some of the work reported herein has been carried out by students at the University of York. In addition to those listed as co-authors, we acknowledge the invaluable contribution of Thomas Butterfield. The Leeds Teaching Hospitals NHS Trust has provided the medical input. In addition to the co-author, we acknowledge the assistance of David Brooke in setting up the project and of the team of cytologists who graded a large set of cell images.

The work was funded by the UK Engineering and Physical Sciences Research Council, grant number EP/C512413/1. We thank the Council for its positive attitude in extending the project's duration and acknowledge the help of John Grogan MP in obtaining this.

SONIFICATION AND VISUALIZATION OF NEURAL DATA

Mindy H. Chang, Ge Wang, Jonathan Berger

Stanford University, Stanford, CA 94305, USA
mindyc@stanford.edu

ABSTRACT

This paper describes a method for integrating audio and visual displays to explore the activity of neurons in the brain. The motivation is twofold: to help understand how populations of neurons respond during cognitive tasks and in turn explore how signals from the brain might be used to create musical sounds. Experimental data was drawn from electrophysiological recordings of individual neurons in awake behaving monkeys, and an interface was designed to allow the user to step through a visual task as seen by the monkey along with concurrent sonification and visualization of activity from a population of recorded neurons. Data from two experimental paradigms illustrating different functional properties of neurons in the prefrontal cortex during attention and decision-making tasks are presented. The current system provides an accessible way to learn about how neural activity underlies cognitive functions and serves as a preliminary framework to explore both analytical and aesthetic dimensions of audiovisual representations of the data.

1. INTRODUCTION

Our brains are able to manage a great deal of information, from taking in sensory perceptions to forming decisions and transforming plans to actions. Current research explores how this is achieved by a network of billions of interconnected neurons, communicating through electrical impulses called *action potentials*, or spikes. The activity of single neurons can be recorded through electrodes placed in the brain while subjects (in this case rhesus macaques) perform experimental tasks designed to examine specific cognitive functions. Neural responses are often very diverse, and when trying to understand how a population of neurons might work together, simply averaging across all neurons results in a loss of information, while plotting the raw responses of all neurons can quickly become difficult to interpret. Sonification offers a complementary way to explore the data and in a literal sense ties in closely with the idea of listening to a dynamic conversation among neurons during cognitive tasks.

The idea of listening to the brain has been explored at both the macroscopic and microscopic levels. Electrical signals recorded from the scalp (electroencephalogram, or EEG) have long been studied as a representation of aggregate neural population activity. Sonification of EEG signals has been applied in a variety of contexts: for scientific understanding [1], as a potential diagnostic tool for detecting abnormal brain rhythms in epileptic patients [2], and as auditory feedback for human computer interaction applications [3]. Previous work has also explored sonification of neurons isolated in culture [4, 5].

For recordings from individual neurons in awake behaving subjects, audification of neural spike trains during data collection has long been used as a tool for navigating through different areas of the brain. During *in vivo* single electrode experiments, electrophysiologists often listen to an amplified voltage signal while lowering the electrode into the brain in order to estimate depth and cortical area as well as identify neurons. Once a neuron is isolated, listening to its spike train, which sounds like a series of pops and clicks, provides a fast and convenient way to gauge, for example, how strongly a neuron responds to a particular visual stimulus in real time. The ability to listen to the neural activity while visually paying attention to the stimulus on the screen enables the experimenter to constantly monitor both. Beyond a few neurons, it becomes difficult to hear nuances within the population activity. In the current study we concurrently sonify and visualize activity from a population of neurons along with a schematic of the behavioral task being performed both to try and provide an intuitive way to identify patterns in the data and to explore different ways in which signals from the brain can be used to create musical sounds.

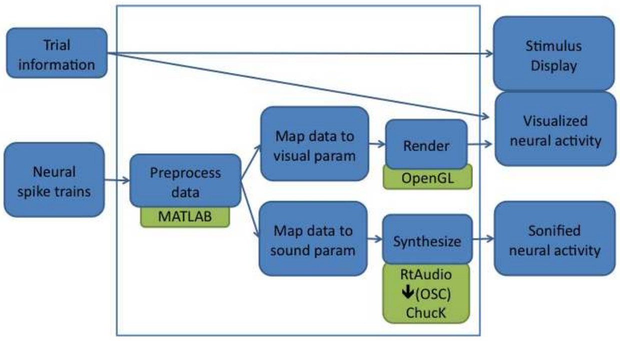


Figure 1: Schematic system diagram.

2. SYSTEM

The current implementation provides a way to explore data after it has been collected. The system enables the user to load neural spike trains and trial information and then listen to and visualize the data as it relates to a behavioral task (Fig. 1). The user can interactively play through entire experimental trials or portions of trials while being presented with a constantly updating schematic of the task performed by the monkeys as well the elicited neural responses.

2.1. Data

In a typical neurophysiological experiment, an animal is trained to repeatedly perform many trials of a carefully controlled task so that multiple instances of neural responses to a particular experimental condition can be analyzed. Three representations of data were explored: single trials, condition averages, and condition average differences. Single trials consist of spike times for each neuron with millisecond precision, whereas condition averages represent the smoothed instantaneous spike rate averaged across multiple trials of the same experimental condition for each neuron. Average differences summarize the difference in spike rate between two conditions for each neuron across time. The user can load different combinations of single trials, average trials, or average difference trials for a particular task.

2.2. Interface

The data are visualized as rasters, which are labeled and stacked as blocks on the left side of the screen. Example screenshots are shown in Fig. 3 and 5. Within a raster, each row represents one neuron's response across time. For individual trials, dots represent a spike at that particular time, whereas for the average and difference plots, spike rate is indicated by the color of the heat map. The average spike rate across all neurons over time for each raster is shown below the raster blocks and highlighted for the current raster.

As a vertical bar moves across time for a given trial, the task screen on the right updates with a schematic of the stimulus that the monkey is viewing at that particular time in the trial. For single trial rasters, the dots representing spikes are dynamically enlarged for the current time. The user can click anywhere on any raster to change the current time and use computer keyboard shortcuts to change certain parameters of the sound, such as the data to sound mapping, musical scale, speed of playback, and data integration time. To change the instrumentation, the user can manually change properties of the sound engine.

2.3. Data to sound mappings

Out of the large space of possible data to sound mappings, three were implemented for the current system, termed *avgRatePitch*, *neuronPitch*, and *eachRatePitch*. The *avgRatePitch* mapping provides the most basic summary of average population activity, where a range of spike rates is mapped to a range of pitches such that higher spike rates correspond to higher pitches. The average firing rate across all neurons is sampled every specified number of samples, and the corresponding note is played, creating a steady stream of single notes. This mapping can be used both for single trials and condition averaged trials. For the *neuronPitch* mapping, which applies to single trials, each neuron is assigned a unique pitch, and a note is played at that pitch each time the neuron spikes. If specified, the neurons can be split into two sets, each set with its own instrument. The *eachRatePitch* mapping focuses on the average spike rate of each neuron over time. Neurons are grouped into 2-4 groups, and each group is assigned an instrument. After a specified number of samples, every fifth neuron within each group is selected, and a pitch corresponding to its spike rate is played, again with higher spike rates corresponding to higher pitches. The neurons within each group are continuously cycled at every sampled time interval.

2.4. Implementation

A software system was designed consisting of a data engine that handles loading, processing, and graphical display of experimental data as well as a sound engine that handles the synthesis of sound parameters mapped from the data. Networked communication sent from the data engine to the sound engine allows for a separation of the extraction of data parameters for sonification from the actual mapping of data to sound.

Initial preprocessing of the data, which included sorting the neurons, creating matrices of spike times, and computing trial-averaged instantaneous spike rates, was done in MATLAB and output as text files. Images of the behavioral task, average activity plots, and labels were generated and saved as .raw image files. The interface was developed in C++ and uses OpenGL / GLUT for the graphical display. Timing of the playback is controlled using RtAudio [6] such that every time a specified number of samples has passed, the appropriate sound parameters are calculated and then sent via Open Sound Control (OSC) [7] to a sound engine. In the current implementation, a ChucK [8] script runs concurrently and handles synthesis. For each OSC message received, ChucK plays a single note with the specified instrument and frequency. Due to the constraints of pitch, the sonified output is stretched in time compared to the actual timing of the data such that the sonification is slowed by a minimum factor of 10.

3. RESULTS

Data from two separate experiments were used to explore how audiovisual displays of neural data might aid in understanding how behavior correlates with neural activity and in achieving different musical aesthetics. Within the context of the two experiments presented, the three different data to sound mappings highlight different aspects of the main effects in the data.

On single trials and average trials, the *avgRatePitch* mapping reflects the average envelope of activity across all neurons such that sharp onsets and offsets of overall neural activity create salient rising and falling of pitch. The amount of sustained neural activity present across a certain span of time can be estimated by the absolute pitch played, but since there is a steady string of notes, the relative intensity of activity as compared to baseline is perhaps less apparent. The constant tempo, single string of notes, and temporally smoothed profile create a steady melody that is easy to follow.

On the other hand, the *neuronPitch* mapping for single trials reflects the amount of participation from the population of neurons since spikes from each neuron have a unique and independent representation (a note played at each spike). While it is not possible to simultaneously track the activity of all individual neurons at all points in time, a sparse sound corresponds to low neural activity while a dense concentration of notes reflects the simultaneous activation of multiple neurons. A persistent sounding of particular notes indicates the elevated activity of specific neurons. For this mapping, there is no rhythmic structure imposed on the sounds. This leads to sporadic bursts of sound triggered by events that drive the activity of the neurons. Since the neurons are represented as independent notes, this mapping showcases the complexity of activity in the population on a millisecond by millisecond basis and creates a more chaotic sound.

For average and difference trials, the *eachRatePitch* mapping provides a blend of average rate and individual neuron information since the discrete sampling of individual neuron spike rates within each assigned group means that a changing subset of neurons in every group is represented at a given time. The regular sampling of activity imposes a steady rhythm on the notes, and the assignment

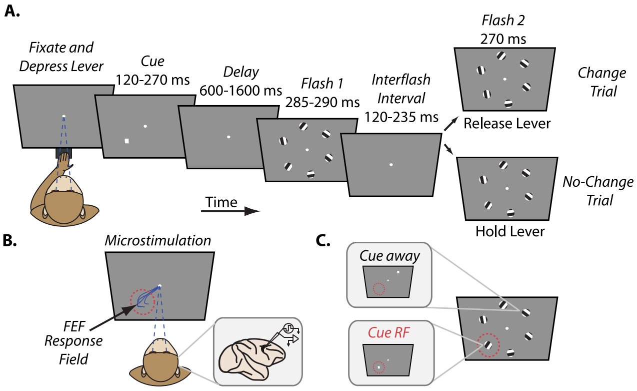


Figure 2: Attention task experiment setup. A. Behavioral task. Each monkey was trained to direct attention to a peripherally cued location in order to detect a localized change across two flashes of a stimulus array. B. Stimulus alignment. The FEF response field (RF) for each recording site was determined by applying microstimulation during a simple fixation task and mapping the evoked saccades. An example set of eye traces from microstimulation-evoked saccades are shown. The array of gratings was positioned such that one grating was centered at the average evoked saccade endpoint. C. Trials in which the monkey was cued to attend to the response field are labeled ”Cue RF,” whereas trials in which the monkey was cued to attend to the opposite array location are labeled ”Cue away.”

of instruments to each group can make the activity of some groups of neurons sound more prominent than others. In this mapping, the same number of notes plays at every fixed interval, creating a structured and continuously flowing progression of chords.

Additionally, the playback speed affects the granularity with which changes in activity across time can be perceived in that slower speeds highlight local changes while higher speeds provide more of an overview of single trial dynamics. The choice of instruments and musical scale also directly affect the overall aesthetic of the sound.

4. EXAMPLES

The following two experiments explore the different response properties of neurons in an area of the brain involved in planning and executing eye movements. Our eyes are constantly receiving sensory input, and visual attention plays a crucial role in how we experience the world. Even though it may seem like we can see everything around us, only a limited amount of information is actually selected for detailed processing. Since we have the highest visual acuity at the center of gaze, our eyes are constantly scanning to bring different visual information into focus. We can also attend to peripheral locations while keeping our eyes fixed, for example while driving and keeping an eye on the road but constantly monitoring the surroundings. A working model for how the brain might resolve these different means of selecting visual information centers on shared neural mechanisms underlying both the control of eye movements and the voluntary allocation of attention.

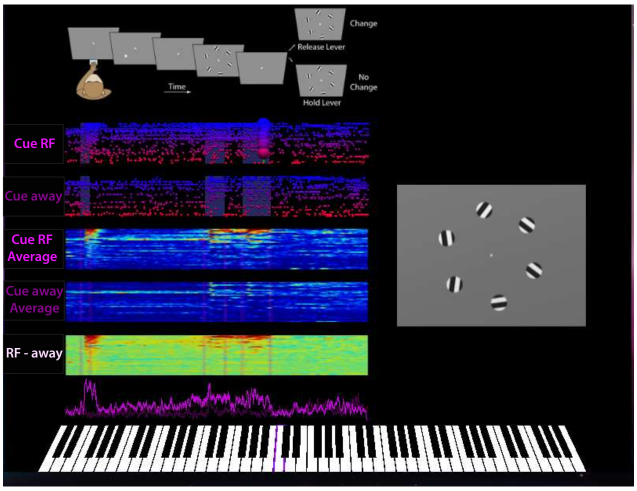


Figure 3: Example screenshot using data from the attention task.

4.1. Spatial attention task

In order to study neural mechanisms underlying visual attention, monkeys were trained to direct and sustain attention at a peripheral location without the use of eye movements (Fig. 2) [9]. During each trial, the monkey maintains fixation at the center of the screen and uses a lever to indicate whether one grating embedded among five distractors changes orientation across two flashes. A spatial cue is given early in the trial, and in order to correctly detect the grating change, the monkey needs to direct attention to the cued location. All six locations are equally likely to be cued, and the cue is always valid.

Single electrode recordings were made in the frontal eye field (FEF), which is an oculomotor area known to play a role in controlling eye movements. The FEF contains a spectrum of visual to (eye) movement responsive cells, which form a map of visual space. For a given neuron, the particular region of space that it represents is called its response field (RF). The RF’s of individual neurons are found by electrically stimulating at the recording site, which causes the monkeys to make a stereotyped eye movement (saccade) towards a particular area of visual space. The comparison of interest is the neural responses when the monkey is attending to the RF of the recorded neurons vs. when the monkey is attending elsewhere. Within this task, individual neurons show vastly different response profiles even though the monkey does not make any eye movements. As a population, the spike rates of these neurons encode whether the monkey is paying attention to a particular area in visual space throughout the duration of each trial.

Fig. 3 shows an example screenshot with two single trials, two condition averages, and one average difference plot comparing neural responses when the monkey is cued to attend the RF vs. cued to attend away. Neurons from independent recordings were combined into example trials and aligned so that their RFs are located at what is schematically diagrammed as the lower left corner of the screen. The trials span four seconds. Vertical lines on the rasters indicate different epochs of the trial, and shading on single trials marks the periods during which a visual stimulus other than the fixation spot is presented.The neurons are sorted such that the more visually responsive neurons are at the top of each raster.

The corresponding video capture, which demonstrates the

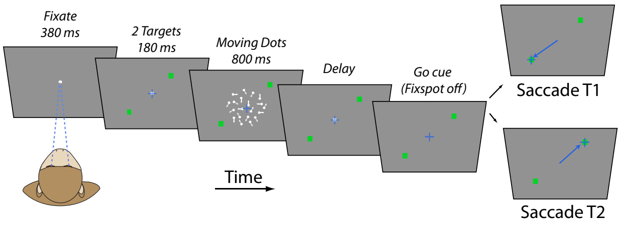


Figure 4: Sensory-guided decision and eye movement task. Monkeys were trained to make perceptual judgments about the average motion of a moving dot pattern and then generate an eye movement towards a corresponding target.

three different data to sound mappings, can be viewed at the link given at the end of the discussion section. A feature that stands out both in the rasters and the sonified output is the increased activity in response to visual stimuli presented in the response field. Furthermore, the neurons sustain an enhanced level of activity when the monkey is attending to the RF location even when the screen is blank without a visual stimulus to drive the cells. During the presentation of the gratings, the neurons also show enhanced visual responses to the grating in the RF when attended vs. not attended even though the visual stimulus is the same in both conditions. At the end of each trial, the level of neural activity quickly falls off.

4.2. Sensory-guided decision and eye movement task

A separate experiment explored the role of neurons from a similar area of the brain during a task that involved perceptual judgments and planning of eye movements (Fig. 4) [10]. In this task, the monkey is shown two targets in the periphery, and a random moving dot pattern appears in the center of the screen for 800 ms. The monkey must determine the direction of motion of the dots and later report its decision in the form of an eye movement to one of the two targets. On different trials, the strength of the motion signal towards one target or the other is varied from 0 to 40%. Once the fixation spot turns off, the monkey can move its eyes to the chosen target. Neurons were recorded from prearcuate cortex, a region of the brain near (and potentially overlapping with) the FEF, using an electrode array.

Fig. 5 shows two average spike rate responses when the monkey is shown a 40% coherence dot pattern and chooses target one (left) vs. target two (right). Although the RFs of individual neurons were not determined prior to the experiment, the neurons generally respond more strongly to what is schematically diagrammed as the left side of visual space. The average spike rate rasters are followed by three average difference rasters, showing the average spike rate when the monkey chooses target two (on the right) subtracted from the average spike rate when the monkey chooses target one (on the left) for 0%, 10%, and 40% motion coherence trials. Each trial shown is event-aligned and spans two seconds, where the first half is centered around the 800 ms presentation of moving dots, and the last half is centered around the time of the saccade. Trials are grouped according to the motion signal strength and the monkey's target choice. The neurons are sorted by motion selectivity during the presentation of the moving dots such that neurons that show the strongest difference in activity between the two directions of motion presented are placed at the top of each raster.

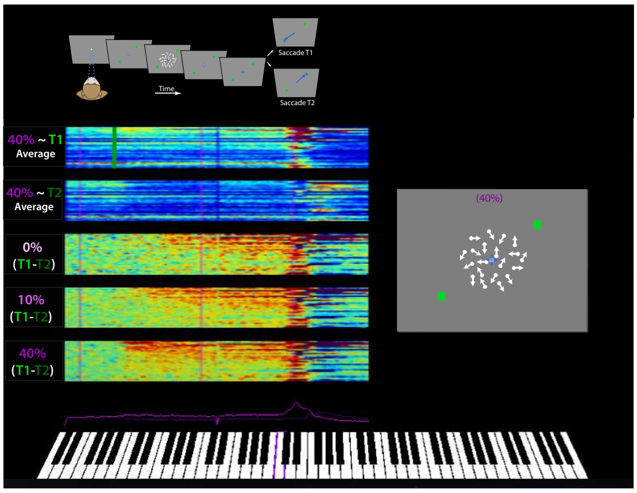


Figure 5: Example screenshot using data from the sensory-guided decision and eye movement task.

The greater the motion coherence of dots, the more information is available to the monkey (and neurons) for deciding on and planning the upcoming eye movement, and this is apparent in both the rasters and sonifed output (see the link at the end of the discussion section for the corresponding video capture). In the average spike rate trials, activity builds up during the dot presentation period as the monkey decides on T1 (in or near the RFs of the neurons on the left side of the task schematic screen) and prepares to move its eyes there. Conversely, activity becomes suppressed as the monkey decides on T2 (away from the neurons' RFs). The diverging response profiles are further illustrated in the difference plots. The neurons show earlier and stronger difference signals when the monkey is presented with increasing motion signals.

5. DISCUSSION

The current system provides a preliminary framework for exploring both data analysis and the creation of biologically inspired musical elements using an integrated audio and visual interface. In the examples provided, the visual display of data rasters and rate traces in isolation already effectively convey spike timing and spike rate information. The addition of sonified output and a dynamically updating task schematic changes the user's interaction with the data such that instead of viewing a static image, the user can step through an experimental trial. While the aim is not necessarily to discover information in the auditory displays that cannot be perceived in the visualizations, the system provides an engaging and accessible means to explore neural data and extract the main effects in each experiment. The sonified output enhances and complements the visualization, providing a multi-sensory means of experiencing and exploring the data.

To date, three types of data to sound mappings have been implemented, and in each case the mapping has been to musical pitch. In order to ensure a relatively constant level of musical consonance, the pitch mappings were scaled to highly consonant pitch collections such as the pentatonic scale. Scaling and filtering pitch mappings, while maintaining a pleasing sonic environment, limits the resolution of the sonified data. Future work will explore how other auditory dimensions such as timbre, spatial location, rhythm, and volume may provide alternative expressive representations of neural activity. Exploration of the space of possible mappings could be directed towards extracting information from the data in an intuitively perceptible manner or towards purely aesthetic goals.

One challenge in sonifying and visualizing neural data for information content is balancing the tradeoff between accurately representing the raw signal and producing a meaningful interpretation of the information contained within the signal. The amount of information contained in a single spike, for example, remains an open question, but as observers we may not be able to easily discern the magnitude of its impact in the context of brain dynamics. Spike rates provide a good estimate of the relative activity level at a given point in time but do not take into account the possible role of temporal patterns in the data. Other potential parameters to extract from the data could include measures of synchrony, trial-by-trial variability, or correlations between neurons. The evolution of neural population activity over the course of individual trials could also be transformed into dimensionality-reduced trajectory representations and sonified to highlight behavior of the network as a whole.

Within each experiment, the network is loosely defined since current technology allows experimenters to sample only a subset of neurons, and it is not necessarily straightforward to determine the number of neurons sufficient to represent the population. Sorting neurons based on properties of the individual neurons (such as visual responsiveness, as used above) imposes particular constraints that may or may not be explicitly utilized in the brain but could help the user in functionally grouping the neurons when viewing and listening to the activity of the population.

Furthermore, brain dynamics on a single trial are rapid and complex in comparison to our ability to perceive and process visual and auditory signals. Depending on the goal of the user, it may be ideal to observe each detail or quickly gauge the neural population response. The ability to listen to neural population data in realtime would potentially be useful during the course of an experiment to monitor neural activity on a trial-by-trial basis and evaluate the quality of the data as it is being collected. A more interactive interface would also allow the user to choose data to sound mappings best suited to feature different aspects of the data.

Sonification and visualization offer tools to relate the activity of neural populations to cognitive tasks by decoding neural signals into features that can be grasped perceptually. As data sets become increasingly complex with more neurons and more simultaneously recorded brain areas, new methods for extracting information from the high-dimensional data may play an instrumental role in conveying insights about how the brain works to neuroscientists and non-neuroscientists alike.

Video captures of the examples described above are available online at: *https://ccrma.stanford.edu/∼mindyc/sovnd/demo/*

6. ACKNOWLEDGMENT

Special thanks to Ge Wang, Katherine Armstrong, Tirin Moore, Rachel Kalmar, John Reppas, Bill Newsome, and Jonathan Berger. This project was developed during the Autumn quarter in 2009 at CCRMA (Stanford University Center for Computer Research in Music and Acoustics) as part of a course taught by Ge Wang entitled "Music, Computing, and Design I: Software Design and Implementation for Computer Music." Data for the attention task was collected with Katherine Armstrong in the lab of Tirin Moore. Data for the decision task was formatted by Rachel Kalmar and collected by John Reppas in the lab of Bill Newsome. All experimental procedures were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals. This work was supported by the National Research Service Award Grant F31NS062615.

7. REFERENCES

[1] T. Hermann, P. Meinicke, H. Bekel, H. Ritter, H. Muller, and S. Weiss, "Sonification for eeg data analysis," in *Proc. International Conference on Auditory Display (ICAD)*, 2002.

[2] G. Baier, T. Hermann, and U. Stephani, "Event-based sonification of eeg rhythms in real time," *Clinical Neurophysiology*, vol. 118, no. 6, June 2007.

[3] T. Rutkowski, F. Vialatte, A. Cichocki, D. Mandic, and A. Barros, "Auditory feedback for brain computer interface management - an eeg data sonification approach," *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 4253, 2006.

[4] E. Miranda, L. Bull, F. Gueguen, and I. Uroukov, "Computer music meets unconventional computing: towards sound synthesis with in vitro neuronal networks," *Computer Music Journal*, vol. 33, no. 1, Spring 2009.

[5] G. Weinberg and T. Thatcher, "Interactive sonification of neural activity," in *Proc. International Conference on New Interfaces for Musical Expression (NIME06)*, 2006.

[6] http://www.music.mcgill.ca/∼gary/rtaudio/.

[7] http://www.audiomulch.com/∼rossb/code/oscpack/.

[8] G. Wang and P. Cook, "Chuck: a concurrent, on-the-fly audio programming language," in *Proc. ICMC*, 2003.

[9] K. Armstrong, M. Chang, and T. Moore, "Selection and maintenance of spatial information by frontal eye field neurons," *J. Neurosci.*, vol. 29, no. 50, pp. 15 621–15 629, December 2009.

[10] R. Kalmar, J. Reppas, S. Ryu, K. Shenoy, and W. Newsome, "Ensemble activity underlying movement preparation in prearcuate cortex," in *Frontiers in Systems Neuroscience*. Poster at: Computational and systems neuroscience, 2010.

PARAMETER MAPPING SONIC ARTICULATION AND THE PERCEIVING BODY

David Worrall

School of Music, College of Arts and Social Sciences
Australian National University
Canberra, ACT 0200
worrall@avatar.com.au

ABSTRACT

In data sonification research, there is a well-known perceptual problem that arises when abstract multivariate datasets of a certain size and complexity are parametrically mapped into sound. In listening to such sonifications, when a feature appears, it is sometimes difficult to ascertain whether that feature is actually a feature of the dataset or just a resultant of the psychoacoustic interaction between co-dependent parametric dimensions. A similar effect occurs in visualisation, such as when parallel lines can appear more or less curved on different backgrounds. Couched in psycho-philosophical terms, we can ask whether this failure is related to classical phenomenology's inability to produce an eidetic science of essential invariant forms that involve no assertion of actual material existence, or to there not yet having been found some generalisably acceptable limits from heuristically tested mappings. This paper discusses the nature of this problem and introduces a sonification research project based on embodied, non-representational phenomenal models of perception.

1. PARAMETRIC MAPPING SONIFICATION (PMS)

The use of discrete sounds for auditory alerts and alarms presents sound designers primarily with differentiation problems: both between the sounds themselves and between the sounds and the background environment in which they function. Though related in subtle ways, these discrete auditory displays do not address another problem: how to acoustically represent data *relations* for interpretation by listeners, for the purpose of increasing their knowledge of the source from which the data was acquired. That task can be recast as one of how to use sound to create mental ‘objects’ for active contemplation, as distinct from how to correctly elicit a timely response to already well-differentiated auditory stimuli.

The term 'data sonification' is usually reserved for a collection of techniques for exploring datasets that have an equally–spaced metric in at least one dimension and in which there are sufficient data points to afford continuous aural interpolation between them[1]. Such dataset representations are most commonly used to learn more about the systems that produced them. Applications range from monitoring the real-time operation of machines, capital–market trading, geographic and demographic features, weather and the environment and so on; as tools to assist in the discovery of features and new regularities, and to assisting those with visual impairment to gain access to large quantities of information normally presented graphically.

Parameter mapping is the most widely used sonification technique for representing multi-dimensional data as sound. Parameter mapping sonifications (PMSs) are sometimes referred to as sonic scatter plots [2][3] or nth–order parameter mappings [4]. Typically, data dimensions are mapped to sound parameters: either to physical (frequency, amplitude), psychophysical (pitch, loudness) or perceptually coherent complexes (timbre, rhythm). PMSs can have both analogical and symbolic components. Analogic variations in the sound can result when mapping from a large data domain into a small perceptual range or when data is specifically mapped to acoustic modifiers such as frequency or amplitude modulators. PMS is sometimes referred to as *multivariate data mapping,* in which multiple variables are mapped to a single sound. Scaletti describes one way of implementing it by “mapping of each component of a multidimensional data point to a coefficient of a polynomial and then using that polynomial as the transfer function for a sinusoidal input” [4]. Within an overall analogic mapping, symbolic representations such as auditory beacons [5] can be used to highlight features such as new maxima and minima, or absolute reference points in a sonification such as ticks to indicate the regular passing of time.

2. FOR *SONICULATION* OR MUSICAL EXPRESSION?

It is useful to distinguish data sonifications made for the purposes of facilitating communication or interpretation of relational information in the data, and data-driven music composition, ambient soundscapes and the like—the primary purpose of which is the expression of musical knowledge and broader cultural considerations, whatever they may be. The current use of the term “sonification” to include such cultural concerns is unfortunate because it blurs purposeful distinctions, yet today, the the older expression “scientific sonification” seems unnecessarily restricted. So, for situations in which the distinction is considered important, the portmanteau term *soniculation* (from sonic + articulation) has been introduced to mean the representation of data with sound with the principal and overriding imperative of making the structural characteristics of the data as clear and explicit to a listener as possible—even at the expense of other aesthetic considerations, if necessary[1].

Needing to maintain this distinction is not to suggest that there are not commonalities. In fact, as discussed later in this paper, the two activities can provide insights that are mutually useful. What is important is to maintain a critical awareness that, because the purposes of the activities are different, so will their epistemological imperatives and consequences, such as, for example, in tool design and useability.

3. "THE MAPPING PROBLEM"

A contemporary overview of the current range of sonification and other auditory display techniques is available[1]. The technique discussed here, parametric mapping sonification (PMS) has a number of positive aspects, which Scaletti first outlined in some detail [4]. Many data dimensions can be listened to simultaneously. It is very flexible and the mappings can be easily changed, allowing different aural perspectives of the same data. In addition, acoustic production can be assigned to sophisticated tools originally developed for computer music synthesis. These are readily available and permit many quite sophisticated parameter mappings to be synthesised in real-time.

Frysinger provides a useful overview of the history of the technique[6], and Flowers highlights some of its pitfalls and possible future directions. An experienced multivariate data sonifier, he observed that while "the claim that submitting the entire contents of 'dense and complex' datasets to sonification will lead to the 'emergence' of critical relationships continues to be made, I have yet to see it 'work'" [3]. The main limitation of PMS is co-dependence, or lack orthogonality (linear independence) in the psychophysical parameter space. Linear changes in one domain produce non-linear auditory effects, and the range and variation of such effects can differ considerably with different parameters and synthesis techniques. These perceptual parameter interactions can produce auditory artifacts that obscure data relations and confuse the listener. Kramer suggests that, although a truly balanced multivariate auditory display may not be possible in practice, given powerful enough tools, it may be possible to heuristically test mappings to within acceptable limits for any given application [5].

In many discussions of data sonification, the distinction between *data* and *information* is often lost. In fact, the expression *data sonification* itself promotes an elision and in doing so, implicitly supports the idea that information can automatically "pop-out" of a sonification once an optimal parameter-mapping of the dataset is found. The purpose of this paper is to argue why this is unlikely (except perhaps for those who have had advanced aural training acquired over many years), and to argue that it is necessary to search for general solutions outside of explicitly representational paradigms.

4. THE AUDITORY OBJECT

The historical record of the study of perception clearly reveals the overwhelming dominance of arguments based on the visual appearance of spatial objects; sounds not being considered as objects in themselves but as secondary properties of spatial objects and not essential to their ontology [7].

In tracing the roots of this "visualism" in pre-Socratic Greek thought, Ihde concludes, citing Aristotle, that it is as old as our own cultural heritage: "Above all we value sight … because sight is the principle source of knowledge and reveals many differences between one object and another."[8]. So the dawn of modern science was essentially a silent one and yet-to-be-captured sound still quite mysterious. One of Descarte's undervalued attributes was his honesty [9]:

> *As to other things such as light, colours, sounds, scents, tastes, heat, cold and the other tactile qualities, they are thought by me with so much obscurity and confusion that I do not even know if they are true or false, i.e. whether the ideas which I form of these qualities are actually the ideas of real objects or not [or whether they only represent chimeras which cannot exist in fact].*

The idea that an aural event could be objectified and studied in its own right, that is independent of the means of its production, evolved slowly and in parallel with the development of the concept of a musical work as reproducible from notation [10] and eventually with the use of various sound recording devices invented during the nineteenth and twentieth centuries.

A distinction between the physical sounds (noumena) and aural events (phenomena) is not just a philosophical one. It is important that these two types of objects are not conflated as there is an enticement to do when the (software) tools designed to produce soundwaves are also used to produce abstract aural phenomena, that is, immanent objects. One difficulty that arises when tools from one task domain are appropriated to another is the implicit transfer of the epistemological assumptions of the former to the latter; an idea expressed in the saying "to a person with a hammer, everything looks like a nail". In the current context, this translates to the assumption that tools used to produce sounds for computer music are appropriate, or at least adequate, for producing data sonifications. In fact, the particular situation is even more convoluted as the tools designed to make computer *music* have themselves embedded an epistemology of *music* which privileges the production of the *sounds* of music over its other aspects, such as gesture and temporal evolution. This "timbre object fetish" in computer music can be understood as having an historical basis in the early relationship between computer music and artificial intelligence research, both of which have continued the doctrine of isolating *res cogitans* from *res extensa* and prejudiced the former over the latter (*cogito ergo sum* – "I think therefore I am").

5. UNDERSTANDING THE AURAL PHENOMENA

As perceptual phenomena, it is appropriate to make a distinction between those sonifications that result from the excitation of physical objects (or synthetic simulations that closely approximate them, such as homomorphic modulation and those based on physical modeling principles), and those, such as a parametrically mapped datasets, that are artifacts of perceptual processes in which elementally composed soundpoints are assembled in such as way that the psychophysical continuity of at least some of the parametric dimensions conflates the perception of those soundpoints into a single immanent object or perceptually coherent auditory scene.

The reason such a distinction is important is that physical objects obey physical laws that human beings have evolved to recognize the effect of with negligible attentional effort, whereas sound structures synthesised from numerical datasets may not. In the physical-modeling case, data is used to excite a "self-contained" resonator (an integrated unity obeying physical laws)[1], or perhaps less convincingly, the data itself is used to construct a physical model that is "resonated" by a listener interacting with it [11][1]. In the case of objects synthesised from numerical datasets, the elementally composed soundpoints are 'presented' to listeners in ways that it is hoped afford their perception of the form of the dataset. Perception is thus understood as human behaviour and soniculated PMSs as sound constructions 'imprinted' in multi-dimensional psychophysical space to elicit a perceptual behaviour which affords the cognition of the form of that structure as an (auditory) object in the listener.

While, in his ground-breaking overview, Bregman described the basic elements and dimensions of analytic and synthetic listening in terms of auditory stream integration and segmentation [12], the current PMS model doesn't work very well and there is yet to be written a generalized exposition appropriate for many sonification tasks: how to synthesise perceptual cohesion while maintaining aurally differentiable soundpoints. It remains a task of soniculation research to develop robust models of listener's perceptual behaviour that can be reliably reverse-engineered to produce affordances that solicit listener's to behave in ways that assist them to get enough 'grip' on these sound structures for them to be perceived as cohesive auditory objects.

Exactly how these perceptual mechanisms work is open to speculation and investigation. In-keeping with the Cartesian tradition, there have been two kinds of investigation, which we label 'mental' and 'empirical'. We review these two approaches before offering a critical discussion that leads to a proposal for a different approach.

5.1. The mentalists

According to Kant's understanding, what exactly is meant by *information* is embedded in relationships between the sensation, perception and apperception of phenomena; what he called *appearances (Erscheinungen):* things as they are for humans, as opposed to things as they are 'in–or–of–themselves' (*Ding an sich)* otherwise known as *noumena*. From this perspective, information can be simply characterized as phenomena, or thoughts about phenomena in the mind of some person [13].

Following Kant and the Idealists, Brentano and his students Meinong and Husserl investigated the perceptual world as a rational or mental construction of a perceiving subject. Their phenomenological method (a contemplative and descriptive psychology as distinct from the newly developing natural or empirical psychology) entailed "bracketing off" (with an attitude Husserl called *époché*) phenomena ('things-as-we-know-them') from the physical world (Kant's 'things–as–they–really–are'), in and attempt to discover the underlying structures and forms of the objects produced by intentional mental processes; firstly in the mind of the perceiver and secondly as sharable with others—a characteristic Husserl called *intersubjectivity.* Brentano expressed it like this [14]:

> *Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity.*

Brentano's goal was to outline the criteria for distinguishing mental and physical phenomena. He used the terms *mental or intentional inexistence* to refer to what today is sometimes understood as a characteristic of consciousness: the mind's capacity to *refe*r or be *directed* to objects that exist solely in the mind. Meinong's concern was with the intentional *relation* between the mental act and an object. He maintained that such a relation existed even when the object external to the mental act towards which it is directed doesn't exist, such as Pinocchio, Orpheus, Unicorns and the Fountain of Youth. Earlier, Hume had considered the concept of non-existent objects contradictory, Kant and Frege considered it logically ill-formed though later, Russell adopted the idea [15].

In cases of temporally extended objects ('events') like melodies, Brentano argued such objects towards which we are directed do not immediately vanish from consciousness once the mental act is over. They rather remain present in altered form, modified from *present* to *past*. Every mental phenomenon triggers an 'original association' (*proteraesthesis)*, a kind of memory which is not a full-fledged act of remembering, but rather a part of the act that keeps lively what was experienced a moment ago. When I listen to a melody, for example, I first hear the first tone. In the next moment I hear the second tone, but am still directed towards the first one, which is modified, though, as past. When I hear the third tone, the second tone is modified as past and the first is pushed back even further into it.

Though Peirce thought that there can be no perceptual objects without a unifying factor that distinguishes them from the 'play of impressions', Husserl's aim was to develop an *eidetic* science; one of essential, invariant phenomenal forms that involves no assertion of actual material existence. However, he struggled to keep them conceptually separate from Plato's Ideas.

In the 1960s, following the lead of Descarte, Kant and Frege, and a misapplication of Shannon's information theory to *meaning* [16], Minsky lead a team at Massachusetts Institute of Technology aimed at modeling intelligent behaviour artificially, using symbolic representation and the predicate calculus. Their atomistic approach was abandoned after its failure to represent the background knowledge and specific forms of human "information processing" which are based on the human way of being in the world. This way of "being-in" turned out to be syntactically and thus computationally unrepresentable using currently conceivable techniques [17] [18].

5.2. The empiricists

In ecological terms, the objects and the environment in which they reside, afford listener exploration. For Gibson[19], following Gestaltists such as Wertheimer, Koffka, Kohler and Mach, these affordances were thought to be *in* the physical objects and our observation of them consists of us (somehow) forming a representation of them in our heads; the Gestaltist's task being to empirically search for the means by which the brain more–or–less unconsciously perceives the forms of objects received from sense data.

The Gestaltists discovered perceptual invariants such as the figure–ground phenomena, which they considered as arising directly from the physical nature of sensations derived from the noumenal world. However, they were not able to extend the idea past sensations. It is a characteristic of such phenomenal forms that their properties can remain unchanged when the objective stimuli upon which they rest undergo certain modifications. This phenomenon of identity is part of a much more general issue in topology and mathematical group theory; of invariances with respect to transformations of the primitive elements out of which a form is constructed. The mathematical concept of transformability corresponds to the concept of transposability in perception. So by accepting "form" as a primitive concept, Gestalt psychology made an attempt to free psychological theory from contingency on the mere mosaic of perceptions.

Not all group-theoretic transformations of perceptual objects are equally cognized, nor are the same transformations as easily perceivable in different sense modalities. For example, symmetry group transformations of pitch and temporal structures, such as transposition, inversion and retrogradation, occur frequently in music though they seem not to be all equally evident to the casual listener: under non-extreme pitch transposition and tempo acceleration a melodic structure remains strongly invariant; pitch contour inversion and rhythmic retrogradation are common occurrences in some musics but are not as strongly invariant, while rhythmic inversion seems not to be perceptually invariant or even generally defined.

6. SUB-CORTICAL NEURAL ACTIVITY

Whilst a considerable amount is known about the structure and functions of individual neurons, the fundamentals of how macro effects emerge from populations of neurons are still largely unknown, despite considerable effort over the last decades. As the field develops, there is a growing realisation that the phenomena associated with "consciousness", "nonconsciousness" and "cognition" are too diverse to continue to be meaningfully subsumed under the same ill-defined terms [20]. For example, given the verifiable presence of nonconscious antecedents to an intention [21], it is unclear how formed our decisions are when we become aware and think of ourselves as mentally ‘‘creating’’ them.

6.1. Neural correlates of consciousness

The search for the neural correlates of consciousness has been aided by the ease of Functional Magnetic Resonance Imaging (fMRI) of cortical activity. However, it is suggested by Churchland and others [22][23] that the ready availability of such technologies has contributed to a cortical "chauvinism" that tends to concentrate on conscious perception at the neglect of the role they have in servicing behaviour. Specifically that, in service of keeping the body alive, the nervous systems of animals, as *movers*, function to service planning, deciding and executing these plans *in movement*. Importantly, much of the brain's input is consequent upon the dynamical feedback loop between observed phenomena and an organism's own movements, exploratory and otherwise. This loop extracts vastly more information about the causal properties of the external world in a given time interval, leading to greater predictive prowess, that is, skills regarding the causal structure of the world, than could a purely passive system.

Time is an essential component of causal knowledge, and predicting durations, interception intervals, velocities, and speeds of various body movements is critical to an animal's survival. Efference copy (being aware that a movement is one's own and not the world's) is also thought to be critical, as perhaps is the nonconscious "analysis" and memory of the movement of other movers, such as in predator–prey/pursue–evade relationships, for example. In contradistinction to the conventional wisdom that ‘‘the sensory pathways are purely sensory", according to the Guillery and Sherman hypothesis, messages to the thalamus and cortex also carry information about ongoing instructions to the organism's motor structures [24]. Consequently, as a developing organism begins to interact with the world, sensory signals also "carry" gestural predictions: as an animal learns the consequences of a particular movement, it learns about what in the world will probably happen next, and hence what it might do after that.

Damasio's studies of efference copying of one's own thoughts and empathy with others provide even more evidence for this thesis that perception, learning and memory are not just cerebral processes but are embodily integrated into an organism as, what Polanyi called, tacit knowledge [25][22].

6.2. Mirror neurons

Kohler et al.'s finding, not only that that certain neurons in the ventral premotor area will fire when a monkey performs a single, highly specific action with its hand: pulling, pushing, tugging, grasping, picking up and putting a peanut in the mouth etc., but that that "mirror neurons" will also fire when the monkey in question observes another monkey (or even the experimenter) performing the same action, offers some neurological basis for a theory of cultural inheritance, "mind reading" empathy, imitation learning, and even the evolution of language [26]. As Churchland observes, by shifting perspective from "visuocentricity" to "motor–sensory-centricity," the singular importance of temporality takes center stage in an hypothesis that "time management," for want of a better term, is the key to the complexity of tasks of thalamic nuclei, and very probably also to a range of conscious phenomena [20].

More recent studies have demonstrated that a mirror neuron system devoted to hand, mouth and foot actions, is also present in humans. Buccino, Solodkin, and Small review this literature and that of the experimental evidence on the role of the mirror neuron system in action understanding, imitation learning of novel complex actions, and internal rehearsal (motor imagery) of actions [27]. The finding that actions may also be recognised from their typical sound, when presented acoustically has important implications for embodied soniculation research. Besides visual properties, it was found that about 15% of mirror neurons, called *audio-visual mirror neurons*, also respond to the specific sound of actions performed by other individuals even if only heard [26]. It has been argued that these neurons code the action content, which may be triggered either visually or acoustically. Phillips-Silver and Trainor demonstrated an early cross–modal interaction between body movement and auditory encoding of musical rhythm in infants [28]. They found that it is primarily the way adults move their bodies to music, not visual observation, that critically influences their perception of a rhythmic structure. Their results suggest that while the mere visual observation of a conspecific's goal-directed movement (e.g., reaching for an object or hand–to–mouth action) is sufficient to elicit a neuronal representation of the action, this does not transfer to the domain of metrical disambiguation [29]. So it appears that either this type of rhythmical body movement is not an example of the kind of object-directed action that activates the mirror neuron system or the information provided by the mirror neurons is not strong enough to influence the later-recalled auditory metrical representation of a rhythmic pattern.

In an experimental study of gestures, subjects of various ages were able, with a high degree of accuracy, on only hearing different individual human's walking and running on various kinds of surfaces, to determine their sex [30]. A consequential inference is that differences in ambulatory action, presumably resulting from relatively small differences in skeletal anatomy, is tacitly 'available' to listeners. Also consequent to these findings is the need for better models of multi-modal sensory input, particularly with respect to the integrative functions of vestibulation and proprioception, which some empirical evidence suggests are available to listeners though aural means alone [31][30].

7. CRITICAL DISCUSSION

While knowledge of the structure and functions of individual and clusters of neurons is increasing, there are billions of them, each with tens-of-thousands of connections so there is no certainty, even when the overall functioning of the neural system is significantly better understood that it currently is, that such an understanding will be able to adequately account for the ability to synthesise perceptual objects. In fact, if the *rate* at which pulses are transmitted turns out to be the minimum unit in an account of the relevant activity of the nervous system [32] and the diameter of an axon, which might be a function of the recency of a signal passing down it, plays a crucial role in processing information by acting as a filter [33], there is no reason to believe that information processing at neurological-level can ever be formally described [17].

The mentalist approach has failed to find any means by which mental representations can be reliably accumulated for conscious reflection, at least not without a good deal of training and effort. This somewhat explains some of the difficulties reported in PMS research that the vast majority of ordinary listeners, for whom a low conceptual loading is necessary for continued engagement, are precluded from making 'sense' of them just as the attempt to use computers to develop an 'artificial intelligence' based on computational theories of mind that rely on a classical reductionist approaches such as "mind is to software as brain is to hardware" failed to be able to understand even the simplest stories because of the unrepresentability of the background knowledge and the specific forms of human "information processing" which are based on their way of being in the world. This suggests that, when compared with the ease with which everyday sounds are identified; the ease with which a myriad of melodies are learned, remembered and identified, that the mentalist approach is inadequate at best. More likely, that it is just wrong.

The relatively recent availability of tools to abstract sound from its origin in the physical action of objects, and the development, alongside that of "good-old-fashioned-AI (GOFAI)[34] of seminal software for computer music [35], has blurred the functional distinction between sound and music, much as often occurs between data and information, and information and meaning. Sound recording enabled Schaeffer, building on the philosophical foundation of Husserlian (that is mentalist, *époché*) phenomenology, to propose a musical analysis based on *reduced listening*, that is listening to sounds for their own sake, as sound objects, by removing their real or supposed sources and meanings [36]. It is of particular interest in the light of the previous discussion of the role of time and causality in perception, that while Schaeffer does discuss tempo and temporality, he makes almost no reference to pulse and rhythm.

8. THE PERCEIVING BODY

Husserl's pupil Heidegger was critical of the subject/object split that pervades the Western tradition and that is in evidence in the root structure of Husserl and Brentano's concept of intentionality, that is, that all consciousness is consciousness of something, and (the idealist notion that) there are no objects without some consciousness beholding or being involved with them. Heidegger encompassed terms such as "subject", "object", "consciousness" and "world" into the concept of a mode of "being-in-the-world" as distinct from an essentially Positivist "knowing" of objects in the universe that is required for navigating the environment–measurement, size, weight, shape, cause & effect etc. His Being-in-the-world is characterized as "ready-to-hand"[37]:

> *. . the kind of dealing which is closest to us, not a bare perceptual cognition, but rather that kind of concern which manipulates things and puts them to use; and this has its own kind of 'knowledge.'*

In other words, participatory, first-hand experience: familiarity, tacit know-how, skill, expertise, affordance, adaptability etc. Heidegger argues that our scientific theorizing of the world is secondary and derivative and he exposes an ontology that is far broader than the dualistic Cartesian framework. He stresses the primacy of the readiness-to-hand, with its own kind of knowing or relating to the world in terms of what matters to us. It follows, from Heidegger's perspective, that human action is embodied, that human knowing is enactive, and participatory.

The Hungarian scientist and philosopher, Polanyi proposes a type of participative realism in which personal knowledge plays a vital and inescapable role in all scientific research, indeed, in all human knowing [38]:

> *Let us therefore do something quite radical … let us incorporate into our conception of scientific knowledge the part which we ourselves necessarily contribute in shaping such knowledge.*

By stressing the tacit nature of participatory knowing, Polanyi claimed that “we know more than we can tell”. In this way he emphasised knowledge that is implicit to tasks, situations and attitudes. He used the term *tacit knowledge* to refer to those things we can do without being able to explain how, that is, in the absence of explicit rules or calculative procedures. The "indwelling" nature of tacit knowledge is important in the development of the skill of reflexivity, such as needed in the sifting through and interpretation of qualitative data.

Heavily influenced by both Husserl and Heidegger, Merleau-Ponty produced a much more developed understanding of the body and its role in non-conceptual perception [39][40]. As the only major phenomenologist of the first half of the twentieth century to engage extensively with the sciences, he was able to systematically demonstrate the inability of the mentalist and empiricist explanations to adequately account for observed phenomena. In doing so, he produced a theory of perception in which the body and the world are entwined; in which perception occurs through the “intentional tissue” of the “body schema” (*schéma corporel*); much as epigenetic alterations occur in a phenotype by the osmotic transduction of molecules through semipermeable membranes.

Todes builds on Merleau-Ponty's work by beginning to work out a detailed phenomenological account of how our embodied, nonconceptual perceptual and coping skills open a world to us. He then works out twelve perceptual categories that correspond to Kant’s conceptual categories, and suggests how the nonconceptual coping categories can be transformed into conceptual ones [41].

9. SO WHERE IS THE BODY IN SONICULATION?

The reduction of music to noises–in–the–head is supported by the wider cultural practice of using visual terminology to describe aural phenomena. Such a privileging of the visual over the aural too easily promotes the unwarranted prejudicial masking of one dimension over another. The implication of the privileging of visual experience, especially when it is conceived principally to be by stationary beings of stationary *ob*jects that are *ob*served[1] and only then perhaps with movement, is to, albeit subtly, privilege the spatial over the temporal; sound 'objects' over gestural dynamics. Applied to soniculation, such an epistemology weakens the strongest ontological scaffolding that supports temporal perception as primary means by which information can be transduced through sound to the perceiving body.

A new movement-encompassing action-based approach to the relationship between sound and sensibility began in the 1980s [42]. Methodologies include the use of abductive as well as inductive inference, and are contributing to new perspectives on how to approach the relationship between sensibilities [31] [43]. In some ways this can be seen as a return to the Aristotelian integration of sound and sensibility through *mimesis* and related to the Kantian problems of openness and *endness* in the containment of beauty in formal structures and the empathic relationship within them through movement and action [13].

It seems reasonable to suggest that continuing to flip-flop between the mental-empirical antinomy, in anticipation that one or the other will eventually provide an applicable model for PMS, may be a forlorn hope. There seems little point in speculating whether or not listening will ever regain the relative importance to humans it enjoyed prior to the European En*light*enment, but there are signs of a growing recognition that the resolution of the mind/body dilemma will not be solved by dispensing with the body.

In many ways, the tradition of emphasising disembodied cognition over alternative approaches has never really been totally applicable to musical sensibility. Clearly humans have the capacity to create, transmit, receive, transform and most importantly recall certain types of immanent objects using sound: music can afford them all! The idea that musical involvement is based on the embodiment of movement and the bodily sensing of music, has a long history, of which the traditional connection between dance and music is but a gross example. Truslit studied the body movements of musical performers and suggested they were articulations of inner movements in the music itself [44]. Central in Truslit’s approach to musical movement are the notions of dynamics (intensity) and agogics (duration). If the music has the dynamo-agogic development corresponding to a natural movement, it will evoke the impression of this movement. He makes a distinction between rhythmic movement and the inner movement of the music. In contrast to rhythmic movement, which is related to individual parts of the body, the inner movement forms the melody via the vestibular labyrinth of the inner ear and is related to the human body as a whole. Both Nettheim [45] and Clynes [46] also make a connection between music and gravitational movement, based on the idea of a dynamic rhythmic flow beyond the musical surface.

Empirical musicology, including the mensural study of performance practices, together with neurophysical analysis of 'embodied' instrumental performance, is becoming recognised as at least as important for understanding musical ideas as notated structural abstractions [47][48]. There is growing interest in human/machine interfaces that enable musicians to produce computer-generated sounds under nuanced gestural control [49][50][51].

Both empirical musicology and gesturally-controlled computer-music performance are of relevance to this investigation. However the former, is deficient in being largely analytical and the latter, in being little interested in empirical evaluation. Between the two, it may be possible for future soniculation researchers to recognise the limitations of the current PMS paradigm, to accept that musical information can be intelligible, that is capable of being *soundly understood*[2], through temporally-encoded “second-order” structures and undertake research to ascertain the viability of a variety of embodiment models under controlled conditions. The beginning described in the next section is but one example of an empirically approach which may or may not prove fruitful.

[1] From L. *ob*-"over"+ *servare* "to watch, keep safe" and *ob*-"against"+ *jacere* "to throw,” as in a jet [49].

[2] O.E. *understandan* "comprehend, grasp the idea of," probably lit. "stand in the midst of," from under + standan "to stand". O.E. under, from **nter*- "between, among" (cf. Skt. *antar* "among, between," L. *inter* "between, among," Gk. *entera* "intestines"[52].

10. TOWARDS A GESTURE-ENCODED SOUND MODEL

A programme of research has begun that seeks to empirically demonstrate whether or not the perceptual access to the structural and informational content of multivariate datasets through sonification based on a model that incorporates the aural transduction of known temporal embodiment affordances such as human gestures, is superior to one based on elementally composed aural *ob*jects that are *ob*served and rationally conceptualised. Philosophically, the is an approach based on an embodied phenomenology of perception first enunciated by Merleau-Ponty [39] and extended by Todes [41].

An extensive search of the literature has not revealed any other approach that addresses the issue of how to use the innate structures of the human body, expressed through gesture and transmitted aurally, to improve the "eyes-free, hands-free" tacit grasping of ideas and information contained in the increasingly large and complex datasets that are becoming a part of our daily lives—from climate and the weather to fluctuations in the financial markets and traffic flow. The research we are currently undertaking is to develop a model of (human) physical and sonic gesture correlates. The task is essentially to apply captured biomechanical data with sound-derived components (timing, spectral morphology etc) and known psychophysical principles as inputs to an iteratively trained Dynamic Bayesian Network (DBN). This Gesture-Encoded Sound Model will then be used to produce an active filter for transducing multivariate datasets to sound synthesis and control parameters. The approach renders a datastream to sound not only using observable quantities (inverse transforms of known psychoacoustic principles), but latent variables of a DBN trained with gestures of the physical body movements of performing musicians and hypotheses concerning other observable quantities of their coincident acoustic spectra. The research on the model will be integrated as an extension to *SoniPy* [53].

11. REFERENCES

[1] D.R. Worrall, “An introduction to data sonification,” in R. T. Dean (ed.), *The Oxford Handbook of Computer Music and Digital Sound Culture*, Oxford: Oxford University Press, 2009.

[2] J.H. Flowers, D.C. Buhman and K.D. Turnage, “Cross-modal equivalence of visual and auditory scatterplots for exploring bivariate data samples,” in *Human Factors*, Volume 39, 1997, pp. 341-351.

[3] J.H. Flowers, “Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions,” in *Proceedings of the First Symposium on Auditory Graphs*, Limerick, Ireland, July 10, 2005.

[4] C. Scaletti, “Sound synthesis algorithms for auditory data representation,” in G. Kramer (ed.), *Auditory display: Sonification, Audification, and Auditory Interfaces*. Santa Fe Institute Studies in the Sciences of Complexity, Proceedings, Volume XVIII. Reading, MA: Addison Wesley Publishing Company, 1994, pp. 223-251.

[5] G. Kramer, “Some organizing principles for representing data with sound,” in G. Kramer (ed.), *Auditory display: Sonification, Audification, and Auditory Interfaces,* Santa Fe Institute Studies in the Sciences of Complexity, Proceedings, Volume XVIII, Reading, MA: Addison Wesley Publishing Company, 1994, pp. 185-221.

[6] S.P. Frysinger, “A brief history of auditory data representation to the 1980s,” in *Proceedings of the First Symposium on Auditory Graphs*, Limerick, Ireland, July 10, 2005.

[7] J. Locke, *An essay concerning humane understanding,* 2nd edition, P. H. Nidditch (ed.), Oxford, 1690/1975.

[8] D. Idhe. *Listening and Voice:Phenomenologies of sound*. 2nd edition. Albany: SUNY Press, 2007.

[9] R. Descartes, *Meditations*. J. Veitch (trans.). 1641/1901, Part V. Accessed 13 July 2008 at http://www.wright.edu/cola/descartes/meditation2.html.

[10] L. Goehr, The Imaginary Museum of Musical Works. Oxford, UK: Oxford University Press, 1994.

[11] T. Hermann and H. Ritter. “Listen to your data: Model-based sonification for data analysis,” in G. E. Lasker (ed.), *Advances in intelligent computing and multimedia systems*, Baden-Baden, Germany, Int. Institute for Advanced Studies in System research and cybernetics, 1999, pp. 189-194.

[12] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: The MIT Press, 1994, pp. 395-453.

[13] I. Kant, *Critique of pure reason,* 2nd Edition. N.K. Smith (trans.) of *Kritik der reinen Vernunft*. London: Macmillan, 1787/1929.

[14] F. Brentano, “Descriptive psychology,” in *Brentano's lectures of 1890-1891*, London: Routledge, 1891/1995, p. 88.

[15] M. Reicher, “Nonexistent Objects,” in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy,* 2006. Accessed 2 February 2008 at http://plato.stanford.edu/archives/fall2008/entries/nonexistent-objects/.

[16] C. Shannon, and W. Weaver, *The Mathematical Theory of Communication*, Urbana, Ill: The University of Illinois Press, 1949, p.3 and p.99.

[17] H. Dreyfus, *What computers still can't do,* Cambridge, MA: MIT Press, 1992.

[18] D.J. Chalmers, *The Conscious Mind: In search of a fundamental theory,* New York: Oxford University Press, 1996.

[19] J.J. Gibson, *The ecological approach to visual perception.* Boston, MA: Houghton Mifflin Company, 1979.

[20] P.S. Churchland, “A neurophilosophical slant on consciousness research,” in V.A. Casagrande, R. Guillery, and S. Sherman, eds. *Cortical function: a view from the thalamus*. *Progress in Brain Research* , Volume 149, Amsterdam: Elsevier, 2005.

[21] B. Libet, “Unconscious cerebral initiative and the role of conscious will in voluntary action”, in *Behavioral and Brain Sciences*, 8, 1985, pp. 529–566.

[22] A. Damasio, *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain,* Orlando Fl. Harcourt, 2003.

[23] R.R. Llinas, *I of the vortex: From neurons to self,* MIT Press, MA: Cambridge, 2001.

[24] S.M. Sherman and R.W. Guillery. *Exploring the Thalamus,* San Diego, CA: Academic Press, 2001.

[25] A.R. Damasio, *The Feeling of What Happens,* NY: Harcourt Brace, 1999.

[26] E.C. Kohler, M.A. Keysers, U. L. Fogassi, V. Gallese and G. Rizzolatti. “Hearing sounds, understanding actions:

Action representation in mirror neurons," in *Science*, 297 (5582), 2002, pp. 846–848.

[27] G. Buccino, A. Solodkin and S.L. Small. "Functions of the Mirror Neuron System: Implications for Neurorehabilitation," in *Cognitive and Behavioral Neurology*, Volume 19, Number 1, 2006.

[28] J. Phillips-Silver, and L. J. Trainor, "Hearing what the body feels: Auditory encoding of rhythmic movement," in *Cognition 105*, 2007, pp. 533–546. Amsterdam: Elsevier.

[29] M. Wilson and G. Knoblich, "The case for motor involvement in perceiving conspecifics," in *Psychological Bulletin*, 131(3), 2005, pp. 460–473.

[30] R. Bresin and S. Dahl, "Experiments on gestures: walking, running, and hitting," in D. Rocchesso and F. Fontana (eds.), *The Sounding Object*, 2003. Accessed on 25 October 2008 at http://www.soundobject.org.

[31] F. Varela, E. Thompson and E. Rosch, *The Embodied Mind,* Cambridge, MA: The MIT Press, 1991.

[32] J. von Neumann, "Probablistic Logics and the Synthesis of Reliable Organisms from Unreliable Components," in *Collected Works*, A.H. Taub, (ed.), NY: Pergamon Press, Volume 5, 1963, p 372.

[33] W.A. Rosenblith, "On Cybernetics and the Human Brain," in *The American Scholar*, Spring 1966, p. 247.

[34] J. Haugeland, *Artificial Intelligence: The Very Idea,* Cambridge, MA: The MIT Press, 1985.

[35] M.V. Mathews, *The Technology of Computer Music,* Cambridge: Cambridge, MA: The MIT Press, 1969.

[36] P. Schaeffer, *Traité des objets musicaux: Essai interdisciplines*. Paris: Editions du Seuil, 1966.

[37] M. Heidegger, *Being and Time,* J. Macquarie (trans.), of *Sein und Zeit,* Oxford: Blackwell, 1927/1962, p.95.

[38] M. Polanyi, *The tacit dimension,* Garden City, N.J.: Doubleday, 1975, pp. 28-9.

[39] M. Merleau-Ponty, *The phenomenology of perception,* C. Smith (trans.) of *Phénoménologie de la perception*", Paris: Gallimard, Oxford: Routledge & Kegan Paul, 1945/1962.

[40] M. Merleau-Ponty, *Le Visible et L'Invisible*, Paris:Gallimard, 1964.

[41] S. Todes. *Body and World*, Cambridge, MA: The MIT Press, 2001.

[42] N. Cumming, "The sonic self: musical subjectivity and signification," in *Advances in semiotics,* Bloomington, Ind: Indiana University Press, 2000.

[43] H.R. Maturana and F. J. Varela, *The tree of knowledge: the biological roots of human understanding,* Boston: New Science Library, 1987.

[44] B.H. Repp, B.H. 1993. "Music as motion: a synopsis of Alexander Truslit's (1938) *Gestaltung und Bewegung in der Music*," in *Psychology of Music*, Volume 12, Number 1, 1993, pp. 48–72.

[45] N. Nettheim, "How musical rhythm reveals human attitudes: Gustav Becking's theory," in *International Review of the Aesthetics and Sociology of Music,* Volume 27 Number 2, 1996, pp. 101– 122.

[46] M. Clynes, S*entics: the touch of emotions*. New York: Anchor Press, 1977.

[47] E.F. Clarke, "Empirical Methods in the Study of Performance," in E. Clarke and N. Cook, (eds.), *Empirical musicology: Aims, methods, prospects,* Oxford: Oxford University Press, 2004, pp. 77-102.

[48] R. Pelinski, "Embodiment and Musical Experience," in *Transcultural Music Review* Nr 9, 2005. Available at http://www.sibetrans.com/trans/trans9/pelinski-en.htm Accessed 7 June 2009.

[49] T. Winkler, "Making motion musical: Gestural mapping strategies for interactive computer music," in *1995 International Computer Music Conference*. San Francisco: International Computer Music Association, 1995.

[50] G. Paine, "Gesture and musical interaction: interactive engagement through dynamic morphology," in *Proceedings of the 2004 conference on New interfaces for musical expression,* Hamamatsu, Shizuoka, Japan, 2004, pp. 80–86.

[51] G. Paine, "Towards Unified Design Guidelines for New Interfaces for Musical Expression," in *Organised Sound*, Volume 14 Number, 2009, Cambridge UK: Cambridge University Press, pp. 142-155.

[52] *Online Etymology Dictionary*, D. Harper (ed.), 2001, http://www.etymonline.com/ Accessed 12 February 2010.

[53] D.R. Worrall, "Overcoming software inertia in data sonification research using the SoniPy framework," in *Proceedings of the Inaugural International Conference on Music Communication Science*, Sydney, Australia, December 5-7, 2007.

SPATIAL AUDIO VS. VERBAL DIRECTIONAL CUES: AN EXAMINATION OF SALIENCE AND DISRUPTIVENESS WITHIN A SIMULATED DRIVING CONTEXT

Jane H. Barrow and Carryl L. Baldwin

George Mason University
Psychology Dept., Arch Lab
4400 University Dr., MS 3F5, Fairfax, VA 22030, USA
jbarrow1@gmu.edu and cbaldwi4@gmu.edu

ABSTRACT

A spatial auditory Stroop paradigm was used to examine the effects of verbal-spatial cue conflict on response accuracy, reaction time, and driving performance. Participants responded to either the semantic meaning or the spatial location of a directional word, which was either congruent (i.e. the word "right" being presented from the right) or incongruent (i.e. the word "right" being presented from the left), while following a lead car in a simulated driving scenario. Greater performance decrements were observed when participants were attempting to ignore a semantically incongruent verbal cue relative to incongruence from the spatial location of the cue. Implications for the design of spatial auditory displays are discussed.

1. INTRODUCTION

Spatial auditory cues and verbal directional cues are both viable options for alerting the human operator to danger in a variety of environments [1], [2], [3], [4]. The question of interest here is which type of information is more salient, and therefore more disruptive when irrelevant or erroneous. Understanding the roles that verbal and spatial components of an auditory message play is an important step in learning to design cues and alerts for a variety of display applications, not just for alerts to potential dangers in the environment.

For the purposes of this paper, the verbal and spatial location components of auditory alerts are examined in terms of their spatial orienting effectiveness within the driving domain. Automated systems are increasingly being implemented in modern automobiles in an effort to increase safety, and because driving is a visually demanding task, the auditory modality is ideal for presenting supplementary information to aid the driver [5]. There is not yet a consensus in the literature as to the relative benefits of alerting drivers to the relevant location of critical events with spatial versus semantic audio cues. Ho and Spence [2] found that participants responded faster to verbal directional cues than they did to non-verbal directional cues, indicating that the semantic information provided by the verbal directional cue was processed more quickly than spatial information provided by the non-verbal directional cue. However, when the two cues were combined to create a congruent verbal-spatial directional cue, participants responded faster still. This finding indicates that having redundant information can actually speed reaction time (i.e. both verbal and non-verbal/spatial directional cues provide the same information). This observation is supported by research on multi-modal redundant targets [3], [6], [7]. However, Ho and Spence did not investigate the effect of an incongruent, or conflicting verbal-spatial directional cue.

Clearly, designers would not intentionally use conflicting pieces of information to relay spatial information to the driver. However, technologies are not infallible and modern vehicles are not a silent environment. Many drivers utilize GPS navigation systems while driving, which provide verbal directions about when and where to make turns in their route. When directional information is being provided by two different sources, especially when these sources may be using different forms of auditory directional cues, there is an increased possibility that directional information from one source could conflict with directional information from another source. Wang, Pick, Proctor, and Ye [4] touched on this very issue in their investigation of driving responses to a Side Collision-Avoidance System (SCAS) when navigation signals were present. They found no differences in reaction time to the SCAS warning when the navigation signal corresponded with the SCAS warning and when it conflicted, but the navigation information was provided visually while the SCAS warning was provided aurally. Research using cross-modal Stroop paradigms has shown that when auditory and visual cues conflict, there is a significant lag in reaction time to the target when presented with an invalid auditory cue but not when presented with an invalid visual cue [8]. This suggests that visual information is easier to ignore than auditory information, which could explain why there was no difference in reaction times for conflicting and non-conflicting cues from the navigation and SCAS systems in Wang et al.'s study. Participants could have been prioritizing the auditory SCAS warning.

To further investigate this issue, the current study utilized a spatial auditory Stroop task originally used by Pieters [9]. The paradigm consists of verbal directional information presented from either a congruent spatial location (i.e. the word "right" presented from the right) or an incongruent spatial location (i.e. the word "right" presented from the left). Participants would either be responding to the spatial location of the stimulus, or the semantic meaning of the stimulus. Participants performed this task while following a lead car in a simulated freeway environment on a desktop driving simulator. This allowed us to examine the relative interference of the spatial component of the stimulus and the semantic meaning of the stimulus.

It was hypothesized that performance on the dependent measures would be better in congruent trials than in incongruent trials, and that performance would also be better in the location auditory task than the semantic auditory task, based on the nature of the auditory system. It was also hypothesized that due to the predicted preference for responding to location information over semantic content of an auditory cue, incongruent trials where a participant was performing the semantic auditory task (and therefore ignoring location information) would result in poorer performance on the dependent measures.

2. METHOD

2.1. Participants

Voluntary participation was obtained from 18 undergraduates (16 female) with a mean age of 19.69 years (SD = 2.02) enrolled in a university on the east coast. All participants reported normal or corrected-to-normal vision and passed an audiometric assessment of their hearing, indicating that their puretone hearing level was less than 24 dB across 250-8000 Hz. All participants were fluent in English.

2.2. Materials and Apparatus

Auditory stimuli consisted of the words "right", "left", and "house" spoken in a naturalistic female voice, digitized and then presented in either the right channel, left channel, or both channels. All auditory stimuli were presented at a level approximating 60 dB from free field computer speakers. The speakers were placed 42 inches apart, with the participant seated directly between them.

The simulated driving task required participants to follow a lead car while maintaining a consistent headway, speed, and lane position on a four-lane freeway with no ambient traffic. When the participant began driving, the lead car began to move forward, then sped up to maintain a constant speed of 65 mph. Images of common brand logos were presented on billboards on both sides of the road during the driving simulation. Two series of billboard images were constructed so that no images were repeated from one condition to the next.

2.3. Experimental Design and Tasks

2.3.1. Auditory tasks

Trials consisted of the words "right" or "left" coming from the right or left speaker. Stimuli were the same in the two auditory tasks, with the exception of control trials, but the instructions changed the nature of how the task was performed. Each task consisted of congruent, incongruent, and control trials as detailed below. Reaction time and accuracy were recorded for both tasks.

In the **semantic task**, participants were instructed to respond to the semantic meaning of the word by depressing a key representing "right" if they heard the word "right" and vice versa for the word "left", regardless of the spatial location of the word. Congruent trials occurred when the semantic meaning of the word matched the presentation location (i.e. the word "right" came from the right), and incongruent trials occurred when the semantic meaning of the word did not match the presentation location (i.e. the word "right" came from the left). A control trial occurred when the word "right" or "left" came from both speakers, eliminating the directionality of presentation location.

In the **location task,** participants were instructed to indicate the spatial location of the word presented by depressing a key representing "right" if they heard a word presented from the right and vice versa for a word presented from the left, regardless of the semantic meaning of the word. A control trial in this task consisted of the word "house" coming from either the right or the left speaker, eliminating the semantic meaning of the spoken word in terms of directionality.

2.3.2. Driving Task

Participants were instructed to follow the car in front of them at what they deemed to be a safe following distance, while maintaining a speed of 65 mph and their lane position. In the event that the participant lost the lead car (fell too far behind to safely catch up), they were instructed to maintain their speed and lane position, and not worry about trying to catch up to the lead car. Average speed and lane deviation were measured.

2.3.3. Billboard Task

Participants were instructed to remember as many of the logos on the billboards as possible while performing the other two tasks. The experimenter clearly indicated that this was the lowest priority task – participants were asked to focus on maintaining their driving performance and their speed and accuracy on the auditory task. Participants received two scores: one for the number of correct, freely recalled logos, and one for the number of logos recognized in a subsequent recognition test that included both old and new logos.

2.3.4. Design

A 2x3 mixed-factorial design was used to examine the effects of response type (semantic vs. location) and congruency (congruent, control, or incongruent). Dependent measures were reaction time and accuracy for the auditory tasks, deviation from average speed and lane position for the driving task, and the number of correctly recalled and recognized logos for the billboard task.

2.4. Procedure

Upon entering the laboratory, participants were given an audiometric assessment and then completed a demographic questionnaire, way-finding surveys and the Edinburgh Handedness Inventory [10]. For the first block of the experiment, the experimenter verbally gave instructions to the participant on how to perform the auditory task, allowed the participant to practice the task, and then gave instructions to the participant on how to perform the driving task, followed again by practice. The participant then practiced both tasks together. The experimenter gave verbal instructions on the billboard task, reiterated the instructions for the auditory and driving tasks, then started the experimental trials. At the end of the experimental trials, the participant completed the NASA-TLX [11] with instructions to rate workload only on the auditory task. Next, the participant freely recalled the images that he or she remembered from the billboards, and then went through a slideshow of images to indicate which images they had seen in the driving scene and which were novel. The participant was offered a break, and then followed the same procedure for the second block of the experiment, minus the practice session for the driving task, since it did not change. The order of auditory tasks was counterbalanced across subjects, as were the driving scenes. Additionally, a baseline was taken of the participant's response time to each word in the auditory task (without the presence of spatial information). In half the participants, the baseline was taken prior to starting the first block of the experiment, and in the other half, the baseline was taken after the second block.

	Condition	Trial Type	Mean	SD
~~RT (ms)~~	~~Location~~	Congruent	921.34	187.56
		Incongruent	973.39	198.65
	~~Semantic~~	Congruent	915.87	139.32
		Incongruent	948.70	147.79
%Correct (percentage correct)	Location	Congruent	.96	.05
		Incongruent	.87	.10
	Semantic	Congruent	.98	.02
		Incongruent	.95	.02

Table 1: Descriptive statistics for auditory tasks.

3. RESULTS

Two participants (both female) were excluded from the analyses due to computer failure during the experimental session, which resulted in incomplete data being recorded. Examination of the baseline data revealed that participants responded significantly faster to the word "house" than they did to either "right" or "left", F(2,30) = 27.32, p < .05, but that there was no difference in response time to the words "right" and "left". This observation indicates that the digitized word "house" may have been more acoustically salient, resulting in people consistently responding to it faster. We excluded all control trials from the analysis due to this confound, and only examined the differences between congruent and incongruent trials.

3.1. Auditory Tasks

Descriptive statistics for reaction time and accuracy to the auditory task trials can be found in Table 1. As predicted, a two-way repeated measures MANOVA revealed that accuracy was better in the congruent trials than the incongruent trials, F(1,15) = 18.23, p < .05. Accuracy was also significantly better in the semantic condition than in the location condition, regardless of the congruency of the trial, F(1,15) = 13.13, p < .05. This was interesting,

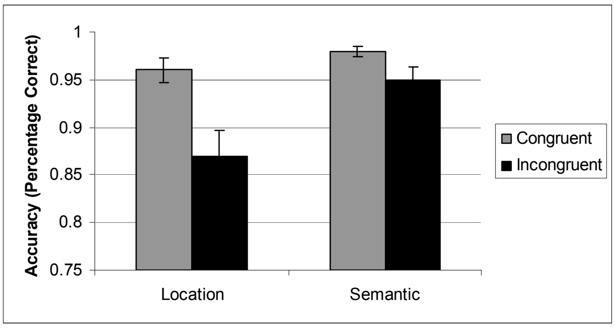


Figure 1: Accuracy for each response condition plotted as a function of congruency. Error bars represent the SEM.

since we originally predicted that people would be faster and more accurate in the location condition. Further analysis revealed an interaction between response type and congruency that approached significance, F(1,15) = 2.98, p = .11. The trend in the data indicated that incongruent semantic information tended to disrupt performance to a greater degree than incongruent location information (see Figure 1 and 2). Both reaction time and accuracy suffered to a greater degree from the incongruent semantic information. This suggests that the overall superior performance in the semantic condition may have resulted primarily from the absence of detrimental effects of incongruent spatial location information.

3.2 Driving and Billboard Tasks

Driving data (average speed and lane deviation) and billboard logo recall and recognition were analyzed using two one-way repeated measures MANOVAs. Comparisons were only made between performance on the semantic auditory task and the location auditory task. No significant differences were observed for any of these measures.

4. DISCUSSION

The results of this study support those of Ho and Spence [2], indicating that congruent verbal-spatial directional information leads to a faster response than non-spatial information. Additionally, verbal directional information results in a faster response than non-verbal directional information. Accuracy data in the current study show that participants were more accurate when responding to verbal (semantic) information relative to when they were responding to non-verbal (location) information, and the reaction time data show a similar trend. Wang et al. [4] found no difference in reaction time to an auditory collision avoidance warning whether it conflicted with visually presented navigation directions or not. In the present study, we did not manipulate congruence of visual stimuli but focused entirely on the congruence of auditory information. Our results indicate a marginally significant difference in accuracy for incongruent trials depending on whether semantic or location information was being attended. Specifically, participants demonstrated a trend for greater interference from incongruent semantic information when responding to the location of a word, relative to incongruent spatial information when responding to the physical location of a sound. This further supports Ho and Spence's [4] results that show the importance of verbal directional cues relative to non-verbal spatial directional cues in terms of improving performance. However, these results raise an important caveat. Depending on the reliability of the system, using semantic methods of presenting spatially predictive information may pose significant problems. Incongruent semantic information may cause greater disruption. It may be more difficult to ignore the semantic content of a conflicting stimulus rather than its spatial location, thus potentially negating the benefit of semantic spatially predictive cues.

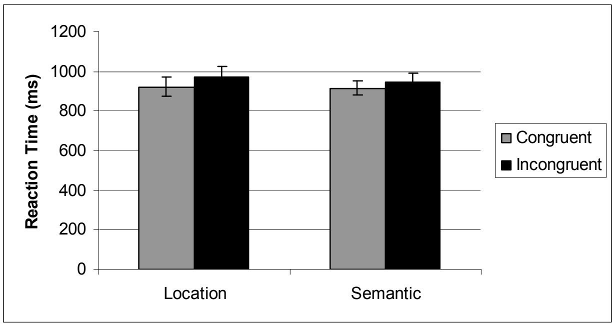


Figure 2: Reaction time for each response condition plotted as a function of congruency. Error bars represent the SEM.

These findings support previous research demonstrating the salience of semantic information, but also illustrate the potential for that information to disrupt response to other tasks. In imperfect systems, it might be wiser to use spatial audio information, which is valuable in directing attention, and less disruptive, particularly in situations where the other tasks being performed require directional judgments and may be equally important to the nature of the spatial audio alert.

5. ACKNOWLEDGMENTS

The authors would like to thank Dan Roberts and Tony Layton for their assistance with data collection, and their feedback regarding the experimental tasks and experimental set-up. The authors would also like to thank David Kidd for his invaluable help in programming the driving scenarios.

6. REFERENCES

[1] D.R. Begault, "Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation," in *Human Factors*, vol. 35, no. 4, pp. 707-717, 1993.

[2] C. Ho & C. Spence, "Assessing the effectiveness of various auditory cues in capturing a driver's visual attention," in *Journal of Experimental Psychology: Applied*, vol. 11, pp. 157-174, 2005.

[3] R.S. Tannen, W.T. Nelson, R.S. Bolia, J.S. Warm, & W.N. Dember, "Evaluating adaptive multisensory displays for target localization in a flight task," in *International Journal of Aviation Psychology*, vol. 14, pp. 297-312, 2004.

[4] D. Wang, D.F. Pick, R.W. Proctor, & Y. Ye, "Effect of a side collision-avoidance signal on simulated driving with a navigation system," in *Proceedings of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, Portland, Oregon, 2007, pp. 206-211.

[5] T.A. Dingus, M.C. Hulse, & W. Barfield, "Human-system interface issues in the design and use of Advanced Traveler Information Systems," in W. Barfield & T.A. Dingus (Eds), *Human Factors in Intelligent Transportation Systems*, Mahwah, NJ: Lawrence Erlbaum, 1998.

[6] R.S. Bolia, W.R. D'Angelo, & R.L. McKinley, "Aurally aided visual search in three-dimensional space," in *Human Factors*, vol. 41, pp. 664-669, 1999.

[7] M. Gondan, B. Neiderhaus, F. Rosler, & B. Roder, "Multisensory processing in the redundant-target effect: A behavioral and event-related potential study," in *Perception & Psychophysics*, vol. 67, no. 4, pp. 713-726, 2005.

[8] A.R. Mayer & D.S. Kosson, "The effects of auditory and linguistic distractors on target localization," in *Neuropsychology*, vol. 18, pp. 248-257, 2004.

[9] J.M. Pieters, "Ear asymmetry in an auditory spatial Stroop task as a function of handedness," in *Cortex*, vol. 17, pp. 369-380, 1981.

[10] R.C. Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory," in *Neuropsychologia*, vol. 9, no. 1, pp. 97-113, 1971.

[11] S.G. Hart & L.E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in P.A. Hancock & N. Meshkati (Eds), *Human Mental Workload,* Amsterdam: North Holland Press, 1988.

Sudden Ionosphere Disturbance Sonification Project for Increased Accessibility by the Blind and Augmented Cognition

Marty Quinn
Design Rhythmics Sonification Research Lab,
92 High Rd, Lee, NH USA, 03861
marty@drsrl.com

Deborah Scherrer
Stanford Solar Center
Stanford University
HEPL-4085
Stanford, California 94305-4085 USA
dscherrer@solar.stanford.edu

ABSTRACT

The Sudden Ionosphere Disturbance (SID) Sonification Project provides a way to query multiple stations across the globe and hear the Sun's impact on the ionosphere, an electrically charged layer 50 miles above the Earth, as a kind of Sun-Earth symphony. Nine instruments represent stations, pitches in one of over fifty scales (Spanish-Gypsy default) represent disturbance values and panning location represents longitude of the reporting station. Lower values map to lower notes and higher values to higher notes. A differentiation/integration 'cognitive mixing' slider provides for four levels/combinations of instrument assignments alternating between unique (differentiated) vs one instrument (integrated), and panning positions alternating between panned, based on longitude (differentiated), and center panned (integrated). Station data can be muted and played forwards or backwards in slow to very fast tempos. Other options include focusing on change in the data over similarity by lowering the volume of notes or not repeating notes if they remain unchanged.

EYES-FREE METHODS FOR ACCESSING LARGE AUDITORY MENUS

Raine Kajastila

Aalto University
Department of Media Technology
P.O. Box 15500, FI00076 Aalto, Finland
raine.kajastila@tkk.fi

Tapio Lokki

Aalto University
Department of Media Technology
P.O. Box 15500, FI00076 Aalto, Finland
tapio.lokki@tkk.fi

ABSTRACT

Two interaction methods for eyes-free control of a mobile phone or a media player are introduced. The methods include a gestural pointing interface and a touchscreen interface to a spherical auditory menu where feedback is provided using spatially reproduced speech. The methods could facilitate the eyes-free use of devices and also make them accessible for visually impaired users. The effectiveness of gestural and touchscreen interaction is compared to traditional visual interface when accessing large menus. Evaluation results prove that moderately fast and accurate selection of menu items is possible without visual feedback. Combining eyes-free interfaces, positions of menu items in 3D and a browsing method with a dynamically adjustable target size of the menu items allows the use of large menus with intuitive easy access.

1. INTRODUCTION

The design and use of audio-only and eyes-free interfaces has been emerging in recent years. They can bring better usability in situations where eyes-free operation is necessary [1]. Such cases include the competition of visual attention, absence or limitations of visual display, or reduction of battery life [2]. With proper design, an audio interface can be even more effective than its visual counterparts [2]. Although audio interfaces are not yet widely used in public, major companies have already realized their potential [3]. Furthermore, they are important as assistive technology for visually impaired users.

This paper presents two interaction methods that allow a reasonably fast and accurate way to navigate spherical auditory menus with large number of menu items. This work builds upon the author's previous work [4], which is presented in Figure 1. The user points or tilts the control device to different directions to browse an egocentric auditory menu. As the user browses the menu items, they are read out loud and the sound is reproduced from the correct 3D direction. Fast browsing is enabled with reactive interruptible audio design [2] and the accuracy in selection is enhanced by the dynamic movement of menu items and an expansion of the selection area. The initial study [4] proved that this type of gestural interaction with auditory menus is efficient and intuitive. The second novel interaction method uses touchscreen input with auditory menus. With good design, auditory menus can be combined to work seamlessly with visual menus [5], thus making eyes-free use of devices intuitive and easy. Touchscreens can also be a barrier for visually impaired users [6], but visual touch screen menus can be made easily accessible by using audio feedback.

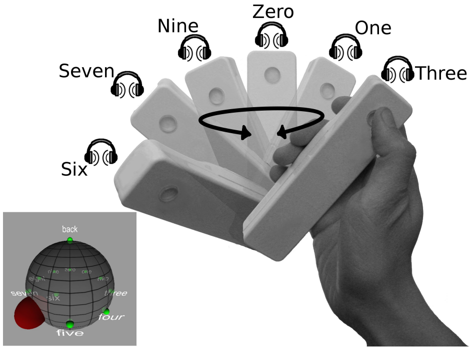


Figure 1: Gesture interface utilizing a mobile device mockup, as introduced in [4]. Auditory menu items can be accessed by pointing or tilting the device to desired direction. Browsing can be continued to the neighboring menu items by rotating the wrist.

The two menu configurations and the introduced interaction methods make it possible to browse auditory menus with large number of items (>100), for example, a contact list in a phone or a playlist in a music player. Furthermore, the interaction methods can be used to handle all basic controls of a modern mobile phone or a music player that contains either accelerometers or a touchscreen. It is also possible to construct a small multi-functional device consisting of only one button and an internal rotation sensing devices, for example, accelerometers or alternatively a touch surface device without a screen. Such robust devices without visual displays can be inexpensive and have low energy consumption but still offer the same functionalities as similar devices with a visual screen. A good example of a such a device is Apple's iPod shuffle [3], which gives feedback to users using synthesized speech. Our interaction and browsing techniques enable more sophisticated control of devices such as the iPod shuffle.

The novelty of the presented techniques lies in advanced auditory menus that can be used with two parallel interaction methods. In the gesture interaction, simple and intuitive wrist rotations are measured with three accelerometers. Touch interaction extends the circular touchpad implementation of earlier work [2] to be applied in touchscreen devices. Both methods can be combined with various browsing techniques and provide a way to access a high number of selectable content especially in 360-degree spherical auditory menus. Furthermore, menus enable the interoperability of the visual and auditory menus where the control logic remains the same in visual and auditory menus, as described in author's previous work [5].

2. RELATED WORK

The implemented interaction methods mainly build on auditory menus controlled with gestures or a touchscreen.

2.1. Auditory menus

In previous studies, different interaction methods with audio objects in menus and different spaces surrounding the user have been presented, and their input methods range from normal keypads and touch interfaces to gestural interfaces. Hiipakka and Lorho [7] used cursor keys for a spatial audio interface for music players. Their approach enabled a browsing system where spatialized sounds along the horizontal axis informed the user about context, menu structure and interaction possibilities.

Pirhonen et al. [8] tested a prototype eyes-free touch interface for a music player, in which finger sweeps on the screen controlled the playing of the music. The interface was provedn to be effective in eyes-free situations, and the results of the study pointed out that immediate audio feedback is crucial for user confidence.

Savidis et al. [9] presented a method of pointing interaction where a data glove, head tracker, voice recognition and headphones were used to produce a modifiable circular audio environment. They used the concept of auditory windows, where a subset of four sound objects was simultaneously played in a spatially larger area, while others were suppressed closer together.

An egocentric circular auditory menu, which is also applied in this paper, has been proposed many times earlier. Brewster et al. [1] used a directional head nodding interface to study four simultaneously playing menu items located around the user. They found egocentric menu design better than exo-centric and the selection method using the head-tracker was also successful in a mobile experiment. Brewster et al. [1] hypothesized that more than 8 simultaneously playing menu items would be difficult to handle with the system in their experiment. Circular auditory menu structures have also been applied in nomadic radio by Sawhney and Schmandt [10] and in diary application by Walker et al. [11].

Marentakis and Brewster [12] studied audio target acquisition in the horizontal plane with the aid of orientation trackers. The experiment focused more on how target width, distance and user mobility affects random target acquisition with a gestural pointing interface. They concluded that a pointing interaction with 3D audio is successful with mobile users. They also suggested that audio elements with feedback in egocentric audio displays could produce efficient design.

The user studies of circular auditory menus with touch input by Zhao et al. [2] showed that auditory menus can outperform typical visual menus used in iPod-like devices. The menu used by Zhao et al. is similar to the one presented in this paper. Their key elements include: 1) a touch interface similar to the iPod, where menu items are mapped to a circular touchpad; 2) direct reactivity to user touch input that gives control to the user without waiting periods; 3) interruptibility of the audio, where only one sound is played at a time, but its playing can be interrupted if user chooses to continue browsing; and 4) menu items that can be accessed directly without browsing through all items.

2.2. Gestural and touch control

Different types of tilting interfaces have been presented mainly for visual displays and writing applications. TiltType [13] has a writing interface where the tilt direction of the device can be used two-handedly to specify letters with the aid of 4 buttons. Wigdor and Balakrishnan [14] proposed a similar system in mobile phones, where tilting direction and numeric keypad press define the output character. Oakley and Park [15] presented a one-dimensional tilt menu system for mobile phones with tactile feedback. Tian et al. [16] studied a circular tilting menu using a pen, where the pen tilt direction was used to select visual menu items. One of the first tilting interactions was presented by Rekimoto [17]. He utilized a FASTRACK position and orientation sensor and applied tilting and a two-button-device to browse menus on a visual display.

Recently, the interaction with wrist rotations of horizontally held arm has been studied by Crossan et al. [18]. Their multipart mobile device consisted of a SHAKE sensor pack attached to the user's hand as a wristwatch and a Nokia N95 as visual feedback. They concluded that horizontal wrist rotation is a quite accurate and feasible control method, but simultaneous walking makes control more difficult.

The Wii Remote has been used in audio only music browsing by Stewart et al. [19] for moving in a large music collection located in surrounding 2D or 3D spaces. According to the authors, their Wii interface was not really usable. It only allowed the definition of general moving directions and was given negative evaluations by most of the users. The interaction methods presented in this paper would radically improve the usability of the music browser, because it would allow an exact definition of the direction of movement.

One important aspect of eyes-free auditory menus is their suitability as assistive technology for blind users. Guerreiro et al. [6] have implemented a gesture-based text entry method for touchscreen devices. In their NaviTouch interface, all the letters are accessed through vowels. The user first slides his finger vertically to find vowels that are read out loud. After hearing any of the vowels (e.g. A), the user can slide his finger vertically to find consonants that are after that particular vowel in the alphabets (e.g. B or C). The user makes one L-shaped gesture for each successful consonant selection. With this approach, the alphabets cannot be accessed directly, which slows down the writing process.

Another good example of touchscreen input is Slide Rule [20]. Kane et al. used the iPhone for eyes-free browsing of lists, selecting items, and browsing and changing music tracks. Kane et. al. also used similar L-shaped touch-gestures for browsing music tracks in Navitouch [6]. In the experiment, 10 album names were placed vertically in a list. Each item on the list could be listened to one at a time. The user first found a desired album with a vertical finger-swipe and continued the finger movement to the right to heari the track names read out loud. The multi-touch capability of the iPhone was utilized by tapping the screen with a second finger to select the desired track. Although the songs can be accessed by using only one continuous touch-gesture, this method does not solve the problem for music libraries holding hundreds of albums; queezing them into a vertical list would be difficult.

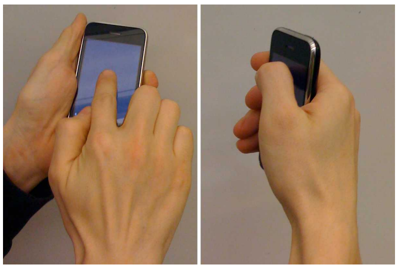


Figure 2: Two interaction methods was implemented as prototype applications for the iPhone. No visualizations of the auditory menu was shown in the device.

3. INTERACTION METHODS

In this work, two eyes-free interaction methods, wrist rotation gestures and touch-gestures of the finger are applied. The prototype implementation was aimed at the iPhone, but the interaction methods can be applied with other devices as well. Next, the novel control methods are introduced.

3.1. Gestural control method

Accelerometers along three axes can be used to sense the orientation of a device relative to the direction of earth's gravity. Problems can occur when the device is rotated horizontally along one axis, because accelerometers cannot sense that type of motion. This scenario can be avoided by holding the device as illustrated in Figure 2 and rotating it as show in Figure 1 [4]. The device can be used like a joystick by tilting the device slightly and rotating it 360 degrees with a gentle wrist gesture. To ease the targeting of menu items, the tilt angles of the device between 5 and 90 degrees are clamped to 90 degrees, i.e., zero elevation. This allows smaller wrist movements and prevents any tedious turning and twisting of the wrist. This control method is especially suitable for horizontal 360-degree menus, because all directions can be reached with equal effort and ease.

3.2. Touch-surface control method

A touch-surface (or a screen) can also be used to access a circular auditory menu, see Figure 2. Sectors extending from the center of the surface represent the menu items, as shown in Figure 3. The user can access any item directly by placing a finger on the surface and can continue browsing with a circular finger sweep. A selection is made by removing the finger from the surface. The center of the touch-surface is a safe area from where the finger can be lifted without making a selection. As explained later, special actions can be assigned to the center of the touch-surface.

4. SPHERICAL AUDITORY MENUS

The auditory menu described in this paper has similarities to the works of Brewster et al. [1] and Zhao et al. [2]. The key element is the use of interruptible audio and immediate reactivity to user input with an auditory display. The spoken menu items are played one by one while browsing a menu and the user has the ability to jump to the next item thus stopping the playback of the previous one. With slower motion, the user can hear all menu items one by one. When jumping to the other side of the menu, no sound is heard until the gesture or finger movement has slowed down. Thus, the user is in control and he/she can adjust browsing speed according to his/her own abilities.

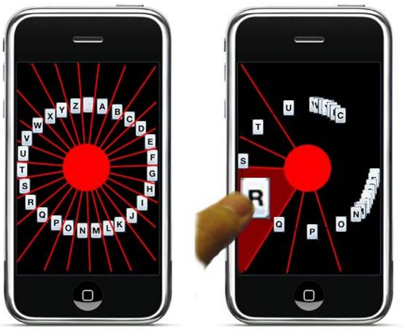


Figure 3: The menu items are defined as dynamically changing sectors on a screen. The visualization is exaggerated and artificially created. It is not needed in eyes-free auditory menu.

When browsing faster, the user hears only the beginning of the sounds. Because the short sounds (or phonemes) represent the first letter(s) of the names, they help the user keep track of the position in a large menu. This feature was recently evaluated as beneficial and was suggested to be named “spindex” [21]. In the author's previous works [4][5] and in the auditory menu described in this paper, the spindexes are automatically generated when the user browses the menu. This is achieved by the auditory menu's instant reactivity to users' gestures by using prerecorded names or fast text to speech synthesis. By slowing down the browsing speed, the user can adjust the length of the spindex thus enabling an efficient search method for menu items starting with a same letter, letters or even word.

To enhance the selection accuracy, a dynamically adjusted target sector, where the item is active (played), is applied. As visualized in Figure 3 (left), if none of the items is active, the menu items have an even distribution of the target area. When a menu item is active, its target area expands in both directions reaching a 1.9 times larger target area. The value of 1.9 was chosen to leave a big enough target area for the neighboring menu items, because they shrink, allowing space for the expanding sector. This is done to facilitate easier browsing and selection by reducing undesired jumping between tightly packed menu items.

When a selection is made by removing the finger from the screen, it is important to give feedback to the user. There are many suggested feedback sounds, e.g., auditory icons [22], earcons [23] and spearcons [24]. The implementation presented in this paper, uses a fast replay of the selected menu item mixed with a short auditory icon. A short clink-sound is played immediately after the selection, followed by the fast replay of the selected menu item. The playback time of the sound is shortened considerably, but the user can still easily recognize the content. The clink-sound further clarifies that the selection was made. The changed pitch also indicates a feedback sound, not another menu item. In this way, the user gets immediate feedback and he/she can easily double-check whether a correct selection was made. A modified sound sample of a plucked guitar was used to inform the participants if

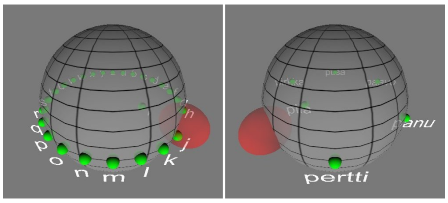


Figure 4: The two-layer menu. The alphabets are always found from the same position in the first menu layer. The second layer holds 6 names in alphabetical order and they are spread evenly.

they moved their finger to the center of the screen or pointed the device up. This was done to notify the participant in the case of an advanced one-layer menu that menu items have been spread evenly around the user.

The correlation between the touched screen location and the reproduced sound directions can help the user associate the sound to the specific menu item location [12]. Proper design can also improve the performance as the spatial menu item configuration becomes familiar to the user. Furthermore, each menu item is heard from a different spatial direction making it easier to distinguish them when browsing with increased speed. The binaural implementation for headphone reproduction applies head-related transfer functions (HRTFs) measured in house in an anechoic environment with a dummy head. In addition, the sounds are processed with a simple reverberation algorithm, which helps in externalization of auditory menu items. Informal listening tests suggest that spatial audio makes the use of the system easier. However, no user tests have been done regarding how much (if any) the mono sound reduces the usability and whether it is possible to use the system with only one headphone or loudspeaker.

The goal of this work is to study the selection speed and accuracy of auditory menus with large number (>100) of items. We designed two alternative auditory menu layouts containing a contact list of 156 (26 x 6) names. As explained below, one menu layout is a more traditional auditory menu with two layers and other uses a novel approach to fit all 156 names to one menu level. Earlier studies with different interaction methods have suggested that egocentric auditory menus could contain at maximum 5 [25], 8 [1], or 12 [2] menu items in usable scenarios. This is probably due to limitations in interaction devices, browsing methods or simultaneously played sounds. With our auditory menu layout the number of names displayed to the user can be dramatically increased compared to the number of names in Slide Rule [20].

4.1. Menu with two layers

The general layout of the two layer menu is visualized in Figure 4. The first menu level consists of 26 letters from A to Z, which are always found in the same locations and are placed in alphabetical order. When none of the menu items is selected, the target sector width in the first menu level is 13.85 degrees (360 / 26). After an item is activated by touching or pointing, its target sector expands to occupy 26.3 degrees (13.85 x 1.9). In the implemented prototype, the user can select a menu item by releasing the finger from iPhone's touchscreen. By selecting a letter, the user can advance to the second layer of the menu. The second layer holds six names starting with the chosen letter. They are evenly spread around 360 degrees and have a target area of 60 degrees. The target area of an active item is now 114 degrees. The names are listed in alphabetical order. For testing purposes, after a name selection, the menu automatically jumps to the first layer. In this prototype, there was no option for going back in the menu structure.

4.2. Menu with hundreds of items in one layer

The novel one-layer menu layout can handle very large number of items, as shown in Figure 5. The applied browsing method combines the benefit of a-priori known item positions in a static menu with large menus. In this approach, when none of the items is selected the menu items are in their absolute positions defined by alphabetical order. For example, all the names starting with the letter A are placed in alphabetical order to the sector that occupies the letter A in the menu shown in Figure 4 (left). Thus, the user can point to or touch desired position and hear one of the names starting with that letter. When none of the items is selected, the target sector width for each menu item is 2.31 degrees (360 /156).

When the user targets a particular item, its neighboring items are spread around evenly with a spacing of 40 degrees, and items farther away are grouped together (see Figure 5, right). If the desired menu item was not directly found, the user can continue browsing items with a rotating hand gesture or a circular finger sweep. The next item is always found 40 degrees forward and the previous one 40 degrees backwards, respectively. Now the target sector width is significantly larger - 76 degrees (40 x 1.9) for the active menu item. Such dynamic item placement greatly reduces undesired jumping between items. This spreading can also be seen as a different implementation of the fisheye distortion concept [26, 27]. In the preliminary tests, the spacing was set to 40 degrees, but further study is needed to find the optimal spacing for different interaction methods.

Initial testing suggested that it is possible to access the desired item without extensive training. This browsing method can significantly facilitate the browsing of alphabetically ordered large menus, such as a music playlist or an address book containing hundreds of items. In the visual domain, this is similar to the possibility of jumping directly to names starting with the same letter in an address book.

5. USER EXPERIMENT

A mobile user experiment was conducted to ascertain the performance of the introduced eyes-free menu browsing techniques. The test environment is depicted in Figure 6 (left). A participant navigated his way around four aligned chairs while choosing names from a large menu. The auditory menus were reproduced with headphones and the names to be chosen with different interaction methods were listed on the large projection screen.

5.1. Participants

Nine participants completed the experiment. All of them were males between 23 and 43 years old. The participants volunteered for the experiment, and they had no previous experience with the interaction methods and auditory menus used. Three participants owned iPhones. Additionally, three participants were left handed.

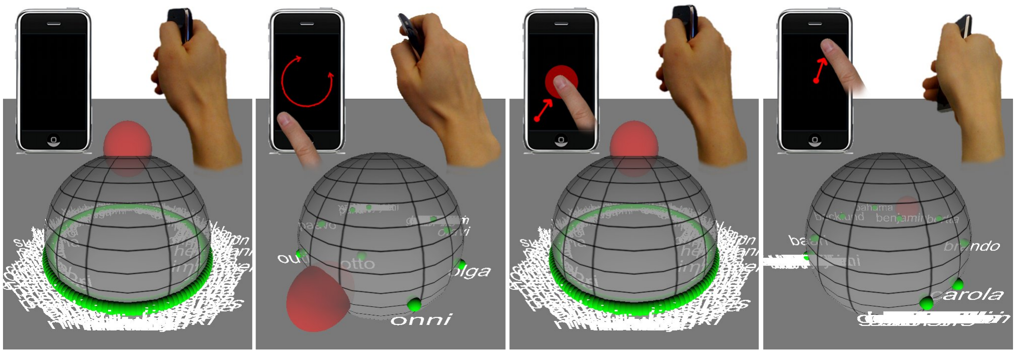


Figure 5: Method for browsing large eyes-free auditory menus. It combines qualities of absolute positioning of menu items and browsing easiness of smaller menus. The browsing method can handle hundreds of items that can still be accessed fast.

5.2. Apparatus

An iPhone was used as the test device. The iPhone was chosen because it had the desired features and only one device for all interaction methods was needed. The iPhone was connected to a Macbook Pro using a wireless network. The participants used Sennheiser HDR 120 wireless headphones whose transmitter was connected to the laptop's audio output. This setup allowed the participants to be truly mobile without the need for all the software running in the iPhone. The visual reference test was implemented with the iPhone's contact list application.

The connection between the iPhone and the laptop was implemented by using the Open Sound Control (OSC) protocol and a modified version of the free Mrmr software [28] installed on the iPhone. The auditory menu was implemented using Pure Data (PD) [9], which received the raw control information from the iPhone. High quality samples (fs = 44.1 kHz) for the menu items were pre-recorded by synthesizing them with text-to-speech software. All samples started immediately in the beginning of the sound file to ensure fast responses to user actions.

5.3. Procedure

After a brief tutorial of the two interaction methods, the participants were given headphones and an iPhone. The touch interface was introduced to the participants before the gesture interface. Both tutorials consisted of writing a test sentence with total of 25 letters by using the menu shown in Figure 4 (left). This took approximately 5 minutes. The visualization was shown on a laptop screen when writing the first word of the test sentence. Afterwards, the visualization was hidden and the participant practiced the rest of the sentence using audio cues only.

The actual experiment consisted of four tasks using auditory menus and one visual task giving reference time and accuracy. The gesture interface was used with one hand. In the touchscreen experiment, the device was held in one hand and, while the other hand's finger was used for the menu browsing. The two handed method was chosen, because it was expected that some participants (e.g., persons with small hands) might not feel comfortable browsing the menu with their thumb. However, many iPhone owners use their device one handedly. The participants were instructed to select the given names aiming at maximum speed with minimum errors. The names to be searched were shown in groups of 5 on a large projector screen, as shown in Figure 6 (left). Before every task the participants were allowed to rehearse the interaction method using a set of 5 names, which remained the same for all methods. The practice time was again restricted to 5 minutes. In the experiment, 11 slides containing 5 names were used. The next slide was revealed right after the last name of the previous slide was completed. The participants were instructed to carry on to the next name, even if they made a mistake. The names were Finnish first and last names.

The whole experiment including the training periods lasted about 1.5 hours. Throughout the main experiment the participants walked in a figures-of-eight around 4 chairs placed across the room, similar to the method used in [11, 1]. The spaces between chairs were 2, 1, and 2 meters, as seen in Figure 6 (left). The participants were instructed to keep a steady pace when walking. They were reminded to continue walking if they stopped for some reason during the experiment. We designed the experiment so that we would have exact reference time and accuracy from an existing visual application. The experiment did not contain a noisy environment or other unexpected events comparable to oncoming pedestrians in the real world. The aim was not to prove the general validity of the auditory menus, which have already been extensively studied. Instead, the mobile experiment was made under conditions that reveal if the presented interaction methods suffer any setbacks when the user is moving during which, for example, the hands could be shaking.

5.4. Design

The experiment was a simple factorial design, in which five different interaction methods were tested. The used auditory menus are explained in Section 4. The 5 methods were:

- *Reference* (Ref), the normal contact list of the iPhone without any auditory feedback.
- *Touchscreen one-layer* (T_1L), the eyes-free touchscreen input with 3D audio output. All names were directly accessible in one menu layer, see Figure 5.
- *Touchscreen two-layers* (T_2L), the eyes-free touchscreen input with 3D audio output. The subject selected first the first letter and then the name from a submenu, see Figure 4.

REDUCING REPETITIVE DEVELOPMENT TASKS IN AUDITORY MENU DISPLAYS WITH THE AUDITORY MENU LIBRARY

Parameswaran Raman, Benjamin K. Davison, Myounghoon Jeon, and Bruce N. Walker

Sonification Lab
Georgia Institute of Technology
430 Cherry St.
Atlanta, GA 30332
params.raman@gatech.edu

ABSTRACT

This paper explores the process of auditory menus research. Several parts are tedious tasks which must be repeated for minor changes to the experiment. Fortunately many of these parts can be automated with software. We present the Auditory Menu Library (AML), a tool for simplifying experiment construction. The AML provides a cross-platform, configuration-based turnkey solution to studies involving auditory menus.

1. INTRODUCTION

This paper explores auditory menus research from a process standpoint and describes an extensible multi-purpose library to help perform similar auditory menus experiments in the future. By reducing the costs involved with experimental development, we will expand the possibilities of research in auditory menus.

The first part of our investigation explored the various research elements of interest in auditory menus, such as submenu behavior or sound types used. This follows the footsteps of Yalla and Walker [1], but we focus on what has been explored experimentally instead of what conceptually is important in menus.

We then studied the general parameters that researchers might be interested in configuring across various sets of experiments and the steps that are required to prepare for user studies. These will give us an understanding of the breadth of study-based functionality that a supporting software tool should possess. In addition, the repetitive nature of menu design, experimental definition, and platform-specific user interface descriptions begs for a more generic, automated process that will support study replication and extension.

This report summarizes our findings and describes a new tool: the *Auditory Menu Library* (AML). This research tool has been developed specifically to aid sonification experiments in auditory menus. The final section describes the system architecture of the Auditory Menu Library, explaining its design and implementation details, with relevant examples.

2. AUDITORY MENUS RESEARCH

Menu displays in a computer provide a list of choices. When selected, the system performs a particular task. Getting to the correct selection involves an interaction between the display and the user. Finding a particular item in a menu involves a consideration stage, which can take place in the user's mind but often is also a part of the display itself. In visual menus, this is typically done with a highlight over whatever menu item the user has their mouse hovering. Therefore, a menu choice involves both a pre-selection, or hovering, and a selection, some binary action that indicates that the user has decidedly picked the hovered item. In auditory menu display *pre-selection* of menu items is the most informative step. The actual selection often does not render the full representation of the pre-selection.

Auditory menus research has featured several components: speech; non-speech: auditory icons, earcons, spearcons, and spindex; auditory scrollbar; combinations of speech and non-speech sounds; available and unavailable menu items; visuals on or off; push and pull menus; different interaction styles; hierarchical menu structures; and broad versus deep menus. This section explores each of these components in more detail. For an analysis of visual menu structures and their applications to auditory menus, see [1].

2.1. Speech

The naive auditory menu is completely speech-based, activated when the user pre-selects and selects a particular item. Specifically, the item label (the visual text) and sometimes the item's properties (such as "unavailable") are "spoken" by the system. Commercial systems that speak menus use a text-to-speech (TTS) engine. For research, high-quality TTS or human speech is often prerecorded for the menu.

Since visual menus are spatial, and since people can easily move their eyes to "scan" a menu, responses to searching for an item on a visual menu are relatively faster than for a spoken menu. Non-speech sounds alleviate this serial characteristic of speech-only approaches. However, non-speech sounds are not typically designed to fully replace speech in a non-visual auditory menu, and often the study stimuli use speech in combination with non-speech sounds.

2.2. Non-speech sounds

The first studies in using non-speech sounds in auditory menus were attempted in the form of earcons by Brewster in particular [2, 3, 4, 5]. Walker et al proposed a speech-based sound called Spearcons that could be used to improve navigational performance in auditory menus [6]. This was followed by the conception of another concept called Spindex [7] which provides auditory equivalents to the visual concept of bookmarks used in a telephone direc-

Experiment	Device	Prog. Env.
Spearcon and TTS [12]	Desktop	Director, mobile
Spearcon and TTS [6]	Cellphone	Java
Spindex and TTS [7]	Desktop	Director, mobile
Spearcon, Spindex, and TTS, dual task [14]	Car head unit	C# Centrafuse
Auditory Scroll Bar [8]	Desktop	Flash

Table 1: Research with one-dimensional menus.

Experiment	Device	Prog. Env.
Spearcon, earcon, auditory icon, and TTS [15]	Desktop	E-prime
Spearcon and TTS [20]	Desktop	Director
Earcon and spearcon [19]	Desktop	Director
Auditory icon, earcon, spearcon, and speech [18]	Desktop	Flash & Java
Speech menu item availability [21]	Desktop Mac	Java

Table 2: Research with hierarchical menus

tory. Auditory Scrollbars were designed and evaluated by Walker and Yalla in [8]. With such a growing scope for research as evident from the previous citations, it also becomes essential to conduct exhaustive experimental analysis, evaluate and gather feedback from the user community. The rest of this section focuses on some key elements of research in the wide domain of auditory menus, with a view to gather data that would be useful to help identify the prime objectives and development goals for the AML. Yet another use of non-speech sounds in menus involves auditory icons. Gaver's auditory icons provided a way to represent WIMP[1] elements in a non-arbitrary way [9, 10]. Natural sounds could represent incoming mail, folders, or a disk drive. However, not everything in the computer can be metaphorically converted to a natural sound; what is the natural sound of "Save as HTML"? The other non-speech sounds described above sacrifice the naturalness of sounds in favor of categorical and descriptive information [11, 12].

Earcons originated as a way to provide organization information as variables in the sound itself [13]. Elements of the item's role and position in the menu structure were represented. This would arguably facilitate the user's navigation of the menu space, particularly if visuals are not available [11].

The definition of "earcon" seems to encompass any brief, organized sound that is not natural or speech-based. This simple worldview could categorize all menu sounds as having components of earcons, auditory icons, or speech. However, spearcon and spindex are hybrid concepts which provide structural information like an earcon but are rooted in a verbal interpretation of the element.

Spearcons gain some benefit of speech information while reducing the time cost of listening to speech [15]. Speech is an effective way to convey information. Spearcons attempt to maintain elements of the phonemes that would be present in the standard oral output. Spearcons are compressed speech using a type of selective sampling of the speech based on the SOLA (Synchronized Overlap Add Method) algorithm [16, 17], which produces the best-quality speech for a computationally efficient time domain technique.

Spearcons might be useful in more familiar menus, such as a personal cell phone address book, but take some time to learn. Therefore, they are typically placed along with spoken text following the spearcon. The user can skip the full speech once they are familiar with the items. The utility of spearcons has been evaluated many times [12, 14, 18, 19, 6].

Spindex is an indicator of the user's position in a sorted alphabetical menu. Typically this is manifested by the first letter of the selected word. Accordingly, for a name "John Smith", the spindex cue is /dʒeɪ/ or simply /dʒ/ (or /es/ or /s/, depending on sorting). These cues provide the user with an indication of his current position in the list and help him navigate to the desired item much more rapidly. Thus the spindex is most useful when going through large sorted lists [7], since the user could make an educated decision as to move up or down based on the current spindex cue.

2.3. Auditory Scrollbar

An auditory scrollbar is an analog to a visual scrollbar: as the value in the 1-dimensional range changes, the scrollbar changes in pitch. The auditory scrollbar can be designed in four different ways - using single tone, double tone, proportional grouping and alphabetical grouping [8]. The pitch polarity is adjusted based on these approaches. Auditory scrollbars may be useful in menus like a Font menu that has hundreds of fonts listed in alphabetical order where the user cannot navigate one by one. In such cases, it might also make sense to combine auditory scrollbars with other non-speech sound variants like spindex to enhance the user search. Auditory Scrollbars are, therefore, flexible enough to be used in conjunction with other non-speech sounds in order to add value to the user experience.

2.4. Combinations of Speech and Non-speech sounds

The sounds mentioned in the previous sections are often combined in research studies. Spearcons, for example, are typically combined with text-to-speech (TTS), so that a novice can still use the TTS, while someone familiar with the list can leverage the spearcon for quicker selection (for example, in [12, 14]). This brings up several considerations:

- Is there any temporal overlap of items? Typically they are presented serially.
- What sort of interval (quiet gap) should be placed between the sounds?
- In what order will the sounds play?
- Should the display change with experience?
- How should menu item properties be represented?

Quite often, studies are comparing one approach to another, so ensuring an apples-to-apples ordering and interval is a critical aspect of the study.

2.5. Intervals and Ordering

An auditory menu item is always likely to be supported by a diverse set of non-speech sounds like spearcon, spindex, TTS and auditory scrollbar. Since there is a possibility that researchers might want to test all of them together, it is very important that

[1] Windows, Icons, Menus, Pointers interface.

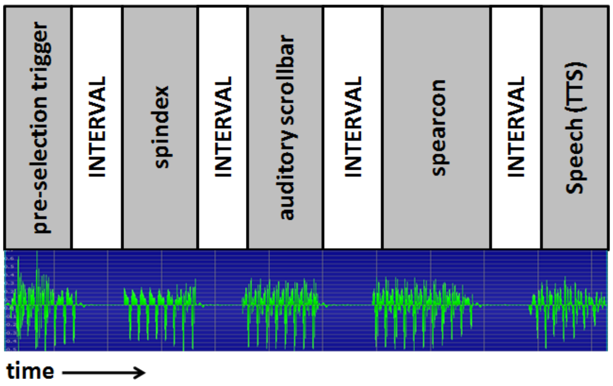


Figure 1: Order of non-speech sounds.

they conform to some standard order in which they are played. For instance, using a spearcon, spindex, TTS and auditory scrollbar on the pre-selection of a menu item might have an order as shown in figure 1.

There can be notable intervals between each of these sounds which is a value that is configurable. Varying this interval value and sometimes the order might lead to interesting results in auditory menus.

2.6. Unavailable menu items

Auditory menus tend to add information such as *unavailable* or *dimmed* to the menu items which are disabled. An immediate example for this is the screen reader software that ships with the Macintosh operating system. However, there are also other rendering possibilities such as using a lower voice or a whisper voice. One can also switch between male and female voices to evaluate their respective effects on menu usage and performance [21].

2.7. Visuals On/Off

Visuals On/Off is an essential experimental setting for almost all the auditory menu experiments which allows the users to get a real-time context of how auditory cues help them where no visual cues are present. Often, when the visuals are on, we tend to rely on them and not focus on the auditory elements fully. Having a way to turn visuals off in almost every auditory menu experiment is thus crucial for better results.

2.8. Push/Pull Menus

Another useful way to classify the auditory menus could be push menus and pull menus, which are studied in context-aware systems (e.g. [22, 23]). These can plausibly be applied to auditory interfaces as different ways of deriving information from the menu items. In the case of push menus, the menu keeps reading out the menu items (using TTS along with the non-speech sounds and other auditory cues) in a specified order; it loops through all the menu items. The user can then select the desired menu item when it is being spoken. This is different from the more common pull menus, where the menus items are played out to the user based on navigation (eg: cursor button presses). Or, in other terms, the user decides which menu items to pull out. In our experiments, we make use of these two variants very frequently and hence, need a way to represent this configurability.

2.9. Interaction Styles

Auditory Menus have been tested on the desktop [12], on mobile phones [6], and with in-vehicle navigation systems [14]. These experiments involve users searching through a one-dimensional menu list (such as a phonebook or MP3 song list) using up and down arrow keys for a particular target name. However, since many smart-phones today use advanced interaction styles such as flick, finger-gestures, tap and wheeling on a touch screen device, the research community should consider how these interaction styles affect the auditory display of menus. There should also be an easy way to customize them as per needs and switch from one style to another for demonstration purposes.

2.10. Hierarchical Menu

Hierarchical menus are more complex to handle because of the variety of information that they contain. Some properties of hierarchical menus include:

- Number of items in a menu
- Available/Unavailable state of the menu item
- Accelerators/Hot Keys for the menu item
- Grouping of menu items and Separators used
- Type of menu item (does it invoke a dialog, is a sub-menu or top level-menu?)

The key challenge therefore is to incorporate all of these characteristics successfully into auditory menus. More importantly, being able to mix and match these features and test them across different platforms and devices is essential for edging closer to practical auditory menus.

2.11. Broad versus Deep Menus

The use of particular displays in a menu may depend on the menu structure. One basic division in menu structures is broad versus deep. There have been many explorations into broad versus deep visual menus [24, 25, 26], and some in auditory menus [27]. Jeon and Walker suggest that very broad menus have not been typically considered [7], and their spindex solution is similar to a visual analog of a visual letter of the current item appearing on the screen, as seen on some iPods when scrolling through long music lists. Changing the display of different structures of menus, visual or auditory, is an open area of study.

3. AUDITORY MENU RESEARCH CHALLENGES

Based on the review above, there are several elements that appear important to a research-oriented developer of auditory menus.

- There are many auditory menu sound types. Selection and generation of the appropriate comparisons of sound types is a key component to testing them. This includes combinations of sound types including intervals.
- There are many menu properties, such as availability, accelerator, and submenus. While most of the research considers only the structural and functional role of a menu item, many

other properties are commonly used in real applications and need to be considered.

- Menu structures can be very broad (i.e., 1 level of 1000 songs), or fairly deep (e.g., 5 levels of settings with 2 to 8 items in each category). The structural definition of the menus is a key component to describing the results of research, since the structure determines, in part the ideal human performance of the task. In addition, certain auditory menu sound types are designed with a particular type of menu in mind. Intuitively a hierarchical earcon is designed for a deep hierarchy, while a spindex succeeds at long, single-dimension lists.

3.1. Roles

Auditory menus experiments require the following roles:

- an experiment *architect* who designs the study. The requirements are made by this role.
- a system *developer* who produces the software framework in which the study is run.
- an *experimenter* who actually runs the experiment.
- a data *analyst* who determines what the study shows.

Multiple roles may be held by a single person, such as architect and analyst. It appears that the system developer is often a different person from the experiment architect, who probably has a more deeper insight into human factors and usability. Therefore, communication of experiment goals is a critical component in attaining what is needed.

System development is necessary but arguably it doesn't need to be as complicated as it typically is. For example, if an experimenter needs two studies, one with spearcons and one with earcons, the menu structure and logging system along with the rest of the system can basically be the same.

3.2. Non-developers

The objective of building the library for auditory menu experiments is to essentially bypass a need for a programmer. Most of the menu definition is configurable. The average time required by a person (not well-versed with programming) to create a menu activity capture program, configure log files, traces and reports to capture the results, and then run the trials should be as minimal as possible. For instance, if the experimenter needs to add an auditory scrollbar sound to her auditory menus, it should be as easy as adding a line in a configuration file.

Currently, most of the tools developed are tightly bound to the developers. For example, an application showcasing auditory menu concepts on a Nokia phone with a phonebook list of items cannot be easily tweaked by changing the names and adding new sounds to it by a non-programmer, because this requires source code modification, compilation, packaging, porting and so on. At best, the names can be configured in a separate file, but the experiment cannot be then run on the desktop or a different cell phone platform. In addition, development in the same platform and programming language may be done in parallel because of lack of understanding of another's code base. This division leads to highly repetitive system designs. Most of the systems described in Tables 1 and 2 were made independently of each other, not sharing a common basic structure. As a result, there was no feedback cycle and similar amount of effort in terms of design and programming was spent in both of them. This could have been avoided by sticking to a generic framework that is reusable over and over again.

Be it Macromedia Director, Java or in-vehicle applications, a change in the program environment to conduct a user study seems a tough job for a non-programmer. Also, it doesn't make sense to keep learning new programming languages just for the sake of modifying the application for re-use. It is unnecessary overhead and time consumption.

3.3. Programmer Effort and Time

"Why do something in two days that will take two months to automate?" Ever since the inception of the ideas around non-speech sounds, several applications were developed to test the usability and evaluate the concepts developed. However, each time a new application was developed from scratch; this resulted in duplication of programmer effort, often with similar pieces of code being rewritten. There is code redundancy and unnecessary delays every time an experiment is performed. Instead of evaluating entirely new systems, developers can be put to use by adding concepts and platform support to a larger auditory menus tool.

3.4. Multiple Devices and Platforms

Beyond accessibility for the visually impaired, auditory menus may provide extra help on platforms with lower visual space and attention than a traditional desktop interaction. Auditory menus have been evaluated on the desktop, in mobile devices, and with in-vehicle systems. Thus, simulations need to be designed on each of these several types of platforms that demonstrate the features available and help the users evaluate them. This is again a repetitive task because though the research questions might be similar, the particular software implementations vary from desktop to a mobile device and to a vehicle (particularly in the user interface). In addition, a replication will work best if a similar process and identical data and stimuli are used. It would be impressive if there were a way to reduce this implementation time and if researchers could have a base toolkit that would expose much of the generic features, which have to be merely extended with minimal effort.

3.5. Replication of Experiments

A research tool should support the replication of a past study. By separating the data from the user interface, different researchers over different places and times can run the same experiments if they have the same data and environment set up. Building auditory menus involves putting in pieces of code, wav files, adding configurations, etc. There needs to be a way to efficiently represent the auditory menus in its entirety without any loss of information and experimental data, and migrate this to any other device. For example, once you create a hierarchical auditory menu on a Macintosh notebook, you should be able to save it in some text file format on the hard drive and use it to load the menu back on a mobile device. It is reasonable to aim for a stable state where non-programmers could store auditory menus simply in the form of plain configuration files and play with it to generate different types of menus, add hierarchies to them, modify the sounds from one type to another, and initialize their experimental settings.

3.6. Existing Toolkits/Libraries for Auditory Menus

Very little prior work has been in this direction to build a toolkit application to aid research auditory menus in particular. Mynatt and Edwards created a process and a software tool called MERCATOR designed to map graphical user interfaces to equivalent auditory interfaces [28]. This work essentially describes how an accessibility tool like JAWS creates accessible spaces. However, Mynatt and Edwards's was a much broader study not covering the specifics of auditory menus. Likewise, Brewster [29] and Davison and Walker [30] provide audio toolkits that extend user interface libraries with the concept of sound. Again, their focus was more general than specific auditory menu design. In addition, these tools are intended for use by end applications, and not specifically a research study, so considerations such as latency, logging, and quality of stimuli are different. There is also need for a data structure to represent an auditory menu of typical hierarchies and make it used in experiments. Thus, the key jobs of Auditory Menu Library are to automate the repetitive task of system construction and to model menu structure. It helps reduce the time between wanting a system and producing it. The next section discusses the system in more detail.

4. AUDITORY MENU LIBRARY

The Auditory Menu Library (AML) is a generic library currently in Java that helps its users represent the concepts of auditory menus using a data-structure, replicate them across diverse platforms and use them for experiments in a much more efficient manner. The AML has been designed considering the varying programming capabilities of major stakeholders involved in auditory menus research. Experiment Architects could just interact with its data and use it for customization to suit their experiments. On the other hand, a System Developer could aim to leverage its Application Programming Interface (API) to define more complex structures. The AML is defined as:

- A YML (a human-readable configuration file) document definition that describes the structure of the menus and settings to be used within the program. This defines the scope of acceptable YML files: those that can be turned into auditory menus by the AML.
- Menu objects that define the menu structure.
- A YML parser that converts a YML file menu definition into the menu objects.
- A hook for user interfaces to turn menu concepts into visual and auditory menus.
- A starter library that has converted the menu concepts into menus on a device like desktop computers and mobile phones.

4.1. Development Goals

There were several goals in creating this software. First, it would be transparent to a novice programmer. Other than working with configuration files, there is little he would have to do to bring up an auditory menu of his choice. Creating another modified menu would be as trivial as copying fragments of an existing menu representation and modifying its item names and other parameters.

Second, for the programmer, it offers methods that he could use to create menus, menu items, menu hierarchies and pack them appropriately into any structure he desires. He could also choose to add auditory information to it, or even extend it to incorporate a new research concept. In addition, and third, the library is extensible in that new menu components can be incorporated into it by relating them to the generic menu types. It is designed as a set of abstract Java classes which can be extended to suit the evolving requirements. This is clearly a one-time programmer effort, after which the architect, experimenter, or analyst could leverage its benefits. For example, defining a nearcon involves the following steps: Create a nearcon concept (Java class), explain programmatically how a nearcon works, and add nearcon parsing to the YML parser. What *doesn't* need to be done is a redefinition of the entire user interface, menus, or other audio interactions.

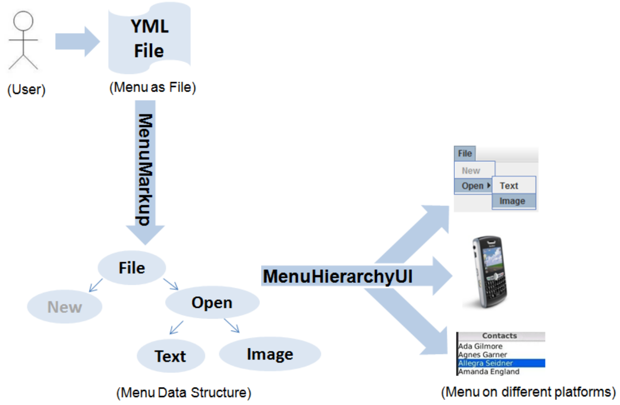


Figure 2: Process of creating menus using AML.

Fourth, the library is robust and portable. The architecture has been made so modular that the menu markup and menu model are totally loosely coupled. This helps since, a programmer would want to literally markup any menu that he has modeled and vice-versa without any glitches. Java was selected as the language since it has a strong cross-platform appeal on desktops. Current work is expanding support to different mobile devices. If a device isn't supported, the developer only needs to create the user interface explanation; the menu concept and YML files remain supported by the current AML.

Fifth, the library has the ability to represent most of the experimental scenarios in auditory menus with an emphasis to reduce developer intervention. As the science of auditory menus progresses and grows with more advanced features, this library can adapt itself and reflect all of those concepts as well.

To summarize, our focus is on the nuts and bolts of building the experiment, including:

- The roles of study development.
- Tools available for each role.
- The software structure that needs to be in place to show the menus and to log data.
- The stimuli creation.
- Log file structure: sufficient information, portable to other programs for analysis.

4.2. Architecture/System Design

The AML has been designed with the objective to achieve the above mentioned development goals and facilitate easy modifications to it in the future. What this essentially means is that all the components of the AML should be as loosely coupled to each other as possible. This would ensure that enhancements made in one do not negatively affect the others. This section discusses many of the key components.

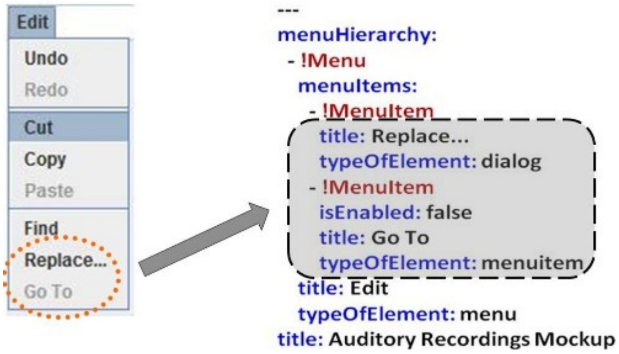
4.3. Menu Model

The AML deals with a variety of platforms and kinds of menus. It is therefore natural for the menu components to be diverse and different. The menumodel is a component that provides the ability to design new hierarchy of menus and represent them as auditory menus.

- **MenuType**
  MenuType is a base class (abstract Java class) from which the basic properties could be inherited to build customized and newer menu components. In other words, the MenuType defines the basic structure and the sub components under it define the functionality. This kind of design also facilitates better abstraction, as we could push the more generic features to MenuType, thereby maintaining the overall abstraction provided by menumodel. The underlying goal therefore is to extract the common features of all menu components and abstract them into this class. This prevents similar behavior from being treated multiple times amongst menu components. For instance, a menu item could be present either as *available* or *unavailable*. The same applies to a menu as well. Therefore, this common feature could be pushed up to MenuType that collects all commonalities.
- **Menu**
  Menu is a component derived from MenuType which represents an auditory menu. It is a collection of items of MenuType. This can be used for top-level menus, sub-menus, contacts list as seen in a mobile device and so on.
- **Menu Item**
  MenuItem is a component derived from MenuType which represents an auditory menu item. It defines most of the auditory menu properties and represents the user actions on selecting them.
- **MenuHierarchy**
  MenuHierarchy contains the menu structure defined in the menu model and is composed of a tree of menus and menuitems. It can be visualized as a single package that bundles all your auditory menu information and it can be then viewed on varying devices/platforms. Once a MenuHierarchy is built using the AML API, it becomes more convenient thereafter to edit it by just modifying the YML file representation of it.

4.4. MenuHierarchyUI

MenuHierarchyUI defines a way to render the MenuHierarchy into a UI specific to the device/platform. This exposes a method named *buildUI()* to accomplish the translation. For example, in the case of a desktop application, we could derive a specific implementation of MenuHierarchyUI termed as *SwingHierarchyUI* which uses Java Swing libraries to render the MenuHierarchy on the screen. Likewise, *BlackBerryMenuHierarchyUI* can be used to put the MenuHierarchy on a Blackberry phone screen. This is important because situations would be different on a mobile platform and we would need to use the J2ME specific UI libraries. Our aim is therefore to insulate the device specific technology from the core concept of auditory menus (which remains the same everywhere).



```
---
menuHierarchy:
 - !Menu
  menuItems:
   - !MenuItem
    title: Replace...
    typeOfElement: dialog
   - !MenuItem
    isEnabled: false
    title: Go To
    typeOfElement: menuitem
  title: Edit
  typeOfElement: menu
title: Auditory Recordings Mockup
```

Figure 3: YML stores other meta-information about the menuitem such as typeOfElement, enabled state,etc .

The SwingHierarchyUI is intended for desktop user interfaces, specifically desktops that use the Microsoft Windows, Mac OS, or Linux varieties of operating systems.

4.5. Menu Markup

Menu Markup converts the YML files into their equivalent MenuHierarchy objects so that they can be displayed on different screens and devices.

- **MenuParser**
  MenuParser is a generic class used to read the YML representation of a menu structure and create a MenuHierarchy out of it. This MenuHierarchy could then be rendered based upon user's choice. MenuParser offers functions like parseFileIntoObject() to accomplish this. This could be inherited by several other classes like StandardMenuParser which could be used to parse an XML file into MenuHierarchy, or XMLMenuParser to parse an XML file.

4.6. YML as a Markup

YML has been adopted as the choice for markup over other commonly used languages like XML, primarily because YML is more human-readable with less meta-information and other unwanted data (refer Figure 3). This contrasts with an XML representation, in which most of the file space is consumed by opening and closing tags. YML's familiar indented outline and lean appearance makes it especially suited for tasks where humans are likely to view or edit data structures, such as configuration files, dumping during debugging, and document headers. It is extensively used in languages like Ruby and Python for storing user configuration.

4.7. Use Example

For example, the following code describes the process of constructing a MenuHierarchy named *Mockup*. As the very first step MenuHierarchy object *hierarchy* is constructed, by passing in a name and the visible state of the menu. Second, a *File* menu is created by adding two menu items *New* and *Open* to it. Third, the top level menu is added to the MenuHierarchy object *hierarchy*. AML provides numerous overloaded constructors for the programmer to set selective properties for the menu and menu items like enabled state, visible state, accelerator, etc.

The newly constructed MenuHierarchy object hierarchy, can either be saved as a YML file using the *dump()* function exposed by Yaml, or displayed on a user interface by using a custom MenuHierarchyUI object so that it can be heard or visualized. SwingHierarchyUI for instance, is a Java class that implements the *buildUI()* method of the MenuHierarchyUI abstract class, in a way so as to render the hierarchy object using Java Swing libraries on a desktop.

```
MenuHierarchy hierarchy = new MenuHierarchy(”MH”, true);
Vector menus = new Vector<MenuType>();
menus.add(new MenuItem(”New”, true, ”menuitem”));
menus.add(new MenuItem(”Open”, false, ”menuitem”));
hierarchy.addMenu(new Menu(”File”, menus, ”menu”));
Yaml.dump(hierarchy, new File(”FirefoxMenu.yml”), true);
SwingHierarchyUI swingUI = new SwingHierarchyUI();
swingUI.buidUI(hierarchy, true);
```

4.8. Salient Features

The library is extensible: a programmer can develop new menu components by extending the generic MenuType provided and adding specifics to it. This applies to other parts of the library as well like the MenuParser which can be extended based on the type of markup language followed.

Being developed in Java, AML is portable across most desktop platforms (Windows, Linux, Macintosh) and also can be used in Java-based mobile platforms such as Google Android and RIM Blackberry.

The greatest strength of AML is its support for configuration of its features. Choosing to turn visuals on or off, specifying which menu items are enabled or disabled can all be done by modifying the YML representation of the auditory menu.

Programmer intervention is required only if a new platform is encountered and AML needs to be extended to support or add features specific to it. Even that is a one-time task.

AML can be also be viewed as an effective research tool that helps you organize your experiments quickly and neatly. It provides support for:

- **Logging.** While running any experiment, there is an enormous amount of information a researcher might want to log about the user and his interaction with the tool. This could be response time, keystrokes, pattern of navigation, menu operations performed, etc. AML lets the person running the experiment choose the data to log and generates periodic log files which can be later examined for more details.
- **Debug Dumps.** These are files containing description about the auditory menu, its sub-menus, inner details, their state before the application crashed and so on. This turns out to be useful while troubleshooting.
- **Reports.** Data reports are essential to be generated for some experiments like auditory menu experiments on handheld devices to study how they perform with different user populations. Specifically, understanding how a new user adapts to auditory menus, measuring the change in his response time and his learning rate could help foster research in a greater magnitude. AML aids in capturing such information by tracking the user interaction.

4.9. Issues Faced

Several issues were encountered in the design and development of AML. It involved tremendous effort to make things generic bearing the various platforms in mind and devices that auditory menus might have to be used on. Even on using Java, numerous JDK version compatability and class file version issues had to be resolved to make AML run across desktop and mobile environments, because not all Java API are the same on all mobile platforms. Porting AML to the Android platform was a relatively easier task than Blackberry since Android is closer to the traditional JDK development supporting latest Java constructs such as iterators and generics. Infact, the way a particular task (like reading a file from the filesystem) is accomplished differs greatly from one platform to another. For instance, during AML's development phase, it was discovered that reading a file on Android seemed almost the same as on Desktop PCs but on Blackberry it was drastically different. Thus, a lot of functionality had to be implemented in different ways on each platform (leveraging features specific to the platform). As a result of these, the AML design had to be refactored by adding more generic features and interfaces.

Several delay issues were found while migrating AML to mobile devices because of the inherent limitations of the device memory and processor speed. As a result, the feedback of the application was sometimes quite different and inferior from the one as observed on a desktop. New solutions are being implemented to address issues such as these as they arise.

4.10. Demonstration Application

A sample application was developed that made use of the Auditory Menu Library to create auditory menu structures in the form of YML files, render it on the desktop user interface and also test the effects of applying the various sound elements like spearcon, spindex and auditory scrollbar to the menus.

This demo application built into the AML displays two different kind of menus - a hierarchical menu and a linear menu, both of them built from a single YML file. The hierarchical menu consists of active and disabled menu items, sub-menus and menu items that invoked a dialog. The linear menu consists of a list of phonebook contacts to which an auditory scrollbar was attached and its combination with other sounds like Spearcons and Spindex could be observed. It is also possible to turn off features selectively to emphasize other specific ones and derive conclusions.

5. CONCLUSION

This paper provided an overview of the history of auditory menus research. It explored the repetitive challenges faced, particularly during system development. Finally, we discussed the auditory menu library, an extensible Java tool designed to support experimenter and programmer development of research-oriented auditory menus. This software can be accessed at http://sonify.psych.gatech.edu

6. REFERENCES

[1] P. Yalla and B. Walker, “Advanced auditory menus,” Georgia Institute of Technology GVU Center, GVU Technical Report GIT-GVU-07-12, Oct. 2007.

[2] G. Leplatre and S. A. Brewster, “Designing non-speech sounds to support navigation in mobile phone menus.” in *Proceedings of the 6th International Conference on Auditory Display*, Atlanta, GA, 2000, pp. 190–199.

[3] S. A. Brewster, “Using non-speech sounds to provide navigation cues,” *ACM Transactions on Computer-Human Interaction*, vol. 5, no. 3, pp. 224–259, 1998.

[4] S. A. Brewster, V. Raty, and A. Kortekangas, “Earcons as a method of providing navigational cues in a menu hierarchy,” in *Human Computer Interaction*, 1996, pp. 167–183.

[5] S. A. Brewster, P. C. Wright, and A. D. N. Edwards, “A detailed investigation into the effectiveness of earcons,” in *Proceedings of the 1st International Conference on Auditory Display*, Santa Fe, NM, USA, 1992, pp. 471–478.

[6] B. N. Walker and A. Kogan, “Spearcons enhance performance and preference for auditory menus on a mobile phone,” in *Universal Access in HCI*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2009, no. 5615, pp. 445–454.

[7] M. Jeon and B. N. Walker, “”Spindex”: accelerated initial speech sounds improve navigation performance in auditory menus,” in *Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society*, San Antonio, TX, Oct. 2009.

[8] P. Yalla and B. N. Walker, “Advanced auditory menus: Design and evaluation of auditory scroll bars,” in *Proceedings of the Tenth International ACM SIGACCESS Conference on Computers and Accessibility*. Halfax, Canada: ACM Press, Oct. 2008, pp. 105–112.

[9] W. W. Gaver, “Auditory icons: Using sound in computer interfaces,” *Human-Computer Interaction*, vol. 2, pp. 167–177, 1986.

[10] W. W. Gaver, “The SonicFinder: an interface that uses auditory icons,” *Human-Computer Interaction*, vol. 4, pp. 67–94, 1989.

[11] S. A. Brewster, P. C. Wright, and A. D. N. Edwards, “An evaluation of earcons for use in auditory human-computer interfaces,” in *Proceedings of the INTERACT ’93 and CHI ’93 conference on Human factors in computing systems*. Amsterdam, The Netherlands: ACM, 1993.

[12] D. Palladino and B. Walker, “Efficiency of Spearcon-Enhanced navigation of one dimensional electronic menus,” in *Proceedings of the 14th International Conference on Auditory Display*, Paris, France, 2008.

[13] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, “Earcons and icons: Their structure and common design principles,” *Human-Computer Interaction*, vol. 4, pp. 11–44, 1989.

[14] M. Jeon, B. K. Davison, J. Wilson, M. Nees, and B. N. Walker, “Enhanced auditory menu cues improve dual task performance and preference with in-vehicle technologies,” in *Proceedings of the First International Conference on Automotive User Interface and Interactive Vehicular Applications*. Essen, Germany: ACM, Sept. 2009.

[15] B. N. Walker, A. Nance, and J. Lindsay, “Spearcons: Speech-based earcons improve navigation performance in auditory menus,” in *Proceedings of the 12th International Conference on Auditory Display*, London, England, 2006, pp. 63–68.

[16] D. J. J. Hejna, “Real-time time-scale modification of speech via the synchronized overlap-add algorithm,” Master’s Thesis, Massachusetts Institute of Technology, Cambridge, MA, May 1990.

[17] S. Roucos and A. M. Wilgus, “High quality time-scale modification for speech,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. New York, NY: IEEE, 1985, pp. 493–496.

[18] T. Dingler, J. Lindsay, and B. N. Walker, “Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech,” in *Proceedings of the 15th International Conference on Auditory Display*, Paris, France, 2008.

[19] D. Palladino and B. N. Walker, “Learning rates for auditory menus enhanced with spearcons versus earcons,” in *Proceedings of the 14th International Conference on Auditory Display*, Montreal, Canada, 2007, pp. 274–279.

[20] D. Palladino and B. N. Walker, “Navigation efficiency of two dimensional auditory menus using spearcon enhancements,” in *Annual Meeting of the Human Factors and Ergonomics Society*, New York, NY, Sept. 2008, pp. 1262–1266.

[21] M. Jeon, S. Gupta, B. K. Davison, and B. N. Walker, “Auditory menus are not just spoken visual menus: A case study of ”unavailable” menu items,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI Work in Progress)*. Atlanta, GA: ACM Press, 2010, p. in press.

[22] L. Barkhuus and A. Dey, “Is context-aware computing taking control away from the user? three levels of interactivity examined,” in *Proceedings of Ubicomp 2003*, 2003, pp. 149–156.

[23] K. Cheverst, K. Mitchell, and N. Davies, “Exploring context-aware information push,” in *Personal and Ubiquitous Computing*, 2002, pp. 276–281.

[24] A. Howes, “A model of the acquisition of menu knowledge by exploration,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. Boston, MA: ACM Press, 1994, pp. 445–451.

[25] J. I. Kiger, “The depth/breadth trade-off in the design of menu-driven user interfaces,” *International Journal of Man-Machine Studies*, vol. 20, pp. 201–213, 1984.

[26] D. P. Miller, “Depth/breadth tradeoff in hierarchical computer menus,” in *Proceedings of the Human Factors Society Meeting*, Rochester, NY, USA, Oct. 1981.

[27] P. M. Commarford, J. R. Lewis, J. A. Smither, and M. D. Gentzler, “A comparison of broad versus deep auditory menu structures,” *Human Factors*, vol. 50, no. 1, pp. 77–89, 2008.

[28] E. Mynatt and W. Edwards, “Mapping GUIs to auditory interfaces,” in *5th Annual ACM Symposium on User Interface Software and Technology*. Monteray, California, United States: ACM, 1992, pp. 61–70.

[29] S. Brewster, “A sonically enhanced interface toolkit,” in *Proceedings of the 3rd International Conference on Auditory Display*, Palo Alto, CA, U.S., 1996.

[30] B. K. Davison and B. N. Walker, “AudioPlusWidgets: bringing sound to software widgets and interface components,” in *Proceedings of the 14th International Conference on Auditory Display*, Paris, France, 2008.

SOUNDSCAPE MAPPING: A TOOL FOR EVALUATING SOUNDS AND AUDITORY ENVIRONMENTS

Iain McGregor, Gregory Leplâtre, Phil Turner and Tom Flint

Edinburgh Napier University,
School of Computing,
Edinburgh, United Kingdom
{i.mcgregor, g.leplatre, p.turner, t.flint}@napier.ac.uk

ABSTRACT

This paper describes a soundscape mapping tool, and provides an illustration of its use in the evaluation of an in-car auditory interface. The tool addresses three areas: communicating what people are listening to, showing how soundscapes can be visualized, and demonstrating how the approach can be used by a designer during the evaluation of an auditory display. The strengths and limitations of this approach are discussed and future work identified.

1. INTRODUCTION

Robare and Forlizzi [1] highlighted a 'lack of design theory' with regard to guidelines for sound design within computing. Despite the dramatic increase in the number of products which replay sound in the last ten years or so, there has been relatively little improvement in how we understand the listeners' experiences. Designers need to consider the context of use, as applications might be used in a wide variety of environments [2]. While product design often explores listening [3], the same cannot be said of the development of auditory displays. This is due, in part, to the relative paucity of formal techniques to measure a design's impact. Available techniques are limited to simple noise pollution measurements [4], the elicitation of interpretations from listeners [5], and 'object-orientated' descriptions [6]. The soundscape mapping tool we present here is designed to evaluate auditory displays in their intended context of use [7]. Our empirical approach to the evaluation of these displays is to study them *in situ* by first, eliciting people's auditory experiences; and then visualising these soundscapes for ease of comparison.

This paper reports the illustration of the soundscape mapping tool through the evaluation of an in-car auditory display. Our interest in evaluating this audio-only interface was in understanding the effect of different auditory contexts on its effectiveness. A small car was chosen as it represented a contained environment that travelled through more complex external auditory environments. In order to provide a consistent experience for all of the participants it was tested in a simulated environment:

- travelling through a busy city centre at rush hour with speech radio playing.
- travelling through a busy city centre at rush hour with both speech radio playing and the auditory display.
- stationary in a quiet location with only the auditory display.

By comparing the findings from these three different contexts we can be confident that both the method and tool are reliable and robust and that they yield ecologically valid results.

2. METHOD

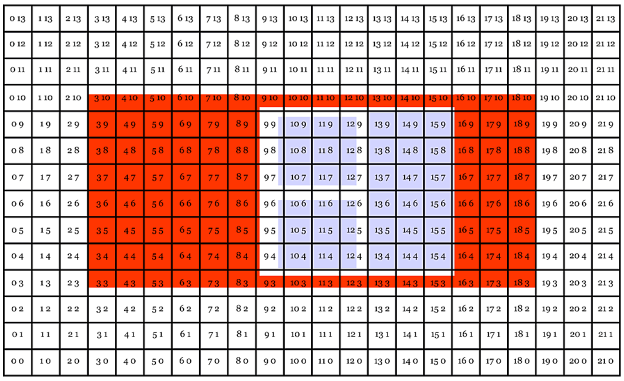
We created a tool for the classification and visualization of soundscapes, that can be used during the evaluation of augmented auditory environments. This tool is based on the results of a series of three studies. The first was an experimental elicitation of concurrent verbalizations by 40 listeners where listeners were asked to describe their auditory environment. The responses were transcribed and coded in order to discover which attributes were important to listeners when describing sound [8]. The second was a questionnaire survey completed by 75 audio professionals where they described the attributes of sound that were important to sound designers [9]. The third study was a soundscape mapping tool based on published methods where 18 listeners' experiences of a shared auditory environment (open–plan office) were compared. The tool was used to represent the experiences of individuals, as well as subsets of users (regular, intermittent and new) of the workspace [10]. The version of the soundscape mapping tool reported here has three distinct phases: capture, classification, and visualization.

- Capture involved the creation of a schematic of the car, recording the sound field, and transcribing the sound events directly from the surround sound recording.

- Classification was conducted by the participants who first listened to the recordings, and then were questioned about the audible attributes of a series of sound events.
- Visualisation involved the creation of a series of annotated soundscape maps based around the physical context of the study.

2.1. Capture

A 20:1 schematic of the car was created, cells were added to the perimeter to facilitate the annotation of external sound events (see Figure 1). A fifteen minute recording was made of the car driving through the city centre, in order to create a consistent soundfield for listeners. This recording was made using a custom eight-channel surround system and then augmented during the experiment with the auditory interface. Eight omni-directional microphones were affixed in suspension mounts inside the car, at approximately head height, and fed into four DAT recorders (see Figure 2).



0 13	1 13	2 13	3 13	4 13	5 13	6 13	7 13	8 13	9 13	10 13	11 13	12 13	13 13	14 13	15 13	16 13	17 13	18 13	19 13	20 13	21 13
0 12	1 12	2 12	3 12	4 12	5 12	6 12	7 12	8 12	9 12	10 12	11 12	12 12	13 12	14 12	15 12	16 12	17 12	18 12	19 12	20 12	21 12
0 11	1 11	2 11	3 11	4 11	5 11	6 11	7 11	8 11	9 11	10 11	11 11	12 11	13 11	14 11	15 11	16 11	17 11	18 11	19 11	20 11	21 11
0 10	1 10	2 10	3 10	4 10	5 10	6 10	7 10	8 10	9 10	10 10	11 10	12 10	13 10	14 10	15 10	16 10	17 10	18 10	19 10	20 10	21 10
0 9	1 9	2 9	3 9	4 9	5 9	6 9	7 9	8 9	9 9	10 9	11 9	12 9	13 9	14 9	15 9	16 9	17 9	18 9	19 9	20 9	21 9
0 8	1 8	2 8	3 8	4 8	5 8	6 8	7 8	8 8	9 8	10 8	11 8	12 8	13 8	14 8	15 8	16 8	17 8	18 8	19 8	20 8	21 8
0 7	1 7	2 7	3 7	4 7	5 7	6 7	7 7	8 7	9 7	10 7	11 7	12 7	13 7	14 7	15 7	16 7	17 7	18 7	19 7	20 7	21 7
0 6	1 6	2 6	3 6	4 6	5 6	6 6	7 6	8 6	9 6	10 6	11 6	12 6	13 6	14 6	15 6	16 6	17 6	18 6	19 6	20 6	21 6
0 5	1 5	2 5	3 5	4 5	5 5	6 5	7 5	8 5	9 5	10 5	11 5	12 5	13 5	14 5	15 5	16 5	17 5	18 5	19 5	20 5	21 5
0 4	1 4	2 4	3 4	4 4	5 4	6 4	7 4	8 4	9 4	10 4	11 4	12 4	13 4	14 4	15 4	16 4	17 4	18 4	19 4	20 4	21 4
0 3	1 3	2 3	3 3	4 3	5 3	6 3	7 3	8 3	9 3	10 3	11 3	12 3	13 3	14 3	15 3	16 3	17 3	18 3	19 3	20 3	21 3
0 2	1 2	2 2	3 2	4 2	5 2	6 2	7 2	8 2	9 2	10 2	11 2	12 2	13 2	14 2	15 2	16 2	17 2	18 2	19 2	20 2	21 2
0 1	1 1	2 1	3 1	4 1	5 1	6 1	7 1	8 1	9 1	10 1	11 1	12 1	13 1	14 1	15 1	16 1	17 1	18 1	19 1	20 1	21 1
0 0	1 0	2 0	3 0	4 0	5 0	6 0	7 0	8 0	9 0	10 0	11 0	12 0	13 0	14 0	15 0	16 0	17 0	18 0	19 0	20 0	21 0

Figure 1: Simplified aerial view of car with grid, red = bodywork, blue = seats

Calibration was achieved by a method borrowed from the film industry, the driver first read off the display of an SPL meter located on the passenger seat illustrating the slow sensitivity peak dbC level, then the driver clapped their hands. The short peak acted as the starting point for the recording, allowing all 8 tracks to be synchronized during the capture process. A handclap by the driver completed the recording, this confirmed whether any of the tracks had drifted during the time period. Each track was subsequently transferred to a Pro Tools LE system, in order to provide a consistent auditory backdrop for the auditory interface.

Sound event transcription included source, action, start time, end time and location. Table 1 contains examples of these. Location was calculated using the perceived central point from the surround sound recording, and notated using x-y coordinates according to the grid. If a sound event moved in relation to the car, the start and end points were documented. Start and end times were also established from the recording, these were rounded down to the nearest second within which the event occurred. In order to reduce the number of events which listeners had to classify, sound events which had the same source, action and location were grouped together.

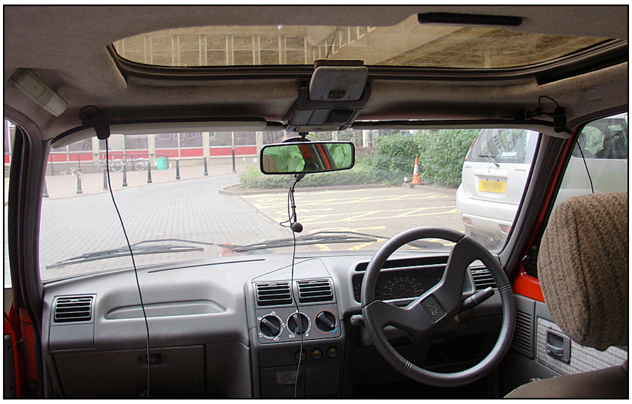


Figure 2: Microphone placement prior to final positioning and calibration, for surround sound recording

All of the captured material was passed to the designer (the second author) so that he could create the auditory display. The designer decided to limit the interface to only three auditory warnings to reduce the cognitive load on the listener. After creating the design he overlaid the new sounds on to the eight channel surround sound recording. This allowed him to control the level, incidence, duration and (perceived) spatial location of each warning. The designer also provided a written description of the different auditory warnings for the listeners' reference. These warnings included, braking distance, dead angle and email message. This final augmented version of the surround sound recording was then split into three versions, one for each simulated environment.

Event	Source	Start	End	x	y	x	y
Engine Idle	Engine	00:00:00	00:00:21	6	6		
Male Speech	Radio	00:00:00	00:00:13	11	8		
Windscreen Wiper	Car	00:00:02		21	9		
Passing Car	Car	00:00:04	00:00:06	9	13	21	13
Siren	Ambulance	00:00:06	00:00:08	21	12		

Table 1: Example sound event transcription

2.2. Classification

A classification was created based on the findings from previous studies [8, 9, 10]. Table 2 holds these ten distinct attributes each with three options. The first six attributes were derived directly from the comparison between audio practitioners and listeners. In the case of *type* rather than specify whether a source was natural or artificial, choices were confined to speech, music or sound effect, with the last representing all sounds which are neither speech or music. *Material* relates to the substance which gives rise to the sound, either gas, liquid or solid, whilst the *interaction* specifies the nature of the sound's generation whether it was impulsive, intermittent, or continuous. *Temporal* reflects the total length of the sound event (short, medium or long) separate to its interaction; *spectral* applies to its pitch (high, mid and low); and *dynamics* to its volume (loud, medium or soft).

Type	Category
Speech	Spoken language
Music	Performed composition
Sound effect	Audible events and actions
Material	Matter
Gas	Airborne
Liquid	Fluids
Solid	Objects
Interaction	Action
Impulsive	Explosion/drip/impact
Intermittent	Whooshing/splashing/scraping
Continuous	Blowing/flowing/rolling
Temporal	Duration
Short	Brief
Medium	Neither long nor short
Long	Extended
Spectral	Pitch
High	High pitch/frequency Treble
Mid	Medium pitch/frequency Alto
Low	Low pitch/frequency Bass
Dynamics	Volume/Loudness
Loud	High volume *Forte*
Medium	Medium volume/level
Soft	Quiet *Piano*
Content	Relevance
Informative	Relevant information
Neutral	Neither relevant nor irrelevant
Noise	Irrelevant/unwanted
Aesthetics	Beauty
Pleasing	Beautiful
Neutral	Mediocre
Displeasing	Ugly
Clarity	Quality
Clear	Easy to hear and comprehend
Neutral	Neither easy nor difficult to hear
Unclear	Difficult to hear and comprehend
Emotions	Feelings
Positive	Acceptance, Anticipation, Joy, Surprise
Neutral	No emotional content
Negative	Anger, Disgust, Fear, Sadness

Table 2: Sound event classification

Establishing whether a sound is informative within an auditory interface has always been important [11], and here the *content* is classified as informative, neutral or (just) noise. Noise being defined, in this case, as an unwanted or undesired sound, rather than unpleasant [12].

Barrass and Frauenberger [13] referred to the importance of the balance which must be struck between the aesthetic and the informative when creating an auditory display. Our earlier work has also been shown that a sound's aesthetics are integral to its functional effectiveness within an auditory display [14]. For this study our treatment of *aesthetics* has been to reduce them to pleasing, neutral and displeasing, rather than the more commonly used terms of harsh, warm, or bright (the latter terms being rather esoteric and requiring 'critical listening skills' [15]).

Clarity applies to the intelligibility of a sound and is rated according to whether it is clear, neutral or unclear, although in professional practice it is normally described as either poor or good. *Emotions*, which in this case are considered in terms of positive, neutral or negative, are not normally associated with sound design, although Johannsen [16] argues that if a sound has been 'well-designed' appropriate emotions should be evoked.

For this small illustrative study, 10 volunteers from the staff and students within the University participated. Each of the participants was familiar with the inside of a car and with driving, and had no known hearing impairments. Each candidate sat in the centre of eight compact loudspeakers and four sub bass units (see Figure 3). Each speaker location corresponded to the equivalent position of an omni-directional microphone during the recording. This ensured that all of the timings for the audio cues remained consistent, making it a more accurate spatial representation of the interior of the car.

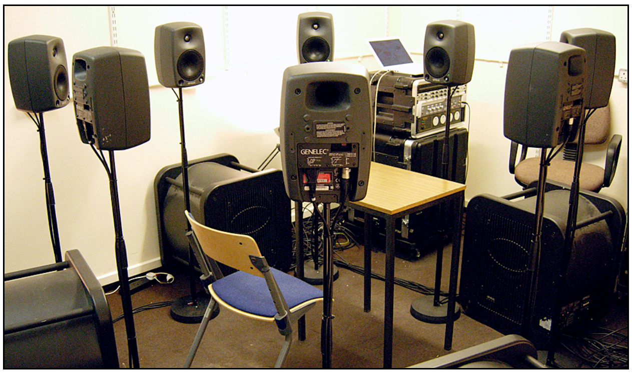


Figure 3: Surround sound reproduction apparatus

Each listener participated individually. They were first asked to read a set of guidelines and invited to ask any questions that they might have. They then listened to the three sounds created by the designer while consulting the printed descriptions. Whilst this meant that that the listeners were primed, which created a risk of a higher rate of recognition, it was necessary for them to have an understanding of the meaning of the sounds as all of other sound events were potentially familiar. The presentation of the second, third and fourth recordings were pseudo-randomised in order to help mitigate the effects of fatigue and the learning effect. After the first sequence participants were asked to use the outline of the car (overlaid with a grid) to record and classify their experiences. Participants were questioned after the replay of the recording so that their responses closely reflect what they had been listening to. Once all of the responses had been elicited, descriptive statistics were applied to them. Aggregated coordinates were derived by using a median rather than a mean, so as to reduce the effect of outliers skewing the data. We also adopted the heuristics that if 50+% of the subjects were aware of a sound event then it was included in the combined map.

2.3. Visualization

Servigne *et al.* [17] have suggested that 'graphic seminology' would be appropriate for displaying sounds, proposing that smiling faces overlaid onto a map could be used to display participant's preferences. And in this spirit we created a set of symbols in order to visualize the listeners' experiences. These symbols may be found in Figure 4.

Each sound event was given a code by the first author and the combination of shapes, colours and symbols were overlaid onto the grid according to the x-y coordinates provided by the participants. If two or more sound events had identical coordinates then they were spaced evenly across the cell so that they remained visible. For ease of interpretation the grid, numbers and interior of the car were removed. The outline of the car was retained in order to provide some indication of orientation and scale.

Sound *type* was represented through either: a series of letters for speech, quavers for music, or a loudspeaker symbol for sound effect. The material was illustrated through the border colour, cyan magenta and yellow (CMY) which were applied to the spectral representation. This allowed colour values to be absolute in both printed and onscreen forms. The interaction was depicted using border dashes, impulsive had short dashes, whilst intermittent had longer, and therefore fewer dashes, whilst continuous was a single dash with no gaps. This approach was chosen so that it visually suggested the length of the sounds' interaction. *Temporal* attributes represented using a fill gradient, a radial gradient was used to suggest a short event, which visually is associated with a droplet falling on to a liquid. A medium event was portrayed with a linear gradient which suggested a more gradual change, and a long event was a solid colour which implied that there was either none or minimal change. The gradient started with the spectral fill colour and then progressed to a pure white and then back to the original fill colour. Fill colour was used for the spectral attribute, red was used for high, green for mid and blue (RGB) for low following the practice of auditory professionals [18].

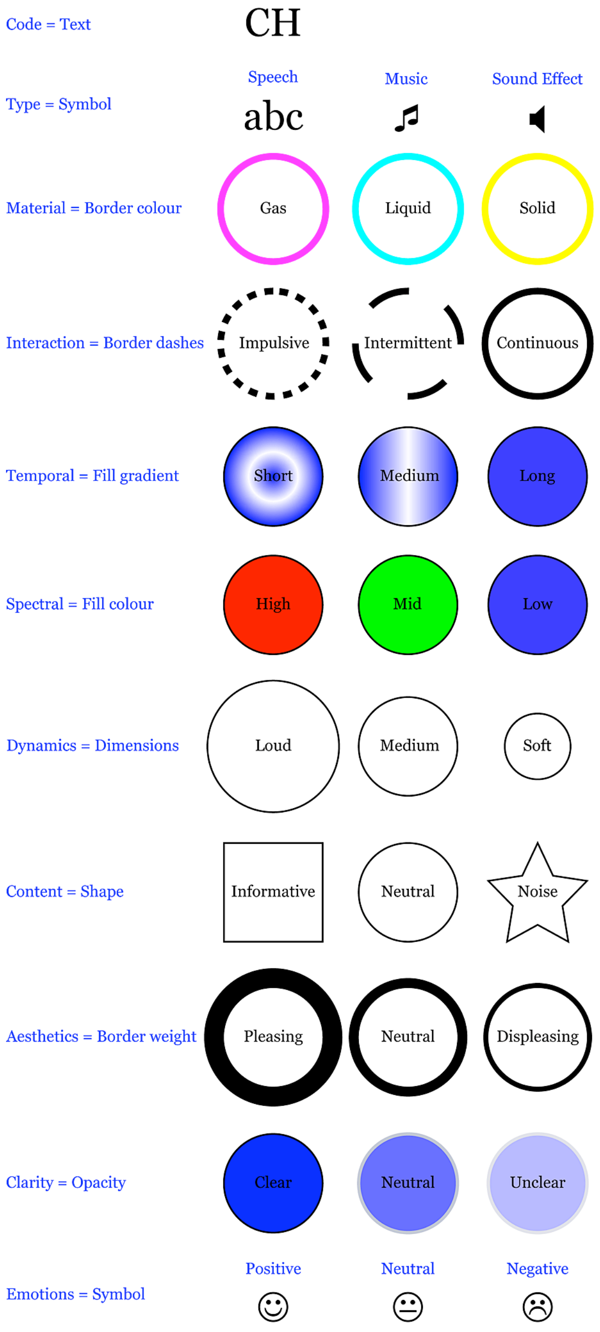


Figure 4: Visualisation key

Dynamics were illustrated using the scale of the shape, a soft sound was half the size of a medium one, and a loud sound event was 1.5 times the size of the medium and three times that of the soft. A square was used to signify informative, a circle for neutral and a star for noise. The three distinct shapes do not share any stroke angles, making it easier to differentiate between them when sound events are overlapped. Aesthetics were denoted by border weight, pleasing was represented with a thick line which was double the width of the neutral and four times the size of the displeasing. The *clarity* of a sound event was shown through the opacity of the shape, clear (=100% opaque), neutral (=66%) and unclear (=33%). Finally, 'emoticons' were used to represent positive emotions (a smile), neutral for neutral and a frown for negative.

3. RESULTS

The recording was relatively simple to transcribe, participants appeared to find the sound events straightforward to classify. The visualisations yielded informative results that showed clearly what participants were listening to.

3.1. Capture

Within the five minutes of audio recording 157 separate audible events were notated, these were identified as having been generated by 49 different sources. Sources such as the car's radio generated more than one type of sound event, so by grouping together sound events according to their source and the event it was possible to reduce the total down to 65. This was augmented by the designer with a further 3 sound events which were grouped together as a simple auditory display.

Sound events were generated from the car under study (28), passing vehicles (28), people (5), a dog and some scaffolding. Within the car, the engine passing through different states (idling, accelerating, cruising and decelerating) was recorded, as well engaging and releasing the handbrake, changing gear and a wide range of vibrations. There were 11 distinct types of sound from the radio, these were split into speech, music and laughter. Outside of the car 27 different vehicles were noted along with a siren, vehicle passes, brake squeals, indicators and windscreen wiping. The remaining sounds included screaming, talking, rustling of clothes, barking and scaffolding being struck.

Regarding the spatial cues, all of the sounds associated with the car could be identified to specific points within the outline of the car. The majority of the passing vehicles were located on the driver's side, which is at the top of the map, whilst most of the stationary vehicles were found to the rear of the car, which corresponds to the right hand side of the illustration. There were few sound events on the passenger's side and in front. The discrepancy to the paucity of sound events on the passenger's side can be partially explained by the comparatively low level of sounds on the pavement, when compared to the much louder vehicles. The shortage of audible sound events at the front of the car is most probably due to masking associated with the car's engine, which was constantly running throughout the recording.

This list only represents what could be heard on the recording, many more sounds would have been present but were either masked or inaudible due to the method of capture. All notes were made listening to the multi channel recording at the original sound pressure level, rather than over amplifying to enhance barely audible sources. This was done so that it replicated the conditions of the original journey as well as the reproduction levels which participants would have experienced.

3.2. Classification

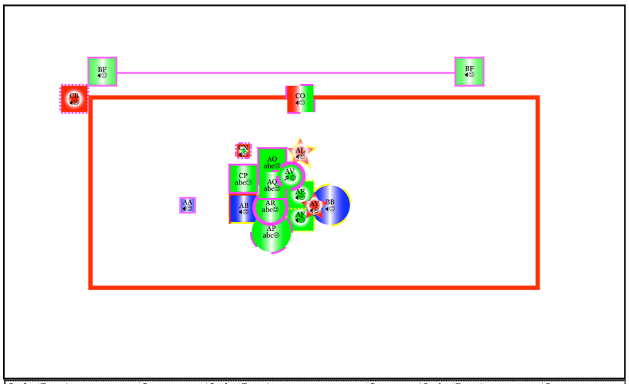
Participants were aware of an average of 30% of the sound events with a range of 38% to 21%. An average of 25% of all of the sound events from the car were heard by the participants the first time they heard the recording compared to 29% for the second. With the auditory display, the average was 94% for the first exposure, compared to 91% for the second, which might be due to habituation, but the difference is too small to draw conclusions from.

Overall there was a high level of awareness for the sounds associated with the car's engine and its handbrake, whereas the other sources such as internal vibrations, and indicating went comparatively unnoticed, except for when all of the wheels passed over a bump together. On the radio the first male voice was discerned, whereas the second, and its associated chanting, was missed. Two out of the three female voices, again on the radio, were identified, as was the interference from a mobile phone, but only one of the pieces of music was attended to. The group laughter was also generally missed, despite being the last thing that was present on the recording. Only two passing cars, and one passing bus were detected, which participants partially explained by the overwhelming urge to listen to the conversation from the radio, even when they were experiencing the identical content for a second time. When listening to the three sound events from the auditory display all of the participants were aware of all of the sounds. When they were listened to in context, then four out of the ten no longer recalled the braking distance cue, and even the designer was unaware of it, despite having added it into the recording himself.

Listeners found it hard to accurately recollect where a sound originated, but were much more comfortable with its orientation in relation to their listening position, although there were the occasional front to back errors. This is not surprising as problems with spatial discrimination are well documented, particularly when the source is not directly in front of the listener [19]. For the classification as a whole there was an average consistency of 80% between individual attributes, with a range of 67% - 98%. The average response of the ten participants was also compared to the combined classification which showed that apart from the *interaction* there was a good level of correspondence between the two sets of figures.

3.3. Visualization

A total of 36 maps were created. Each participant provided classifications for three maps, the car on its own, the isolated auditory display, and the car augmented with the auditory display. The aggregated (combined) classifications were also mapped in the same manner as the individuals' (see Figure 5). In addition it was possible to create a fourth map which represented the auditory display as experienced in context, but isolated from the auditory backdrop.



Code	Event	Source	Code	Event	Source	Code	Event	Source
AA	Engine Idle	Engine	AP	Male 2 Speech	Radio	BF	Passing Car	Car 2
AB	Engine Accelerate	Engine	AQ	Female 1 Speech	Radio	CB	Horn	Car 14
AE	Handbrake Released	Handbrake	AR	Female 2 Speech	Radio	CN	Braking distance	Auditory Display
AF	Handbrake Engaged	Handbrake	AV	Music 1	Radio	CO	Dead angle	Auditory Display
AL	Seat Creak	Driver's Seat	AY	Mobile Phone Interference	Radio	CP	Message	Auditory Display
AO	Male 1 Speech	Radio	BB	Bumps	All wheels			

Figure 5: Visualisation of soundscape for car and auditory display by combined participants

Only sound events which participants stated that they were aware of were included on the maps, otherwise they were omitted. An issue arose when sound events occupied the same coordinates. If their clarity was classified as being neutral or unclear then it was possible to overlap them quite tightly, whilst ensuring that the relevant information was still clearly visible, this was due to their partial opacity. But if all of the sound events were considered to be clear, and therefore opaque, then the amount of overlapping was minimal, as any area that was occluded was therefore no longer visible. Whilst this created problems with accurate positioning on the relevant coordinates, it did visually make it easier to see distinct clear sound events as they occupied a larger area. In contrast clusters of neutral or unclear sound events were visually more complex due to their cluttered nature. A simple solution to allow the inclusion of more sound events within a single grid would be to scale all of the attributes of the shapes down. Monmonier [20] recommended that symbols are moved 'slightly apart' to decrease the amount of overlap, and if this is not possible, then an inset at a larger scale could be used for the crowded area. The code and the type and emotions symbols were always kept at the same scale (8 pt) and opacity (100%) which made them easier to locate and identify.

The maps clearly show the listeners' awareness of sounds located in front and, to a lesser extent, the sides of the listeners. Sound events which were located to the side were normally moving, whilst those in front were almost always stationary. The use of CMY for borders and RGB for fills meant that any combination, even a continuous gas long high sound event which had a continuous magenta border with a solid red fill was clearly legible. Where this does not work as well as hoped was when a sound was classified as displeasing, the thin nature of the border width made it difficult to read the material and interaction, without the ability to zoom. This could be partially rectified by increasing the overall scale of the borders, so that the thinnest is at least 2 points, which is currently the size of the neutral condition.

Shape and size were easy to identify, even when partially occluded due to their symmetrical nature, which meant that the entire symbol does not have to be visible in order to identify its shape. Smaller soft sound events were layered on top of larger loud ones, and semi opaque unclear sounds appeared slightly washed out compared to the stronger colours of the clear ones. When comparing maps it is easy to see what a participant or group are paying attention to, and how this differs from individual to individual. Figure 6 shows the designer's map for the auditory display and the participants' combined responses in situ with the vehicle pre-existing auditory environment subtracted.

It can be seen that the spatial cues have been identified, albeit with slight variation, the email message and the braking distance alerts have remained in front of the driver, but reversed, and the dead angle has been discerned as originating from the right, but not as far back as the designer intended. The type has remained consistent for the braking distance and dead angle, both being considered to be sound effects, but the message has only been classified as speech, rather than a combination of speech and sound effects. This suggests that the sounds contained within the message are passing unnoticed. The material, which in this case was gas, remains constant, whereas the dead angle is perceived as being intermittent rather than impulsive. This shows that the dead angle is thought to be more of a whooshing sound rather than an explosion, which is also possibly due to a close association with the sound which a passing vehicle makes, this is also borne out through the alert being thought to be temporally medium in length rather than short.

The pitch for the two alerts were judged to be high mid rather than just high and the dynamics for the braking distance was considered to be soft but still clear. All of the events were classified as informative and aesthetically neutral, as well as emotionally neutral. It can be seen that the participants experienced the auditory display in context in a manner similar to the designer's intentions.

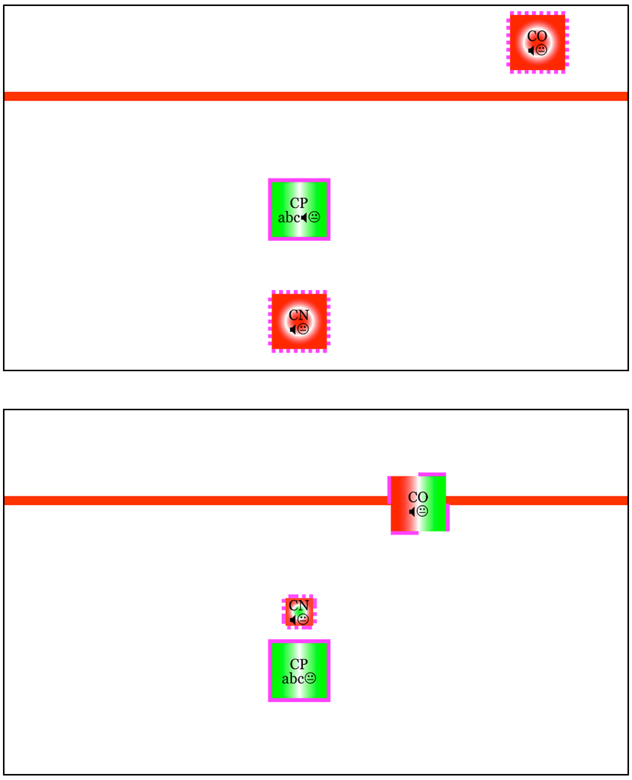


Figure 6: Magnified areas (identical coordinates)of designer's (top) and combined participants' (bottom)soundscape map for the auditory display in context with vehicle sound events subtracted (CN = Braking distance, CO = Dead angle, CP = Message)

3.4. The designer's comments

The designer found this method to be a quick and useful way of interpreting the data. He did, however, identify the need to include *height* channels. There were some other general comments as to the conduct of the studies themselves, observing that for longer duration soundscapes it would be useful for listeners to make notes, interruptions could also be used for longer experiments. He requested a confidence rating for each individual icon, as well as an electronic version where information about how the values were derived was displayed in a side table, on mouse-over of the relevant icon. He also suggested giving the designer a choice of classification scale, as sometimes looser is more appropriate. Some attributes might be better with more categories such as spectral and dynamics, whilst others would suit less, as in informative, where the neutral option could be dropped so that the decision is binary. The inclusion of spatialisation in the form of coordinates was deemed to be appropriate.

The labels used within the classification may require some fine tuning. He found the *temporal*, *spectral* and *dynamics* attributes to be context dependent, but relevant. The issue with the *temporal* attribute is that where a sound event could considered to be high in relation to its source, such as a high note on a cello, which is essentially a bass instrument, or a high tone from a male voice which might be considered to be low pitched in overall terms. It was also suggested that practical examples such as a female voice for the high category might be more helpful than the current examples of 'high pitch/frequency treble'.

With respect to *content*, the need for neutral option was queried and a request for a greater degree of granularity scale with possibly five or seven choices specifying the degree of information, such as moderately informative, informative, highly informative and so on. The use of the term noise was considered to be too ambiguous, noise could be considered as irrelevant and annoying. It was suggested that noise was changed to uninformative for consistency. The description was judged to be imprecise, as the information could be relevant but unwanted, this could easily be improved by removing the term unwanted. This attribute was regarded as the most important for the purpose of interface evaluation, especially with reference to answering the question of how informative it was.

Aesthetics were judged to be relevant, but like content, it would be more useful to have a more discriminating scale. With regards to the descriptions, mediocre was considered to be displeasing rather than neutral, and it was felt that the neutral state did not require a description at all. *Clarity* was regarded as pertinent, and like *type*, *material*, *interaction*, *spectral* and *dynamics* had the correct number of options, at three. Both the terms and descriptions were judged to be suitable. The classification of *emotions* could allow a greater degree of granularity, and the descriptors should be refined. Annoyance is not captured in the descriptor as a negative emotion, and it was queried as to whether surprise and anticipation were positive emotions. Concern was raised about the possibility of aesthetics cancelling out the emotions. There was a tendency for pleasing sounds to be classified as positive, This was even more evident for neutral aesthetics and neutral emotions, but was not the case with displeasing and negative emotions which only coincided fifty percent of the time.

Almost all of the methods of visualizing the attributes were regarded as effective, two suggestions for changes were made. The first was to amend the gradient associated with the temporal attribute so that only a radial gradient was used and that its size varied according to the length of the event. A short event would have a smaller area where the gradient was applied, whilst a long event would have a correspondingly larger area. This would allow for a linear scale as well as addressing the issue of the linear gradient sometimes being difficult to see in conjunction with a low level of opacity. The spectral representation might also be changed from three distinct colours to a continuous scale, in order to allow a greater degree of granularity.

4. CONCLUSIONS

This paper provides an illustration of the use of a soundscape mapping tool. It also showed that the tool could potentially be used by designers for the evaluation of sounds and auditory environments. The process of mapping has allowed a four dimensional auditory environment to be captured in two dimensional form, allowing ease of comparison between a designer's expectations and listeners' experiences. It also represents the effect of listening rather than hearing, where it is clear what is being attended to, and what has become habituated or has been ignored. With the car it was evident that sounds emanating from beyond the rear of the vehicle fell into this latter category, whereas those in front of or immediately surrounding the driver fell into the former. The relevance of sounds were also shown so that unwanted elements such as mobile phone interference and the driver's seat creaking could be silenced or masked, but other sounds such as the engine idling or accelerating, and the handbrake being engaged and disengaged should remain clearly audible as they were considered informative. The next stage of the research is to ask a range of sound designers to use the tool within their professional practise, and then query them about both the attributes and the visualization. This will help establish the tool's suitability for evaluating sounds and auditory environments.

5. REFERENCES

[1] P. Robare and J. Forlizzi, "Sound in Computing: A Short History," Interactions, vol. XVI, pp. 62-65, 2009.

[2] S. Barrass and C. Frauenberger, "A coummunal map of design in auditory display," in Proceedings of the 15 International Conference on Auditory Display, Copenhagen, Denmark, May 18-22, 2009 Copenhagen, Denmark: ICAD, 2009.

[3] S. Bech and N. Zacharov, Perceptual Audio Evaluation. Chichester, West Sussex: Wiley, 2006.

[4] American National Standards Institute., "Acoustical Terminology," American National Standards Institute, New York S1.1-1994, 1994.

[5] J. A. Ballas and J. H. Howard Jr, "Interpreting the Language of Environmental Sounds," Environment and Behaviour, vol. 19, pp. 91-114, January 1987.

[6] G. Bohme, "Acoustic Atmospheres," Soundscape, vol. 1, pp. 14-18, Spring 2000.

[7] M. O. Watson and P. Sanderson, "Designing for Attention With Sound: Challenges and Extensions to Ecological Interface Design," Human Factors,, vol. 49, pp. 331-346, 2007.

[8] I. McGregor, G. Leplatre, A. Crerar, and D. Benyon, "Sound and Soundscape Classification: Establishing Key Auditory Dimensions and their Relative Importance " in ICAD 2006 London: Department of Computer Science, Queen Mary, University of London, 2006.

[9] I. McGregor, A. Crerar, D. Benyon, and G. Leplatre, "Establishing Key Dimensions for Reifying Soundfields and Soundscapes from Auditory Professionals," in ICAD 2007, 2007.

[10] I. McGregor, A. Crerar, D. Benyon, and G. Leplatre, "Visualising the Soundfield and Soundscape: Extending Macaulay and Crerar's 1998 Method," in Proceedings of the 14th International Conference on Auditory Display Paris: IRCAM, 2008.

[11] W. W. Gaver, "Auditory Icons: Using Sound in Computer Interfaces," Human-Computer Interaction, pp. 167-177, 1986.

[12] P. Radomskij, "Measurement of noise," in Noise and Its Effects, L. Luxon and L. Prasher, Eds. Chichester, UK: John Wiley and Sons Ltd., 2007, pp. 13 - 43.

[13] S. Barrass and C. Frauenberger, "A coummunal map of design in auditory display," in Proceedings of the 15 International Conference on Auditory Display, Copenhagen, Denmark, May 18-22, 2009 Copenhagen, Denmark: ICAD, 2009.

[14] G. Leplâtre and I. McGregor, "How to Tackle Auditory Interface Aesthetics? Discussion and Case Study" in Proceedings of ICAD 04-Tenth Meeting of the International Conference on Auditory Display, S. Barrass and P. Vickers, Eds. Sydney, Australia: International Community for Auditory Display (ICAD), 2004.

[15] W. Moylan, The Art of Recording: Understanding and Crafting the Mix. London: Focal Press, 2002.

[16] G. Johannsen, "Auditory Displays in Human–Machine Interfaces," Proceedings of the IEEE, vol. 92, pp. 742-758, April 2004.

[17] S. Servigne, R. Laurini, M.-A. Kang, and K. J. Li, "First Specifications of an Information System for Urban Soundscape", IEEE International Conference on Multimedia Computing and Systems, vol. 2, pp. 262-266, 1999.

[18] D. Gibson, *The Art of Mixing: A Visual Guide to Recording Engineering and Production*, 2nd ed. Boston: Artist Pro Publishing, 2005.

[19] W. D. Grantham, Spatial Hearing and Related Phenomena. In B. C. J. Moore (Ed.) *Hearing* (2nd edition), London: Academic Press, 1995, pp. 297-345.

[20] M. Monmonier. *Mapping it out: Expository Cartography for the Humanities and Social Sciences*. London: University of Chicago Press, 1993.

AUDIO-VISUAL RENDERINGS FOR MULTIMEDIA NAVIGATION

Tifanie Bouchara, Brian F.G. Katz,
Christian Jacquemin

LIMSI-CNRS,
BP 133, 91403 Orsay Cedex, France
tifanie.bouchara@limsi.fr, brian.katz@limsi.fr
christian.jacquemin@limsi.fr

Catherine Guastavino

CIRMMT & McGill University,
555 Sherbrooke St. West,
Montreal, QC, H3A 1E3, Canada,
catherine.guastavino@mcgill.ca

ABSTRACT

Our study focuses on multimodal information access to audio-visual databases, and evaluates the effect of combining the visual modality with audio information. To do so, we have developed two new exploration tools, which extend two information visualization techniques, namely Fisheye Lens (FL) and Pan&Zoom (PZ), to the auditory modality. The FL technique combined coherent distortion of graphics, sound space and volume. The PZ technique was designed without visual distortion but with low audio volume distortion. Both techniques were evaluated perceptually using a target finding task with both visual-only and audio-visual renderings. We did not find significant differences between audio-visual and visual-only conditions in terms of completion times. However we did find significant differences in participant's qualitative evaluations of difficulty and efficiency. In addition, 63% of participants preferred the multimodal interface. For FL, the majority of participants judged the visual-only rendering as less efficient and appreciated the benefit of the audio rendering. But for PZ, they were satisfied with the visual-only rendering and evaluated the audio rendering as distracting. We conclude with future design specifications.

1. INTRODUCTION

With the current development of computer technology, the size of data collection is rapidly increasing. Efficient methods are required to help retrieve a particular document and browse the entire collection. Research on audio-visual information access has traditionally focused on classification and indexing techniques, mainly based on content [1]. This is true for different media document, particularly image [2], audio and music [3], and video [4][5]. However once the data collection is filtered by retrieval methods, there is still a need to find efficient presentation strategies to display the query results and help browse the new dataset. The role of the interface and presentation techniques has received little attention. Our work focuses on user-centered exploration strategies to facilitate interactive information access in these datasets. Most currently available navigation methods are based on vision even for audio or audio-visual data. Our study integrates audio in browsing tools to explore multimedia collections and to determine in what extent audio modality improves exploration.

For the moment, existing systems of video browsing such as video-on-demand systems (e.g. YouTube [6] or GoogleVideo [7]), typically present the user with a simultaneous set of static fixed frame images (called key-frame or poster frame) associated with each video. As such, in these systems, the initial search effort is based on visual feedback, with the user missing the audio content. Although certain systems enable the view of the entire sequence, it is most of the time for the browsing within a unique document [8]. Only few systems can play several videos simultaneously (like the wall of *Blinkx* [9]) and they still not offer an overview of the audio content. In order to address this issue, the user should access simultaneously the audio and video content of the data.

Some auditory displays take advantage of human abilities of simultaneous listening and browsing auditory document. These systems are based on the ability to segregate sound sources played in different location (known as *cocktail party effect* [10]). The *Dynamic Soundscape* project [11] applies this concept and sound spatialization to browse a single audio file. It relies on mapping temporal position within an auditory document to spatial location so the user can listen to different portions of the audio file at the same time.

As presented in the application of Stewart et al. [12] and in the *Audio Hallway* of Schmandt [13], some other interfaces give the user the possibility to explore a collection of several sounds distributed in space around her/him without any visual feedback. On the contrary the *SonicBrowser* [14], improved in the *Audio Information Browser* [15] and the *SoundTorch* [16] are enhanced by a visual icon representation of the sounds. The user can thus browse several sound files simultaneously by navigating through a 2D soundscape. These systems exploit a concept called *aura* for *SonicBrowser* (named *torch* in [16]) consisting of circles defining the limits of user's domain of perception. All sonic objects on the perimeter or beyond are silent, while all the objects inside the disk are simultaneously played with a relative loudness depending on the distance from the center.

The concept of *aura* is derived from visualization techniques ([17], [18]) used in Zoomable User Interfaces (ZUI) (also called multiscale interfaces [19]), and particularly from the *Fisheye Lens* (FL) concept. ZUI provide a powerful way to represent and manipulate large sets of data by managing the level of detail and separating the user point of interest area (focus) from the global view (context). Among these techniques *Pan&Zoom* (PZ) relies on translations and zoom level modifications through which a homogeneous but partial view of the dataset is presented. As a focus-plus-context method, FL presents the whole dataset at a low level of detail and utilizes a movable non-homogeneous distortion (magnification) to a section of the dataset in order to examine the subset at the required level of detail. Such interfaces have been proven beneficial for visual and auditory data browsing. We proposed, developed and evaluated two novel

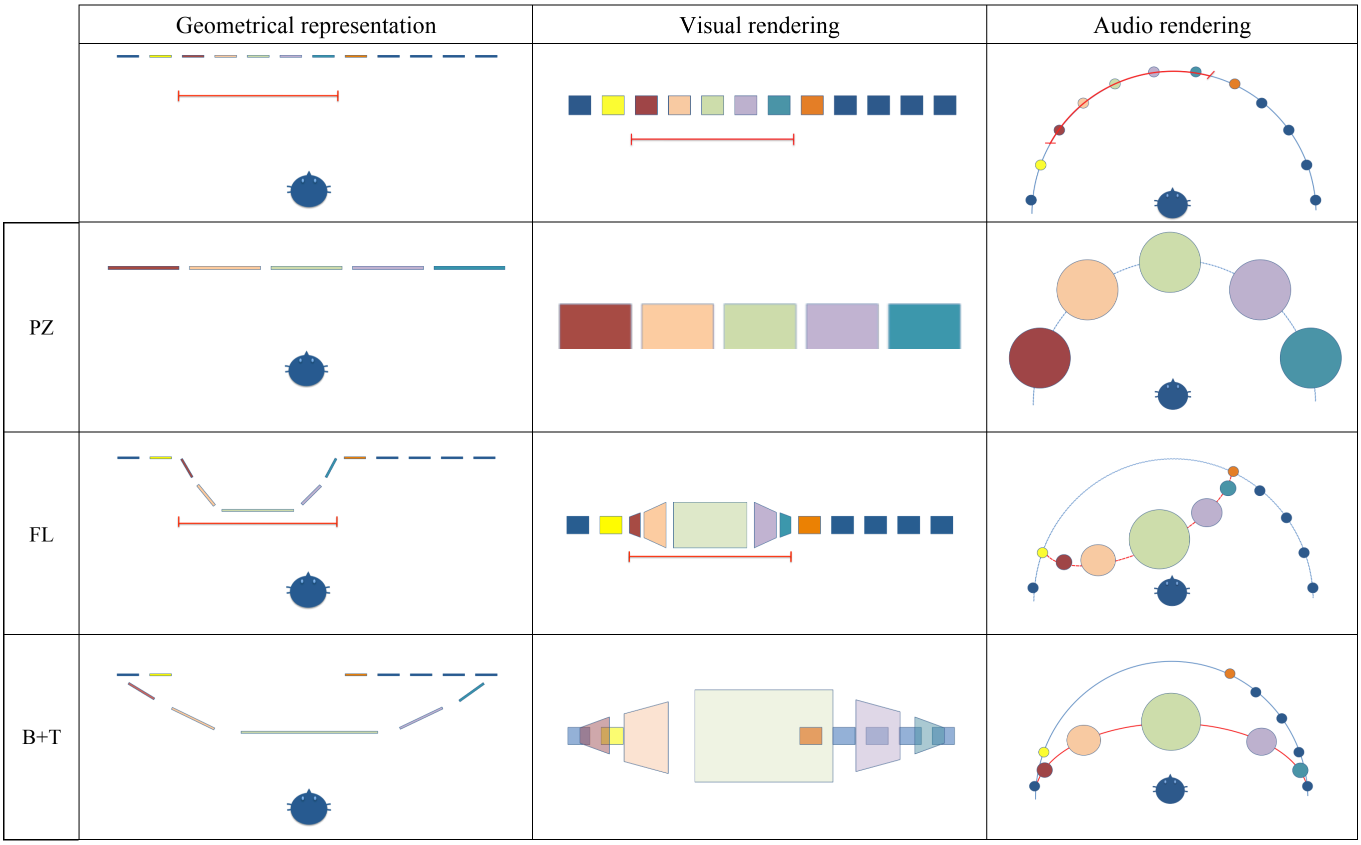


Figure 1: Schema of 3 different rendering techniques: Pan&Zoom (PZ), Fisheye Lens (FL), and Bifocal + Transparency (B+T).

audio-visual exploration techniques, combining two existing visual information access and visualization techniques, namely PZ and FL, with their auditory analogs.

The next section of this paper introduces the design of such audio-visual exploration techniques. Section 3 presents a user experiment comparing two modalities: a unimodal one (only visual) with a bimodal one (audio-visual) for the two different user interfaces, PZ and FL, while Section 4 discusses the results.

2. DEVELOPPING AUDIO-VISUAL RENDERINGS FOR NAVIGATION

2.1. Taxonomy of Zoomable User Interfaces techniques for visual information access

In Zoomable User Interfaces (ZUI) users can focus on a subset of a dataset by specifying the level of detail [17]. One of the most employed techniques is the *Pan & Zoom* (PZ). Zooming allows the user to change the scale of a specific area called *focus,* while information outside this area is discarded. Panning allows the user to translate the viewport. In such an approach, the rendering is homogeneous (without distortion) but there is no global view. As users cannot see the relationship between the visible portion and the entire structure, they can be disoriented by the lack of visual *context*. On the contrary *Focus-plus-context* techniques, combine the focus area and the global view in a single display. Among these techniques *Bifocal Display* superimposes the focus area over the context. Both areas are presented without distortion but the focus masks a part of the context and some information cannot be displayed.

Another option is to distort the rendering as in the *Fisheye Views* [20] (see also [18] for a review on distortion-oriented techniques). Originally this technique consisted of the suppression of non-interesting part of the information according to a threshold. It relied on the calculation of the *Degree Of Interest* (DOI) for each object and was designed for hierarchical information. An improvement of this method was designed for tree structures with the concept of *Hyperbolic Browser* [21] where more space is assigned to a portion of the hierarchy while still embedding it in a much larger context. The concept was also extended to a graphical fisheye lens in [22]. The focus area is enlarged while the rest of the image is reduced proportionally to the Euclidean distance to the center of the lens. This method combines the accuracy of spatial distortion while preserving the simultaneous visualization of the focus and context areas.

Pook et al. [23] suggested a transparency method where the contextual view is a transparent layer drawn over the magnified focus of attention. There is no masking and no distortion, however the large amount of information presented simultaneously results in more efforts for the user to distinguish one view from another.

2.2. Extension to audio-visual renderings

To extend the visualization techniques to the auditory domain and design audio-visual browsing methods, we chose to map different properties of graphical rendering to audio rendering: position of the objects are map from visual position in the screen picture to the spatialized audio rendering and size of the objects are mapped to the sound level. The mapping is presented in Fig.1. for 3 different techniques: Pan&Zoom (PZ), Fisheye Lens (FL), and Bifocal+Transparency (B+T).

The link between graphical and audio space can be seen as projection from the geometrical Cartesian representation (Fig.1 col.1), corresponding to a top-view of the visual rendering (Fig.1 col.2), to a polar representation for audio rendering (Fig.1 col.3). Indeed it is equivalent to say that the graphical rendering is analogous to the front space of spatialized audio rendering. The objects' positions are also coherent, but not congruent, between graphical and auditory renderings.

The mapping between the visual size and the volume of the object's sound is inspired from real life as both are linked to the distance from the user. Thus, we considered that the larger an object is in the visual rendering, the louder the sound of this object must be.

In the first method extending Pan&Zoom, there is no distortion. Also the objects have a homogeneous size and volume and are uniformly distributed in space. The main problems are that only few objects are displayed and no context can be perceived.

The FL design uses a visual position distortion corresponding to an angle manipulation for the spatialization of sounds. The progressive graphical magnification is equivalent to a progressive audio level increase. The main advantage for both modalities is the presence of context. However this results in a graphical distortion that can disturb users and in an audio distortion that can be difficult to perceive, because of the small azimuthal distortion. Indeed it is quite difficult to segregate the different sources inside the lens, as the objects are close to one another.

In the third method B+T, we decided to improve segregation of sound sources inside the focus area. Sound sources should be more spread out so that users can better segregate multiple audio sources [24]. B+T also combines bifocal display and transparency method. The rendering is also similar to FL but the focus area is centered and transparently superimposed on the context. The sources are more distinguishable, however some sources are heard as located in the same direction because of superposition.

Finally graphical and audio renderings can be combined non congruently, associating the graphical rendering from one method with the audio rendering from another, e.g. a PZ visual rendering with a B+T audio rendering.

2.3. Visual rendering implementation

Two graphical renderings methods were implemented: Fisheye Lens (FL) and Pan&Zoom (PZ). They are processed through shaders, small programs that are run on the graphics card [25].

The PZ technique renders only a single portion of the environment. This method corresponds to the manipulation of a camera as described in space-scale diagram by Furnas and Bederson [19]. The camera can be moved through the left/right axe (panning), and through the back/forward axe to change the scale of detail (zooming). Thus only a part of the global view is captured then enlarged to obtain an image of desired size, i.e equal to the screen size.

The FL rendering is divided into three parts: in the center area of the lens, objects are homogeneously magnified, outside the lens objects' size is not modified while in between the size is progressively interpolated. To compute the rendering of FL, three passes are necessary.

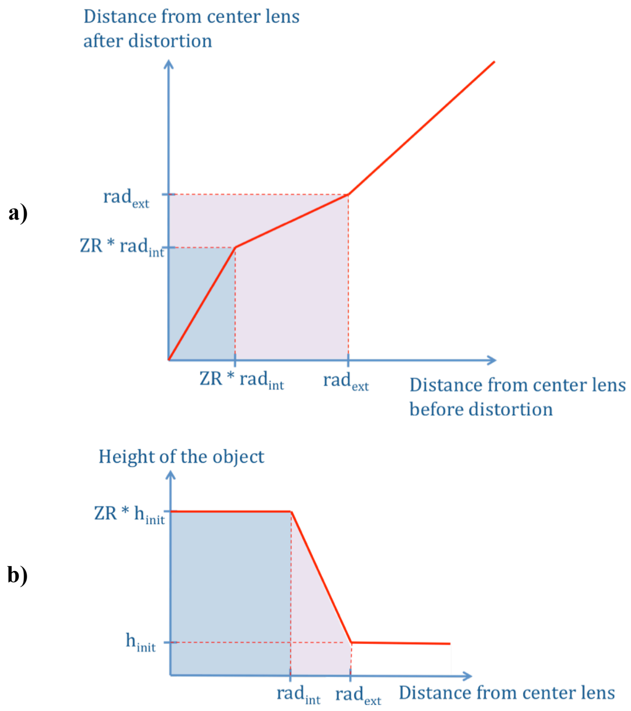


Figure 2: Distortion curves for FL technique. a) Position of objects after distortion. b) Height of object according to their visible position.

```
  if p ∊ d1
     then textFinal <- textZoomed
     else if p ∊ d2
        then textFinal <- f(textZoomed,ZR,p)
        else if p ∊ d3
           then textFinal <- g(textNorm,ZR,p)
           else textFinal<- textNorm
        end if
     end if
end if
```

Figure 3: Pseudo-code of the lens shader.

Then the rendering is processed by taking a zoomed view of the environment and stored as a second texture (*textZoom)* enlarged so that its size equals the screen size. The environment part captured in zoomed view is the one contained inside the lens of radius rad_{ext}. Both textures are then mixed and distorted according to the fisheye strategy: in the focus area (of radius rad_{int}) the zoomed view is used to have magnification without pixellization (when screen size is lower or equal to the texture size), while the normal view is used for the context part. The shader is parameterized to select the parts of the texture that are enlarged and to define the strength of the distortion according to a distortion ratio *ZR*. The deformed texture *textFinal* is then mapped to a quad parallel to the projection plane and is finally rendered on a view port that covers the whole display screen. A white border marks the boundary of the lens and allows the user to easily locate the position of the lens even at low distortion levels. Figure 2 presents the distortion curves chosen to modify the position or the size of the objects in the graphical rendering of FL. Figure 3 presents the pseudo-code for the lens shader creating the distortion: for each pixel *p* of the rendering picture we allocate the corresponding texture depending on which zone *p* belongs to (center, outside or between). *d1, d2* and *d3* are discs delimiting each part of the lens. The radius of the discs are rad_{int}, $(rad_{int+}rad_{ext})/2$, and rad_{ext} respectively. *f* et *g* are two functions of distortion using the curve b (Figure 2).

2.4. Audio rendering implementation

For audio rendering, we considered that a spatial separation among sound sources is necessary for perceptual segregation. The implemented audio rendering was also the bifocal+transparency (Fig. 1). We adapted our audio B+T technique to work with PZ or FL graphics described in Fig 1. The multimodal congruency is thus not respected but the link between audio and graphical renderings is still coherent.

As there is no visual distortion with PZ, we tried to keep a homogeneous rendering for audio. There is no azimuthal distortion in this audio rendering as illustrated in (2). However, we applied a low distortion on the volume (*vol*) to reduce the number of sources played simultaneously (1). The volume distortion is similar to the FL volume distortion (3) but with a maximal lens radius, i.e. equal to the width of the window rendering (screen size if in fullscreen mode).

$$vol = \begin{cases} v_{\min} + \log(ZR), & if \quad |dz| < rad_{\text{int}} \\ v_{\min} + \log(ZR) * e^{-c*|dz - rad_{\text{int}}|}, & if \quad rad_{\text{int}} < |dz| < rad_{ext} \\ v_{\min}, & if \quad rad_{ext} < |dz| \end{cases} \quad (1)$$

$$az_{bis} = az \quad (2)$$

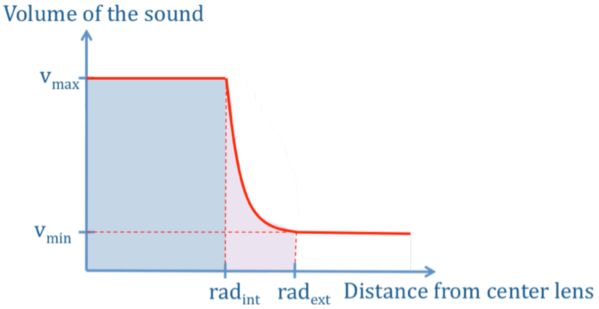


Figure 4: Nonlinear distortion curve used for volume in audio renderings from the visual position of objects.

For FL, the process relies on a distortion on both position and size in graphics and also on both angular position (*az*) and volume (*vol*) in audio. Equations of the proposed auditory distortion are presented in (3) and (4). *dz* represents the visible distance, after visual re-processing, between the object and the center of the lens. *ZR* is the zoom ratio or magnifying scale. v_{min} is the minimal level when no magnification is applied or when sources are out of the focus area. *c* is a constant giving the attenuation of volume between focus and context areas. α_{int} and α_{max} represent azimuth of sources on the internal perimeter rad_{int} and on the external perimeter rad_{ext}. Figure 4 presents the distortion curves chosen to modify the volume of objects according to their graphical position.

$$vol = \begin{cases} v_{\min} + \log(ZR), & if \quad |dz| < rad_{\text{int}} \\ v_{\min} + \log(ZR) * e^{-c*|dz - rad_{\text{int}}|}, & if \quad rad_{\text{int}} < |dz| < rad_{ext} \\ v_{\min}, & if \quad rad_{ext} < |dz| \end{cases} \quad (3)$$

$$az_{bis} = \begin{cases} dz * \dfrac{\alpha_{\text{int}}}{rad_{\text{int}}}, \quad if\, |dz| < rad_{\text{int}} \\ A * dz + B, \quad if \quad rad_{\text{int}} < |dz| < rad_{ext} \\ with \quad A = \dfrac{\alpha_{\max} - \alpha_{\text{int}}}{rad_{\max} - rad_{\text{int}}} \quad and \quad B = sg(dz) \\ az, \quad if \quad rad_{ext} < |dz| \end{cases} \quad (4)$$

2.5. Sound spatialization technique

Sounds are spatialized through a *virtual Ambisonics* technique for the auditory part of our bimodal interface [26]. This mixed method between Ambisonic encoding and binaural decoding allows us to treat simultaneously a large amount of sources without latency while providing a rendering on headphones usable for general public.

However, as the decoding part is independent from the encoding, the diffusion system could be replaced by a more immersive system with loudspeakers like VBAP, Ambisonic or WFS. Furthermore, we chose to use 2D audio renderings in this study but the implementation offered the possibility to extend the methods to 3D sound spatialization.

2.6. Global architecture

The software architecture (Fig. 5) was based on the *SceneModeler* package designed in two different parts: a virtual scene descriptor and a sound spatializer [27]. We used a triangular structure where all vertices (user, visual and sonic components) are connected by interaction links. The scene descriptor tool and the spatializer communicated through OSC messages via UDP protocol [28].

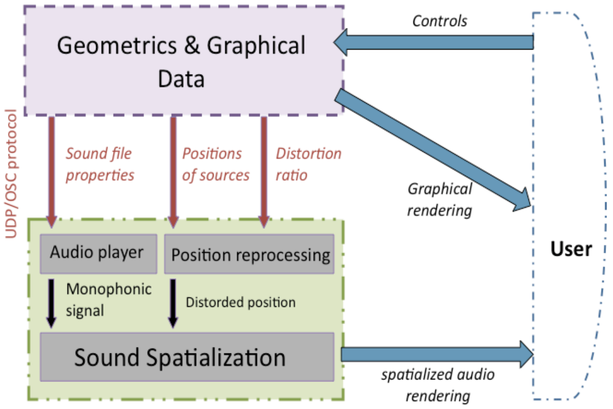


Figure 5: Structure of the interface.

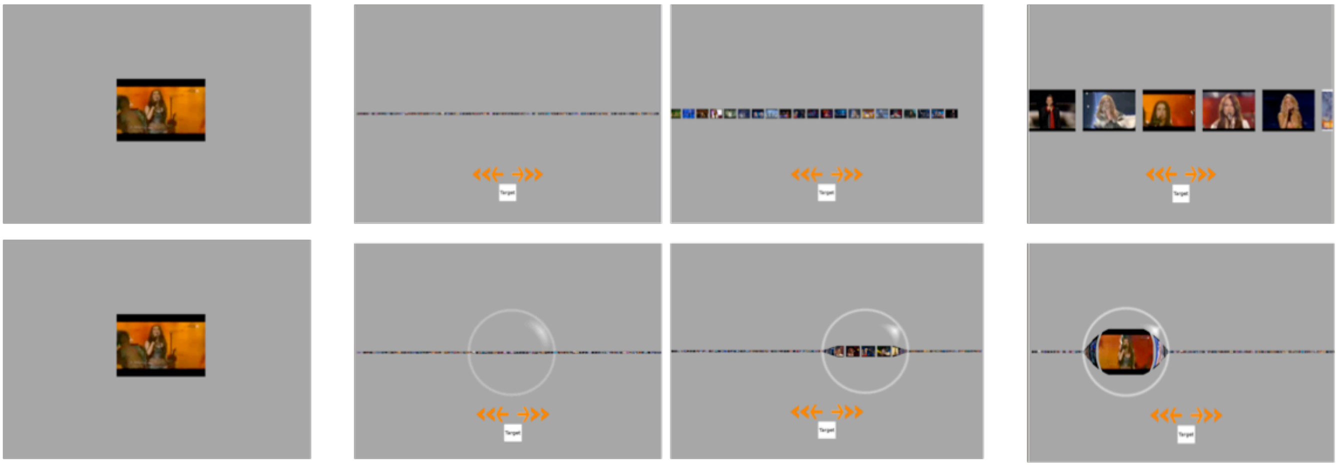


Figure 6. Storyboard of the task with both methods: PZ (top) and FL (bottom).

3. EVALUATION

The aim of the study is to evaluate the possible contribution of audio to browse audio-visual databases. Thus, the experiment is based on a comparison between audio-visual (AV) against visual-only (V) renderings. Two different navigation techniques, Pan&Zoom (PZ) and Fisheye Lens (FL), were tested in order to assess whether the audio influence could be depending in the chosen visual or audio rendering technique.

3.1. Experimental protocol

Participants

Sixteen participants, with basic computer skills and familiar with the use of a mouse, took part in this experiment (11 males, 5 females, mean age 27). They received $15 for their participation.

Design and conditions

We used a 2x2 within-subjects factorial design with 2 modal conditions (V/AV) * 2 methods (PZ/FL).

The graphical rendering of the FL was based, as described earlier, on a Fisheye Lens distortion. The radius of the lens was rad_{ext}= 178 px for a screen size of 1270x940 pixels. The visual rendering was also divided into three parts: the center of the lens (rad_{int} = 2/3* rad_{ext}) was magnified in a heterogeneous way, the external area was a global view and a progressive distortion was used in between the two. For audio-visual presentation, the audio rendering corresponded to the audio bifocal transparency described in (3) and (4). Outside the lens, sources were spatially spread out between -α_{max} and α_{max} = 90°. Sources inside the lens were spread between -α_{max} and α_{max} = 90° too but with the sources belonging to the internal area of the lens spread between -α_{int} and α_{int}=70°. With the highest deformation ratio (ZR=20), three different sources sufficiently separated in space could be heard simultaneously.

No distortion was used for the visual rendering of PZ. The audio rendering was based on the same audio distortion as for FL with a lens radius equal to the width of the screen, i.e. rad_{ext}=1270 px. More sound sources could be heard simultaneously than for FL, up to six or seven sources with the highest zoom ratio.

Hypothesis

Audio rendering can convey redundant information to reinforce visual feedback, or convey additional information to complement visual information. Therefore, we hypothesized that the addition of redundant and complementary audio rendering would enhance navigation and information when browsing an unorganized audio-visual collection. In addition, we investigate the effect of the rendering technique itself, and hypothesize that a more focused audio rendering (used in FL), i.e. with few but relevant sound sources, would be more useful that a less focused audio rendering (used in PZ).

The video collection

To evaluate renderings in a realistic context, we used a collection of 100 video clips from the Eurovision Song Contest from 2005 to 2008 [29]. The videos were selected from the result set of the textual query “Eurovision” on the video-on-demand system YouTube [6]. The video clips were excerpts of singers’ performances, each showing a different singer and a different song. For each video, we extracted a 10-second clip corresponding to a musical phrase. Video clips were then played in a loop. The different clips could easily be distinguished through the visual properties of the singers, their voice and the musical genres provided several clues for identification. Moreover there was a good balance between visual and auditory cues for identification and a consistency between simultaneous visual and auditory components. Finally the videos were selected from the same TV program to ensure homogeneity of the collection.

Videos were stored with a 160x120px size and displayed at 11x8px before magnification. The soundtrack of the videos were extracted from the movie and stored as monophonic signal (left channel only) in 44,1kHz in 16 bits wav files. They included singing voice and instrumental music. To spatialize the sounds we considered that each sound file was attached to the center of the corresponding visual object. All stimuli and audio-visual rendering examples are available on the web [30].

Retrieval task

The task was to watch a video clip and then browse the video collection to retrieve it as quickly as possible. Each trial was divided into three steps represented in Figure 6 : a presentation of the targeted movie, then a step of exploration to find the target by changing scale or distortion level and position of the focus, and finally the selection of a movie with the user clicking on it.

Participants started by clicking on a button to see and listen to the target, a 10 second video clip presented in isolation once with no distortion. At the beginning of the exploration step, the user was presented with an overview of the 100 videos in the collection, arranged in a line in random order at a reduced size and sound level, so that the user could not discern the different clips in this view. The user had to use the zoomable techniques proposed. The minimal size of the videos on the 1270x940 screen is 11x8 pixels while the maximal size is 220x170 pixels. As the videos are very small, thousands of stimuli would have been necessary to fill in the screen resulting in hours of browsing experimental sessions. Hence the line arrangement was preferred.

Procedure

After a training block (on all 4 conditions), the actual experiment was divided into four blocks corresponding to the 4 conditions of the factorial design, namely AV-FL; V-FL; AV-PZ; V-PZ, presented in counterbalanced order using a Latin square design. Each block consisted of 15 trials. On each trial the presentation of the videos was randomized and a new video clip was randomly chosen as a target.

After each block, participants were asked to provide free-format comments and to evaluate for each condition: the perceived efficiency, adaptability, and difficulty. After the experiment, participants were asked to indicate their preferred method, audio-visual condition and combination.

The entire experiment lasted around one hour and half per participant. Participants were invited to take breaks after each block.

Apparatus

For faster computing, we used a distributed multi-platform architecture on two different computers for this experiment. The first one processed only the audio rendering while the second one managed with navigation and graphical rendering. We used the platform VirtualChoreographer on a AMD Athlon 64X DUAL CORE 5000+ 2.60 GHz with a Nvidia 8600T graphic card for the navigation process and graphical rendering and the Max/MSP environment on a MacBookPro 2.4Ghz with an integrated digital sound card for the audio display. The audio rendering was presented on AKG K271 headphones.

3.2. Results

Our dependant variables included completion times, number of errors, adaptability, difficulty and efficiency ratings, collected after each block, as well as overall preference ratings as free format descriptors collected at the end of the experiment. To present the different results we used a color code throughout the paper: PZ conditions are represented in green, FL in blue, and audio-visual conditions are shaded in.

For the statistical analysis we first removed miss trials for which a wrong video was selected (~3.2% among the 960 trials: 5 errors for AV-PZ, 2 for V-PZ, 14 for AV-FL and 10 for V-FL; 240 trials for each condition). Then we removed outliers from the hit trials for each condition and participant (13 outliers for AV-PZ, 11 for V-PZ, 6 for AV-FL and 7 for V-FL). Outliers corresponded to hit trial for which the completion time was more that two standard deviations away from the mean. Completion times were considered only for hit trials.

A 2*2 factorial ANOVA revealed that completion times were significantly lower for PZ than for FL (F(1,890)=8.82; p=0.003) (see Fig. 7). No interaction effect between methods and modality were observed (F(1,888)=0.06; p=0.81). We subsequently report the comparison between AV and V conditions for each method separately.

For the PZ method, no significant effect of modality on completion times was observed (F(1,448)=0.46, p=0.59). However, the analysis of subjective ratings (Fig. 8 and 9) and the free-format comments indicated that participants evaluated the addition of audio rendering negatively for PZ technique. Indeed they rated V-PZ as significantly more easier to use than AV-PZ (t(15)=2.07, p=0.05). V-PZ was also perceived as more efficient than AV-PZ but this difference did not reach statistical significance. In addition, participants commented that PZ method "produced too much overlapping noise when scanning many videos" which is "more a distraction than an aid". They further commented on the difficulty to associate the sound to the right movie as too many sounds were presented at once. Thus, even in the audio-visual condition, the PZ method was "mostly a visual scan instead of audio-visual" all the truer, as visual information is highly reliable in this technique without distortion. Participants also enjoyed the visual rendering providing "visual scanning of many items of the same size" and "like the uniformity when scrolling".

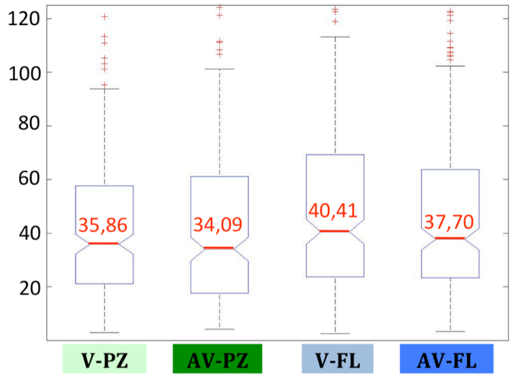


Figure 7. Mean completion times in sec. collapsed over all tasks and participants and grouped by conditions.

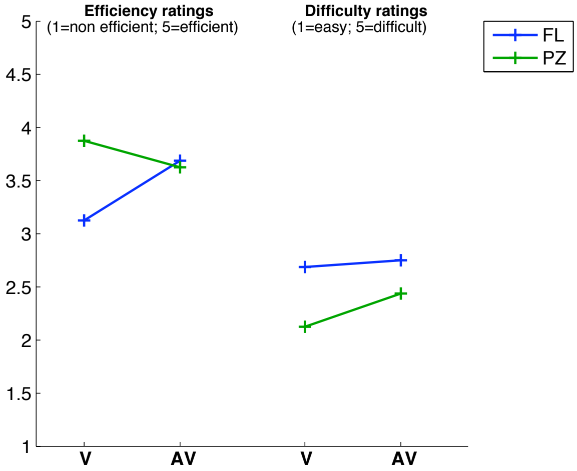


Figure 8. Average of subjective ratings collapsed over all tasks and participants and grouped by conditions.

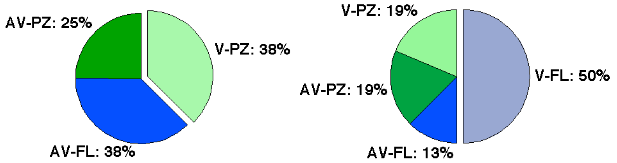


Figure 9: Participants' preferences (N=16): left) most enjoyed combination, right) less enjoyed combination.

As for PZ, while completion times for AV-FL modality were lower than for V-FL (37.70 sec. 40.41), the difference did not reach statistical significance (F(1,442)=0.06,p= 0.80). However the qualitative results (Figures 8 and 9) and comments differ from those of PZ as modality does not affect difficulty while audio improves the perceived efficiency for FL (t(15)=2.52, p=0.02). Participants justified that FL displayed too many sounds, but less than PZ, and that sound was also beneficial when zoomed in: "At first, the audio seemed to be a distracter, but once the magnifier is zoomed in, it is helpful". Thus the lens allowed the user to "visually scan multiple videos while audio scanning just few". In this case audio is used to compensate for the insufficiency of visual information as the "view of zoomed-in-image is limited" with FL.

To summarize, in terms of techniques, PZ is faster than FL and in terms of modality, the addition of audio rendering has a positive effect for FL and a negative effect for PZ. In one of the participants' own words "With PZ there are too much noise but with FL it's funny you've got only 3 or 4 sounds. But it is easier with PZ cause you can see all the videos". In their free comments, 81% of the 16 participants reported relying on audio during the experiment, either to browse sonically the video collection (55%) or only to confirm the visual selection certain ambiguous videos (25%). displays the conditions preferred and most disliked by participants. The majority of participants (63%) preferred bimodal conditions. Similarly, a similar percentage (69%) of participants disliked unimodal conditions. Together, these findings indicate the addition of audio rendering enhances user experience.

4. CONCLUSION

This study aimed to evaluate if the addition audio rendering could improve navigation in audio-visual collections using a multimodal user interface. The first step of the study was to suggest ways of combining audio rendering with existing graphical rendering. Two audio-visual methods related to Pan&Zoom and Fisheye Lens have been implemented in a visual-only mode and an audio-visual mode. They were evaluated with respect to the contribution of audio on video browsing. No significant differences were observed between multimodal and purely visual interfaces in terms of completion times. This could be explained by the predominance of vision in human perception but also by participants' previous experience with visual searching while audio rendering is rarely used for navigation. However, subjects self-reported audio as an enjoyable and interesting way to provide additional information. So we believe that the absence of performance improvement due to the inclusion of audio could be due to compensation between the positive effects (redundant and complementary information transfer) and some negative effects (auditory fatigue and discomfort). Participants also reported the background noise produced by the contextual sources as "annoying". In future instances, to avoid auditory fatigue due to the presentation of non-relevant sounds, we suggest keeping silent all sources outside of the lens (reciprocally outside the screen for PZ) as done by [14] and [16].

Furthermore participants reported a preference to rely mainly on visual rendering for navigation, and for a graphical rendering without distortion with several videos presented at the same time with homogeneous magnification. Participants' ratings and comments reveal the positive effects of conveying information through the auditory modality when focused on few sound sources as in the FL case. Our results suggest also us to design audio-visual renderings differently to benefit from advantages of both modalities. Providing a combination of the homogeneous PZ visual rendering plus the distorted FL audio rendering focusing on few sound sources should improve the navigation step.

Our primary goal was not to compare PZ to FL technique, as we focused mainly on the addition of audio renderings. However our results show that PZ significantly outperformed FL both in terms of completion times and affective reactions. Even thought the same control was used for navigation with PZ and FL, selecting a video might have been more difficult with FL as the lens had to be positioned on the video to select it. With PZ on the other hand, participants could click on and thus select any video displayed on the screen. The advantage observed for PZ could therefore possibly be attributed to interaction control.

Finally, this audio technique proposed here could be improved further by including additional spatial auditory cues to segregate sound sources, particularly elevation. For instance, we could apply azimuthal distortions in the same manner and arrange multimedia objects using a grid instead of a straight line – which is more representative of a real application. The tools could also be extended to an immersive 3D environment. However PZ is a non-egocentric concept that is not really suitable to immersive 3D scenes. On the contrary FL could be interesting to explore these environments.

Participants' positive reactions during the experiment showed the beneficial effect of audio rendering when focused on a limited number of sound sources (3 or 4 at a time). Future studies will investigate other audio design methods for multimodal navigation. Sound level distortion could be combined efficiently with distortion of other sound parameters to increase the effect. Specifically, a simulation of distance and presence, by adding reverberation or varying the high-low frequencies balance, could be used to differentiate foreground and background sound sources, thus directing attention to relevant sound objects and improving audio selection.

5. REFERENCES

[1] M. S. Lew, N. Sebe, C. Djeraba and R. Jain, "Content-Based Multimedia Information Retrieval: State of the Art and Challenges." *ACM Trans. Multimedia Comput. Commun. Appl*, vol. 2(1), pp. 1–19, 2006.

[2] A. Halawani, A. Teynor, L. Setia, G. Brunner, H. Burkhardt., "Fundamentals and Applications of Image Retrieval: An Overview." in *Datenbank-Spektrum, vol. 18*, pp. 14-23, August 2006.

[3] J. T. Foote, "An Overview of Audio Information Retrieval." in *ACM Multimedia Systems*, vol. 7 (1), pp 2-10, January 1999.

[4] R. C. Veltkamp, H. Burkhardt and H.-P. Kriegel (eds.), *State-of-the-Art in Content-Based Image and Video Retrieval*. Kluwer, 2001

[5] O. de Rooij, C.G.M. Snoek and M. Worring, "Query on Demand Video Browsing." in *Proc. of the ACM Int. Conf. on Multimedia (MM'07)*, Augsburg, Germany, 2007, pp. 811-814

[6] http://www.youtube.com

[7] http://www.video.google.com

[8] W. Hürst, "Interactive audio-visual video browsing." in *Proc of the 14th ACM Int. Conf. on Multimedia (MULTIMEDIA '06*), 2006, pp. 675-678.

[9] http://www.blinkx.com

[10] B. Arons, "A Review of the Cocktail Party Effect. *J. of the American Voice I/O Society,* 12, pp. 35-50, 1992.

[11] M. Kobayashi and C. Schmandt, "Dynamic Soundscape: mapping time to space for audio browsing" in *Proc. of the Conf. on Human Factors in Computing Systems (CHI '97), New York, NY, 1997, pp. 194-201.*

[12] R. Stewart, M. Levy and M. Sandler, "3D Interactive Environment for Music Collection Navigation" in *Proc. of the 11th Conf. on Digital Audio Effects* (DAFx-08), Espoo, Finland, 2008. pp. 13-17

[13] C. Schmandt, "Audio Hallway: a Virtual Acoustic Environment for Browsing" in *Proc. of the Symp. on User Interface Software and Technology (UIST'98), 1998, pp.* 163-170

[14] M. Fernström, and E. Brazil. "Sonic Browsing: an auditory tool for multimedia asset management" i*n Proc. of the 7th Int. Conf. on Auditory Display (ICAD'01)*, Espoo, Finland, 2001, pp. 132-135.

[15] E. Brazil, M. Fernstroem, G. Tzanetakis, and P. Cook, "Enhancing sonic browsing using audio information retrieval" in *Proc. of the 8th Int. Conf. on Auditory Display (ICAD2002),* Kyoto, Japan, 2002

[16] S. Heise, M. Hlatky and J. Loviscach, "Aurally and visually enhanced audio search with soundtorch" in Proc. Proc. of the 27th Int. Conf. on Human Factors in Computing Systems, 2009, pp. 3241-3246

[17] A. Cockburn, A. Karlson, and B. B. Bederson, "A Review Of Overview+Detail, Zooming, And Focus+Context Interfaces". *ACM Computing Surveys (CSUR), vol. 41 (*1), 2008.

[18] Y. K. Leung and M. D. Apperley "A review and taxonomy of distortion-oriented presentation techniques" in *ACM Transactions. on Computer-Human Interaction (TOCHI), vol. 1(2), pp. 126-160, 1994.*

[19] G. W. Furnas and B. B. Bederson, "Space-Scale Diagrams: Understanding Multiscale Interfaces" in *Proc. of the Conf. on Human Factors in Computing (CHI '95), 1995,* pp. 234-241.

[20] G. W. Furnas, "Generalized Fisheye Views" in *Proc. of the Conf. on Human Factors in Computing Systems (CHI'86), 1986,* pp. 18-23.

[21] J. Lamping, J., R. Rao and P. Pirolli, "A Focus+Context technique based on hyperbolic geometry for visualizing large hierarchies" in Proc. of C*onf. on Human Factors in Computing Systems (CHI' 95), 1995.*

[22] M. Sarkar and M. H. Brown, " Graphical Fisheye Views. *Communication of ACM*, vol. 37 (12), *pp.* 73-83, 1994.

[23] S. Pook , E. Lecolinet, G. Vaysseix and E. Barillot. Context and interaction in Zoomable User Interfaces" in *Proc. of the 5th Int. Work. Conf. on Advanced Visual Interfaces (AVI 2000)*, 2000, pp. 227-231

[24] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[25] R. Fernando and M. J. Kilgard*, The Cg Tutorial: The Definitive Guide to Programmable Real-Time Graphics,* NVIDIA, 2003. Available on-line: http://developer.nvidia.com/object/cg_tutorial_home.html

[26] M. Noisternig, A. Sontacchi, T. Musil and R. Holdrich, "A 3D Ambisonic Based Binaural Sound Reproduction System." in *Proc. 24th Int. Conf. of the AES: Multichannel Audio, The New Reality,* 2003. *pp. 1-5*

[27] T. Bouchara, "Le SceneModeler: des outils pour la modélisation de contenus multimédias interactifs spatialisés" in *Proc13ème Journées d'Informatique Musicale (JIM'08)*, GMEA-AFIM, Albi, France, 2008. pp. 8-13

[28] M. Wright, A. Freed and A. Momeni, "OpenSound Control : State of the Art 2003" in *Proc. of the 2003 Conference on NIME*, Montreal, Canada, 2003. pp. 153-159

[29] http://www.eurovision.tv

[30] http://www.limsi.fr/Individu/tifanie/downloads/audiovisual_renderings.zip

DEVELOPMENT AND EVALUATION OF A CROSS-MODAL XML SCHEMA BROWSER

Dena Al-Thani

Department of Computer Science,
Queen Mary, University of London
Mile End Road, London E1 4NS,UK
dea1@dcs.qmul.ac.uk

Dr Tony Stockman

Department of Computer Science,
Queen Mary, University of London
Mile End Road, London E1 4NS,UK
tonys@dcs.qmul.ac.uk

ABSTRACT

We describe the development and evaluation of across-modal XML (Extensible Mark-up Language) schema browser. The aim of developing the system is to investigate cross-modal collaboration between users. The browser provides an audio representation of XML schema documents in a way that preserves the structure of documents and supports multi-level navigation. The project has two principle objectives: 1) to overcome the difficulties faced by visually impaired users and sighted people using small screen devices when browsing XML schema files, 2) To explore usability issues when users collaborate using the auditory and visual interfaces of the system. The paper also examines differences between sighted and visually impaired users of the developed auditory interface.

The overall results of the usability evaluations demonstrate that both sighted and visually impaired users were able to perform tasks using the audio modality efficiently and accurately, and the same was true of sighted users interactions with the GUI.

The use of the system to support collaboration where each user employs a different mode (audio or visual) of the system clearly demonstrated that cross-modal collaboration is effectively supported, enabling users to collaborate and successfully complete a complex shared task.

1. INTRODUCTION

It has been more than a decade since XML was first introduced as a standard by the W3C (World Wide Web Consortium). XML is used in many applications, the primary one being data exchange between computer applications over the web. It is also used to store data in semi-structured databases. It plays an important role in modern web searching and in the transferr of data to portable devices such as mobile phones and PDAs (Personal Digital Assistants).

As XML has grown in popularity, it has become necessary to provide a schema language that ensures that XML documents satisfy a pre-specified structure. One of the most popular solutions to this problem has been the development of XML editors that provide a graphical tree representation of XML schema documents [1].
Using tree representations, XML editors provide the user with an overview of the XML document based on the schema that is to simplify the process of creating XML documents.

Screen readers present a linear representation of information, and have few mechanisms to provide overviews of information or to facilitate the exploration of data at different levels of detail [2] [3]. This project seeks to address these shortcomings by providing an auditory representation of XML schema information, and enable its efficient exploration at different levels.

2. RELATED RESEARCH

The way most visually impaired users browse web pages falls into one of two categories: either by using a screen reader such as Jaws or Window-eyes to render the output from a mainstream browser such as Internet Explorer or Firefox, or using a specially developed audio browser. Most audio browsers such as IBM Home Page Reader, Lynx, pwWebSpeak, etc are almost entirely speech based, and very largely loose information about the spatial layout of page elements. In general screen readers render the information on web pages in a very linear fashion [4].

A number of research efforts have examined in detail the role that non-speech sound can play in preserving the spatial information of web pages and improving the bandwidth of computer-human communication. A study by James [5] which examined the presentation of HTML in audio showed that when presenting hierarchical structures, such as heading levels within an HTML document, earcons proved to be more effective than a simple change in sound level. Petrucci Et Al. [6] developed WebSound, an auditory Web browser for blind and visually impaired users. They demonstrated that the use of non-speech sound in graphical interfaces can increase the bandwidth of computer output. They further demonstrated that a 3D immersive virtual sound environment, combined with haptic manipulation of the audio environment, can enable blind users to construct a mental representation of the spatial layout of HTML documents. James [7] developed the AHA Browser, in which auditory icons are combined with musical cues and speech processing to render web pages in which visual formatting is preserved. Murphy et al. [8] developed a multimodal browser plug-in with audio and haptic feedback, to explore how basic concepts in spatial navigation can be conveyed to web users with visual impairments. Using multimodal cues, users were able to successfully navigate a sequence of screens with directions from a sighted user.

The above research projects focused on how non-speech sound, in some cases in combination with haptics, could be used to preserve spatial layout and improve the audio presentation of html based web pages. In this project, we examine how non-speech audio can improve the accessibility of the audio presentation of XML schema documents, in comparison with the speech-based approach of current screen readers. We are not aware of any other studies that have applied non-speech sound to XML document presentation. However, the Auditory Display literature provides some guidance on how the different elements in such an auditory display might be chosen. The work by Brewster [9], on the use of hierarchical earcons, suggests that earcons may provide a good candidate for the representation of the tree structured XML schema documents we wish to represent. In this work, Brewster examined the use of earcons in "communicating hierarchical information". They also investigated how much a user can recall sound representations of hierarchical structures. The results of these experiments indicated that users can recall earcons with a high degree of precision, and so were able to know their position within a hierarchy after only a short amount of training.

Comparative studies of auditory icons and earcons have shown that users can react more quickly to auditory icons than earcons, but the structured nature of earcons enables them to represent more complex information [10]. Clearly speech will continue to play an important role in the display of the XML schema documents in our system, where the specific names or values of XML schema elements must be rendered.

Mynatt [11] synthesized a set of principles and guidelines regarding a non-visual representation of a GUI interface. These can be summarized as follows:

1. Mynatt 1: All the functionality accessible to sighted people using GUI interfaces must be accessible by visually impaired people. That includes icons, images, buttons and spatial location of GUI objects.
2. Mynatt 2: apply good GUI design principles wherever possible such as direct manipulation when implementing a non-visual representation of a GUI.
3. Mynatt 3: Change any interaction device used in a GUI which is not appropriate for use in an auditory interface.
4. Mynatt 4: Mechanisms should be provided to support mutual awareness, i.e. users should be aware of the focus of attention and actions of co-users
5. Mynatt 5: Both non-visual and visual interfaces must support the same mental model. However, Winberg & Bowers [12] argued that having non-visual and visual interfaces which are coherent and have a similar mental model does not guarantee success.

In the design section we will examine the way in which we used the above guidelines to determine the rationale for the design of the XML schema browser.

3. SYSTEM DEVELOPMENT

3.1 Choice of schema language

The XML schema language produced by the W3C was chosen because it is widely used and supports the definition of complex document structures.

3.2 Choice of schema style

XML schema documents are written in numerous different styles. To keep the development work within manageable bounds, while providing an adequate test of the approach, it was decided to develop the cross-modal browser for one specific style of XML schema representation. A number of recent studies aimed to classify these styles and identify their strengths and weaknesses [13]. Among the most popular XML schema organization styles are Russian Doll, Salami Slice, Venetian Blind, and Garden of Eden. These styles differ mainly in the way in which they define complex elements. [13] The chosen style is the Salami Slice. In the salami slice style, the complex elements can only contain references to simple and complex elements which are defined in the first level of the XML schema document. The reason for choosing the Salami Slice style is that according to the W3C schools website, it is the most widely used [14]. Additionally, studies have shown that it is conceptually simpler than the other organizational styles of XML schema. Further, it supports the reuse of elements within the document [13] [15].

4. DESIGN

We first describe the visual interface design. We then detail the design of the auditory interface, before examining the role that the guidelines developed by Mynatt [11] played in assisting the process of mapping from the visual to the auditory design.

4.1 Visual Interface Design

In the graphical interface, two main components need to be displayed. The first is the graphical representation of the XML schema. The second is the control panel which contains the available functionality to support browsing. The screen is divided in to three parts. The top part of the screen contains the buttons for loading an XML file, getting help and listing to sound samples. The lower right part of the screen contains the graphical representation of the XML schema document, and the left part of the screen contains the buttons that are used to get node details and to move from one node to another. Figure 1 shows the initial GUI design.

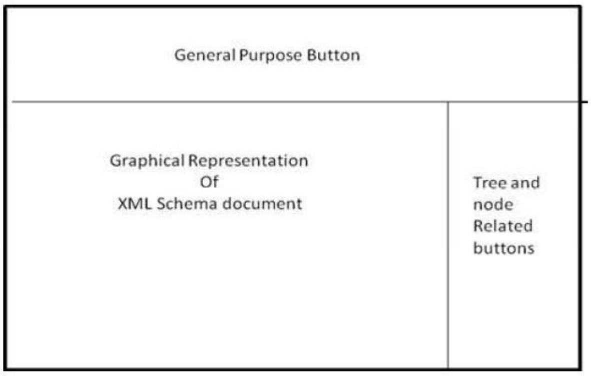

General Purpose Button
Graphical Representation Of XML Schema document
Tree and node Related buttons

Figure 1: The Visual Interface Design

4.2 Auditory Interface design

Because of the complexity of the documents to be represented, we determined to support two views of each schema document:

1. An overview of the XML schema document, to give the reader a sense of the document's size and complexity. This is particularly important given the serial nature of audio. Providing an overview of complex data has been the topic of several research papers [16] [17] in which it was emphasized that it should not be necessary for users to listen to long streams of audio to gather an overall view of the major features of the represented data.
2. A detailed view of the XML schema document in order to present its structure and detailed contents. Audio samples and video demo of the XML schema browser can be downloaded from (http://www.megaupload.com/?d=8W9QV5DU)

4.2.1 Use of Speech

Speech is an essential component in this application, as this is the only way to represent certain types of essential information using sound such as the names of nodes in XML schema documents. Speech was used substantially in both the overview and detailed display components of the auditory interface of the browser. In order to help the user keep track of their current depth in the tree, the pitch of the speaker's voice is modified to represent the level of the element within the XML schema tree. As the user navigates to lower levels of the tree structure, the pitch of the spoken voice is decreased, and visa versa.

4.2.2 Non-speech sound

For Non-speech sound, two representation techniques were used auditory icons and earcons. Samples of the elements used in the auditory display are included with the paper and will be presented at the conference.

Auditory Icons: Auditory icons have been used in many parts of the application to represent a number of components of an XML schema document. In particular these were used because their sounds are very distinct and life-like, and where for example there was a direct word association with the schema element being represented. Examples of the use of auditory icons are as follows:

a) The sound of a car braking is used to represent restrictions and limitations associated with simple elements and attributes.
b) The sound of keys clinking is used to represent ID, IDREF, and IDREFS types in an XML schema. ID usually represents a primary key while IDREF or IDREFS represent foreign keys in a database.
c) The sound of a water bubble is used to represent an attribute; Attributes are always linked with complex elements. In order to differentiate between the child elements of a complex element and its attributes the water bubble sound is used through its association with the bubble symbol often used in visual diagrams. It can be argued that this is a less direct association for a visually impaired user, who may or may not know of the use of the bubble symbol, but it is an attractive and memorable sound which should be relatively intuitive to sighted users of the system using PDAs.

*Earcons:*Auditory icons are easier to recall, but studies [9] have shown that in some cases, performance becomes more efficient when using earcons. Earcons are used to inform the user about the number of child elements of a complex element. They are played in prior to the complex element name. They are produced at runtime. The numbers of musical notes in an earcon represent the number of child elements of the complex elements. In this project Earcons are also used to notify the user that the end of a tree branch has been reached. Earcons were produced using a combination of Csound and Audacity.

Concurrent presentation of auditory information is used to reduce the pace differential between browsing in audio and visual modes. The application provides concurrent audio and visual feedback when a button is clicked. In addition, while traversing within the XML schema tree, concurrent feedback of speech and non-speech sound is heard by the user in order to overcome the serial nature of audio information. For instance, when users navigate to the next element in the tree, non-speech sound heard before the element name indicates that it is a complex element. This non-speech sound also indicates the approximate number of child elements that exist.

4.3 Applying Design Guidelines

As mentioned in the Related Research section, Mynatt 1997 represents one of the most detailed attempts to provide general guidelines about designing auditory interfaces that deliver equivalent functionality and usability as their GUI counterparts. We examine below how we applied Mynatt's [11] guidelines in the development of our system:

- Mynatt 1: All the functionality accessible to sighted people using GUI interfaces must be accessible by visually impaired people. That includes icons, images, buttons and spatial location of GUI objects. This was achieved by mapping visual objects to appropriate auditory objects, principally auditory icons, earcons and static and dynamic speech elements. Through the overview mechanism we tried as far as possible to provide a summary of schema documents which would give some idea of their size and complexity, providing some of the characteristics available to a sighted user when overviewing the document on screen.
- Mynatt 2: apply good GUI design principles wherever possible. The position we took on this guideline was that the strengths and weaknesses of audio and graphical representations are very different, and that what works well in a GUI will not necessarily translate intuitively to an auditory interface. For example, brackets are widely used in schema specifications to indicate nesting, but these were not reproduced directly in the auditory display, but audio users are provided the equivalent information through the audio context described in terms of speech, auditory icons and earcons.
- Mynatt 3: Change any interaction device used in a GUI which is not appropriate for use in an auditory interface. We adhered to this principle by substituting the keyboard for the mouse when navigating schema documents and ensuring common navigation options are supported by hot key combinations.
- Mynatt 4: Mechanisms should be provided to support mutual awareness, i.e. users should be aware of the focus of attention and actions of co-users. We adhered

to this principle by ensuring that the presentation of schema information is always synchronized between the visual and audio interfaces.

- Mynatt 5: Both non-visual and visual interfaces must support the same mental model. We adhered to this principle by using the tree structure as the basis of schema representation in both the audio and visual interfaces. In the audio interface, users start from the top level, and have a choice either to continue to navigate the schema at the level of complex elements, or whether to open up successive amounts of detail on demand. Navigation of the tree structure and synchronous cross modal presentation of information is supported by presenting the element name and highlighting it when the position has changed. The user is also able to change the position at anytime using buttons or equivalent keyboard shortcuts.

Concurrent presentation of auditory information is used to reduce the pace differential between browsing in audio and visual modes. For example, while traversing within the XML schema tree, concurrent feedback of speech and non-speech sound is heard by the user in order to overcome the serial nature of audio information. For instance, when the users navigate to the next element in the tree, non-speech sound heard before the element name indicates that it is a complex element. This non-speech sound also indicates the approximate number of child elements that exist.

5. IMPLEMENTATION

The system was implemented using Java on a PC platform, using the Java Speech API (FreeTTS) and the DOM (document object model) API for representing XML schema documents as tree structures. The earcons were created by calls to the Java Sound API (MIDI), while auditory icons were presented by playing pre-recorded sounds using the Java Sound API. In a number of situations requiring only static speech, we pre-recorded the speech and made use of better quality TTS engines such as Verbose by NCH Swift Software and VoiceMax by Tanseon systems.

In the case of speech sound, Free TTS is used to represent the runtime data which is the XML schema tree. It is used to give information about the current element. When a child element of a complex element is represented, the pitch of the Free TTS voice is slightly lowered to differentiate between a child element and its parent element. From prerecorded sound software, two voices were chosen. The voice of a female was used to represent indicators in XML schema documents and the voice of a male was used represent the buttons. Echo was added to the male voice in order to distinguish between the male voice that represents the buttons and the male voice of Free TTS.

For non-speech sound, MIDI was used to represent complex elements. Two audio representations of complex elements are designed for this purpose. For both representations, the MIDI sound is played prior to the name of the complex element.

The first represents complex elements with child elements that are less than or equal to three. In this representation, the number of the repetitions of the MIDI notes indicates the number of child elements, thus allowing the users to know the number of child elements without needing to go to each child element.

The second represents complex elements with more than three children. In this representation, a major chord of four notes is played using two instruments. The reason for coming up with an alternative solution to represent complex elements with more than three children, is that while prototyping the first representation with participants, it was noticed that repeating the MIDI notes helped the participants to know the number of child elements. However, when the number of child elements are larger than three it started affecting the user's performance time. As the number of notes increased, the time of playing the MIDI sound increased. Therefore, a better alternative was needed.

6. EVALUATION

We were fortunate in having ready access to users and Formative evaluations of early prototypes guided the design of the system, but the results described here come from a more detailed, summative evaluation.

The main goals of the usability experiments are to find out whether the audio representation of the XML schema documents is able to provide a way for visually impaired users and sighted users who used small devices such as mobile phones and PDAs to work with XML efficiently, and whether the audio and visual interfaces together can support cross modal collaboration. In addition, while conducting the usability experiment we also aimed to compare the auditory XML browser interface with a screen reader that visually impaired users use to read XML schema documents.

6.1. Auditory Interface Usability Evaluation

6.1.1Hypotheses

Since the goal of using this approach is to determine the usability of the system, two hypotheses were defined:

*Hypothesis 1***:** Using this interface, users are able to obtain a useful understanding of the nature, application area and major components of a schema document.

*Hypothesis 2***:** Using this interface, users are able to navigate efficiently to appropriate parts of the schema document in order to perform tasks such as information seeking and compare schema elements.

To test the first hypothesis, we needed to examine whether the auditory interface allows the user to have an effective overview of the information presented in the XML schema document. This is tested by asking participants to listen to two audio presentations of the schema by the system, and asking them a set of general questions about the schema. These questions ask about the size and application area of the given schema as well as the numbers of complex and simple elements. The second hypothesis is tested by investigating the efficiency of the navigational features of the auditory interface. This is achieved by allowing users to navigate around the schema as much as they wish, while asking them a set of questions focusing on low level details of the schema, such as finding the details of specific elements, determining the number of IDREFs, and navigating to child elements.

6.1.2 Procedure

Nine sighted and four visually impaired participants were recruited. They were all given a sufficient amount of training prior to conducting the evaluation which took from 15 to 40 minutes. The primary factor behind the variability in training time was the user's previous knowledge of XML.

The participants were asked to use the interface to answer two sets of questions. The first set of questions was used to determine the participant's ability to get an overview of the schema, therefore the questions where quite general such as the number of simple elements, the number of complex elements and the domain the XML document is related to. Whereas the second set of questions determines was used to determine their ability to understand the schema in details. It contains questions that ask the participants to navigate to a certain node and write down the names of its child elements, attributes or primary key. The schema given to the participants in training and evaluation ranged from medium to large, where we defined a medium schema to have from 5 to 10 complex elements and from 5 to 10 simple elements, and a large schema we took as having more than 10 complex and more than 10 simple elements. The maximum schema that was given has 17 complex elements and 25 simple elements.

6.1.3 Analysis of Results and Discussion

General Observation

The participants' overall performance was fairly good with a very low error rate in both set of given questions. For the first set of questions the visually impaired participant error rate was 0% and the sighted participants' error rate was also extremely low, 3.8%. For the second set of questions which examine the participants' ability to understand the schema details the error rates were also low for both groups of participants. The average error rate for visually impaired participants was 20% and the average error rate for sighted participants was 13.8%.

Training times ranged from 15 minutes to 40 minutes while the overall task performance time ranged from 20 to 50 minutes. From direct observation of the interactions and discussion with users, there were a number of external factors that affected the individual training time and performance. These were as follows:

1. Computer literacy: Computer literacy played an important role in the overall user performance time. Two visually impaired and one sighted participant, who were less familiar with HTML and XML than the other participants, took longer to perform the tasks.

2. Experience with auditory display: Sighted Participants who had not previously used an auditory interface showed some hesitation and confusion at the start of the training session. This was expected as studies have shown that representing complex data needs in-depth training (Brewster, 1994) (Vickers and Alty, 1996). In addition, some sighted participants had problems remembering some of the non-speech sounds, in particular the attribute sound, whereas visually impaired participants had no difficulties remembering them.

3. Familiarity with XML documents: Participants who were less familiar with XML documents had some difficulties in differentiating between simple elements and attributes. However, they developed a better understanding as they worked through the tasks. Additionally, it was noticed that participant performance in a given task can sometimes be affected by not having prior knowledge of XML. For instance, in some cases the participants forgot that a complex element has child elements.

Apart from the above factors affecting performance times, the participants performed as expected. All participants were able to relate auditory icons to the XML schema components which they were intended to represent. By listening to the earcon which sounded before the complex element name, they were able to identify the number of child elements belonging to the complex element. Both visually impaired and sighted participants made use of the relationship between the pitch of the voice and their current level in the tree.

Summary of the results of Experiment 1

The main findings revealed that both visually impaired and sighted participants performed well on both the overview and detailed navigation tasks. Both were able to develop a good overall understanding of the XML document with low error rates. This supports the first hypothesis, as all the participants developed an adequate understanding about the schema size and nature of the schema. It was clear that the differences between participant performance times were due to the external factors described in the general observations section.

When examining the participant's ability to get a sufficient understanding of the XML tree details, there are a number of variables that affected the performance of the users and therefore had a bearing on the second hypothesis. Firstly, the scores were affected by the external factors explained previously. Secondly, training played an important role in this part of the experiment. Thirdly, the learning curve had an influence on the subjects' performance. Figure 2 shows the results of three participants in three trials .In each trial the participant was given a different schema, but they were of the same level of complexity. Complexity is defined here as a composite measure incorporating the number of complex elements, simple elements, attributes, and restrictions within an XML schema document. For the three participants the scores in the second trial were higher.

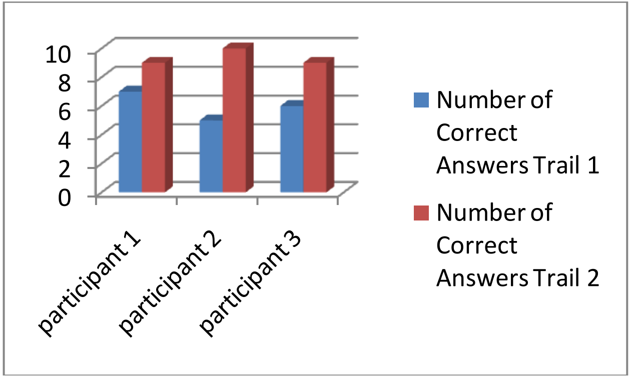


Figure 2: Three Participants Results in Two Trials

6.2 Cross-modal Interface Usability Evaluation

6.2.1 Hypothesis

Since it is a cross-modal system the aim was to enable users using the audio interface only and users using the visual interface only to be able to work together and have a similar mental model of the system. In other words, their understanding about the main components of an XML schema document should be similar and so allow them to work together coherently. The hypothesis to be tested was as follows:

Hypothesis 3: That the two users are able to collaborate and to develop an accurate representation of the XML schema they are browsing.

6.2.2 Procedure

For this experiment we recruited participants in pairs. Three pairs of sighted participants were recruited. Participants were trained on either the visual or audio interface.

The participants were then asked to use the interface on which they were trained until they felt comfortable using it.

Following this, Participants worked in pairs. For each pair, one participant only used the auditory interface and the other participant only used the visual interface. In the experiment, both participants using the different modalities were given the same schema. Their task was to work together to create an XML document that satisfied the structure defined by the schema.

They used different computers and were seated with their backs to each other as they were also informed that they cannot view the other participant's task sheet at anytime and that the only way to communicate with the other participant is via direct conversation. Their conversations were then recorded and analyzed. It was made clear to the participants that they can plan their work collaboratively in the way that suits them, as the experiment is mainly focused on the result of the collaborative work rather than the process of their collaboration.

6.2.3 Analysis of Results and Discussion

Generally, the collaborative task was performed well, with an average time to complete the task of 10.5 minutes. It was observed that the participants in the collaborative work did not face any difficulties while trying to explain information related to a specific element in the schema tree.

An interesting observation was that in all three experiments, the participants using the visual interface started the conversation first and tried to lead the collaborative work.

However, around the middle of the process the participant using the visual interface stopped leading the work and both participants started working together more evenly. The reason might be that both participants were not familiar with auditory interfaces. Therefore, the participant using the audio interface was more hesitant at the start than the participant using the visual interface, allowing the participant using the visual interface to lead the work. Once the audio interface participant became familiar with the interface and had gained more confidence, then both participants took part in the work more evenly.

It was clear that individual and collaborative performances improved with time. The participants' collaboration work became better towards the end of the process, as they became more familiar with the system.

The most important observation relates to the result of the overall task, as both participants were able to create XML documents correctly. Even though participants using the audio interface had not seen the XML schema structure, they were able to create an XML document that satisfied the given schema.

However, the XML documents created by non-XML users were, not surprisingly, less efficient, but nonetheless they did demonstrate a good understanding of the structure presented to them through the interface of the schema browser. An example of the kind of error these users made was in differentiating between attributes and simple elements.

6.3 Usability Experiment comparing schema reading using the XML browser with reading schemas using a Screen Reader.

*6.3.1 Experiment Design Research Hypothesis***:** (*Hypothesis 4)*

The use of speech and non-speech sound in the schema browser to represent XML schema documents is more efficient compared to reading schemas with a screen reader.

6.3.2 Procedure

Four visually impaired participants were recruited. All participants are experienced JAWS screen reader users and all are computer literate. However, they have little knowledge of XML. They were all given 20 to 40 minutes training on the schema browser's auditory interface.

The participants were asked to review three XML schemas using the JAWS screen reader, and three other XML schemas using the audio interface. The XML schemas reviewed using the screen reader are different than the ones reviewed using the XML browser. That is to avoid any bias results, as reviewing the XML schema using one tool, will effect the participants performance when reviewing the same XML schema using the other tool. After reviewing each XML schema, participants were asked to describe the XML schema. They were also asked to give their comments regarding both tools.

6.3.3 *Analysis of Results and Discussion*

Due to the fact that all the participants are experienced screen reader users and have a modest knowledge in XML, external factors such as computer literacy, and knowledge in XML were fairly constant across the 4 users. From the timings collected in the experiment, it was clear that the time that it takes a participant to review and fully understand a schema using a screen reader was longer than the time it takes another participant to review the same schema using the audio interface.

The figure below (figure 3) shows the time it took the participant to review a schema using both tools. Schemas were numbered according to their complexity, with one being the most complex with larger numbers of complex and simple elements and six being the smallest. From the figure it is evident that with larger schemas the average time taken to review using a screen reader was almost double the time it took the participants to review the same schemas using the XML schema browser.

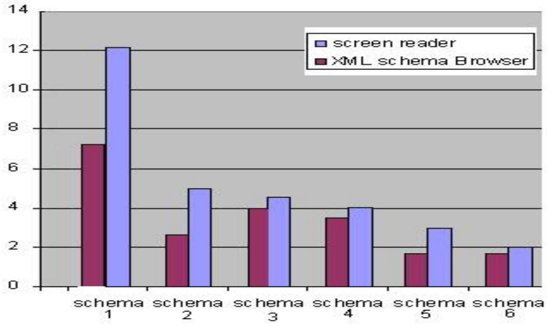


Figure 3: The average time taken using the XML schema Browser and the Screen Reader.

All participants were able to describe the content of schemas with differences depending on the tool they used. It was noticed that with complex schemas participants when using screen readers seemed to be less clear on the overall structure of the schema, whereas participants who were using the XML browser on the same schema demonstrated a good understanding about the structure of the schema as well as its details.

The above outcomes support the hypothesis that within the small number of participants involved, the combination of speech and non-speech sound provided in the schema browser is able to enhance the users' performance in comparison with screen readers.

7. DISCUSSION AND CONCLUSIONS

This paper introduced a novel approach to represent XML schema documents in audio. It also examined a number of ways of representing complex data in audio. Evaluations have examined the use of the audio interface of the browser used alone, and the use of the visual and audio interfaces by 3 pairs of users performing a collaborative, cross modal task.

The results of the evaluations demonstrate that the audio interface was successful in supporting audio browsing and cross modal collaboration for the relatively small numbers of users involved in the trials.

The benefit of the approach taken can be summarized as follows:

1) The audio XML browser helps to overcome the problems visually impaired XML users face when using screen readers. Given the serial nature of sound, screen reader rendering of XML schema contain a number of repetitive and unnecessary symbols that can overload the user's short term memory, which may affect the user's understanding of the structure of the data.

2) The browser enables rapid identification of the XML schema structure, and gives the user the option to get more details on demand.

3) The use of auditory icons and earcons provide a concurrent presentation of the properties of the elements which help to improve use of the communication bandwidth between the computer and human, rather than presenting these elements serially as they are when read with a screen reader.

4) Rather than representing the data serially, the data is represented in three different levels, leaving the user to match the level of detail to the task.

Additionally, Evaluation of the system for cross modal collaboration suggests that once users of the auditory interface have become comfortable with its use, they are able to take a full part in the collaborative task and that both users are able to form a sufficient mental model of the structure of the xml document described by the schema to be able to synthesize a document that accords with the underlying schema.

As well as performing the experiments described here with more users, an important remaining experiment is to test the collaboration between visually impaired and sighted users. However, given the results so far, we anticipate that the results of this experiment will be at least as good as the collaborative experiment described here, as visually impaired users in general will not have the difficulties of lack of familiarity with the auditory interface experienced by sighted users in the early part of the collaborative experiment described above.

8. ACKNOWLEDGMENT

Many Thanks goes to all who helped us from our colleagues.

9. REFERNCES

[1] D. Mertz, *XML Matters: A roundup of editors. IBM's Developer work.*2001. Website: http://www.ibm.com/developerworks/library/x-matters6.html.

[2] L.Brown, S. Brewster, R. Ramloll, M. Burton, M. and B. Riedel. "Design Guidelines For Audio Presentation Of Graphs And Tables". *Proc. of Conf.* on Auditory Display (ICAD 2003), Boston,USA, July2003.

[3] T. Stockman, G. Hind,C. Frauenberger. " Interactive Sonification of Spreadsheets". in *Proc. of the 11th Int. Conf(ICAD).* Limerick, Ireland. 2005. pp 134-139.

[4] T. Stockman and O. Metatla. "The Influence of Screen Readers on Web Cognition", *Proc. of Accessible Design in the Digital World*, 2008.

[5] F. James, "Presenting HTML structure in audio: User satisfaction with audio hypertext," *In Pro, of Conf. on Auditory Display(ICAD)* ,Xerox Parc,1996.

[6] L. Perrucci, E. Harth , P. Roth , A. Assimacopoulos and T. Pun, "*WebSound: ageneric Web sonification tool, and its application to an auditory Web browser for blind and visually impaired users",* In Proc of the 6th Int. Conf. on Auditory Display (ICAD 2000), 2-5 April 2000.

[7] F. James, 1998 "Lessons from developing audio HTML interface*", Proc of the 3rd Int. ACM Conf. on Assistive technologies,* Marina del Rey, CA, United States, pp.27-34. April 15-17, 1998

[8] E. Murphy, R. Kuber, P. Strain, G. McAllister and W. Yu. "Developing sounds for a multimodal interface: conveying spatial information to visually impaired webusers", *Proc. of Int. Conf. (ICAD),* p356-363. 2007.

[9] S. A. Brewster, "Using Non-Speech Sounds to Provide Navigation Cues". *In Proc. Of ACM Transactions on Computer-Human Interaction* ; 5, 3: 224-259. 1998.

[10] M.P. Bussemakers and A. de Haan, "When it Sounds like a Duck and it Looks like a Dog... Auditory icons vs. Earcons in Multimedia Environments", *in Proc of the 6th Int. Conf. on Auditory Display(ICAD)*, Atlanta, US, 2000.

[11] E. Mynatt, ”Transforming graphical interfaces into auditory interfaces for blind users”. *Human-Computer Interaction*,1997, vol.12, pp7-45.

[12] Winberg, F & Bowers, J. (2004) Assembling the Senses: Towards the Design of Cooperative Interfaces for Visually Impaired Users. *In Proc. of CSCW'04*, Chicago, Illinois, USA, November, 2004.

[13] R. Lämmel, “Style normalization for canonical X-to-O mappings”. *Proc. of the 2007 ACM SIGPLAN symposium on Partial evaluation and semantics-based program manipulation.*

[14] http://www.w3schools.com/

[15] Khan, A., and Sum, M., Introducing Design Pattern in XML schemas.2006. Retrieved on: 25th July, 2009. From sun developer network. Website: http://developers.sun.com/jsenterprise/archive/nb_enterprise_pack/reference/techart/design_patterns.html

[16] Zhao, H., Plaisant, C., Shneiderman, B. and Duraiswami, R."Sonification of geo-referenced data forauditory information seeking: Design principle and pilot study". *In Proc of the Int. Conf.* on Auditory Display (ICAD), (2004).

[17] Finlayson J, and Mellish, C., (2005) “The ‘Audioview’ – Providing a Glance at Java Source Code,” *in Pro of Int. Conf. on Auditory Display (ICAD)*, Limerick, Ireland, July 2005, pp. 127-133.

VERSUM: DATA SONIFICATION AND VISUALISATION IN 3D

Kelly Snook

NASA Goddard Space Flight Center,
Scientific Visualization Studio,
kelly.snook@nasa.gov

Tarik Barri

me@tarikbarri.nl

1. INTRODUCTION

Versum (Barri, 2008) is an advanced, interactive 3D audiovisual composition environment which is augmented with a hardware and software front-end system that maps data into the environment for the purposes of exploratory scientific analysis. Originally intended as an audiovisual sequencer for real-time or automatable music and video performance, Versum also provides a unique environment for systematically investigating new data mappings for optimized human cognition of complex datasets.

2. ABOUT VERSUM

Versum provides a theoretically infinite virtual 3D space, which can be inhabited by any number of audiovisual entities. These entities are rendered and manipulated spatially in real-time using a combination of Max/MSP/ Jitter, Java/Processing, and Supercollider exchanging Open Sound Control messages. Currently the entities can take on two basic shapes: spheres and lines, of virtually any size, brightness, and color and can be sonically coupled to a large number of “synths” rendered in supercollider. Defining parameters can vary in time as sequences that can be indefinitely repeated or changed on the fly.

The original Versum control interface consists of three panes: 1) a control window; 2) a navigator window; and 3) an actor window.

The control window displays visual and audio numerical data that provide the user info about any entity that is currently selected. Within this window the user may also change these properties, thereby altering the appearance and sonic output of the entity.

The navigator window provides a 2D cross section through simplified representations of the entities and displays the actor’s view and motion vectors. This view allows for selection of (groups of) entities by clicking on their representations and is zoomable, enabling the user to work on both macro- and microscopic scale. Selected entities can be easily copied, deleted, created and dragged to any position in the virtual space.

The actor window gives a 3D, fully-rendered, visual representation of the virtual space as seen from the viewpoint of the actor which can be seen as a virtual camera with virtual directional microphones attached, moving through the space. The actor’s position, speed and viewing angle can be manipulated with the mouse and several supported controllers. Movements of the actor determine not only what is seen in the actor window, but also what is heard. The actor’s microphones pick up the sounds of entities that are nearby (the closer they are, the louder they will be heard) and each microphone sends its signals to an output on the computer’s audio interface. The amount of virtual microphones can be set to match any amount of audio outputs used, enabling the use of full surround sound setups. As the actor moves, the panning and the volume of these audio signals change.

The three-dimensional nature of both Versum’s imagery and sounds gives the Versum user powerful tools to make explicit use of – and experiment with – the perceptual implications of dynamic spatial distributions of sounds.

We see Versum as a promising experimental environment for data sonification and visualization. Entities and their parameters may be used and adjusted to represent complex datasets, which can be intuitively and efficiently explored by literally moving through them. In the process the user receives information through combined use of visual and sonic pathways, specifically targeted at efficiently exploiting the unique and rich information processing capabilities of the human auditory and visual cognitive systems.

Currently new mappings are explored to determine the best approach for scripts that will automate the process of entering data into Versum, with the objective of eventually creating a scientifically accurate audiovisual representation of the solar system.

3. ABOUT THE COMPOSITION

The sonification/visualization file accompanying this text is the result of musical and visual experimentation with the properties of Versum and its entities in order to explore the consequences of perceiving audiovisual structures and data in this manner. Entities of different sizes and shapes have been used, either moving or static, either pulsating or constant. These entities have been copied, pasted, dragged across space, their parameters tweaked, their velocities adjusted, and so on, to get a feel for the broad scope of new possibilities on the levels of both basic sensory and aesthetic perception. The listener can hear the spatial distributions and perceived volumes of the entities change as the actor moves through space. Also Doppler effects are audible as the actor passes moving entities.

4. REFERENCES

T. Barri, *Versum: Audiovisual composing in 3D,* Copenhagen, DK: ICAD, 2009

TWO OR THREE THINGS YOU NEED TO KNOW ABOUT AUI DESIGN OR DESIGNERS

Myounghoon Jeon

Sonification Lab.
Georgia Institute of Technology
654 Cherry Street, Atlanta, GA, 30332 USA
mh.jeon@gatech.edu

ABSTRACT

This paper presents an overview of the work on Auditory User Interface (AUI) design at the LG Electronics Corporate Design Center, where the author was responsible for all of the AUI designs from 2005 to 2008. The definition and strategy of AUI design is covered, as well as the role of AUI designers at the global company. Details on process, methodology, and design solutions of four big projects are provided. The paper also discusses how a practitioner's perspective is related to theoretical framework of auditory display. The review of this practical AUI design aims to inspire other practitioners and researchers on auditory display and sound design, and to facilitate communication in the ICAD community.

1. PRELUDE

This report introduces several aspects about Auditory User Interface (AUI) design and designers in a major electronics company and specifically aims to answer the following questions: What is AUI design and who are AUI designers? (Section 1); What type of projects do AUI designers do? What is the process of the project? How is the sound evaluation conducted? (Section 2-5); and what skills are needed for AUI design? (Section 1 and 6).

Prelude provides the definition of AUI design; and AUI designers' roles, functions, and partners at the company. This is followed by four Themes, each of which presents an overview of the big project(s) in which AUI designers can have a critical role. In each Theme, first a project concept is briefly explained. In Overall Process, meta-process of the entire project is presented. Then, in the Details and Implications, more detailed procedures, results, and related implications of the project are described. At the end of the Theme, there is a Cadenza section, which covers design considerations and suggestions. Finally, in Coda, the multi-disciplinary characteristics of AUI design are highlighted. The process delineated in this paper is relatively high level, and results are partially redacted due to company confidentiality.

1.1. Definition of AUI Design at the Company

What is AUI design? It can be defined literally by its components: "Auditory," "User," and "Interface."

Auditory. AUI designers create and manipulate auditory entities. They include non-speech sounds such as earcons [1] (e.g., function feedback sound), auditory icons [2] (e.g., analog ring sound and camera shutter sound), music (e.g., background music of visual design and the system booting), warning signals (e.g., user errors and system errors), and speech sounds (e.g., voice guidance and speech recognition).

User. AUI designers have to answer what kind of sounds users prefer and dislike, and what components of sound affect those responses. Since music and sound have a strong impact on affection or emotion [3-6], AUI designers are required to be familiar with emotion-related research as well as information aspects of sound. Therefore, they iteratively evaluate their own sound products using ample methodologies with target users.

Interface. Interface designers are required to know the specifications of the machine or the system and expected to understand system's functions and users' tasks of that particular interface. Because AUI is a type of user interface design, it includes not merely making ring tones of mobile phones, but also systematically planning and applying all of the sounds in relation to the user interface. Thus, it is different from composing artistic music by just personal inspiration.

In brief, AUI design involves the plan, analysis, creation, management, and evaluation of the product sounds. The results of AUI design should be proper for the function and image of the interface and adequate for users' needs and preference. For more academic definition of auditory display and sonification, see [7, 8].

1.2. Roles and Functions of AUI Designers

Despite its increasing importance, the position of AUI designers has not settled well with companies. AUI designers are still missionaries of AUI design itself; they have to teach and preach what AUI is and why AUI designers are needed in companies in addition to User Interface (UI) designers and Graphical User Interface (GUI) designers.

In Korean companies, AUI designers belong to User Interface Team regardless of whether it is under a design department, an engineering, or other department. Similarly, in LG Electronics, AUI designers worked with UI designers at the Corporate Design Center.

AUI designers usually work with three types of designers in the Design Center: Product designers are responsible for the overall concept of the product and outer shape; UI designers devise the user interaction, focusing on the control panel and sketching the blueprint of the overall user interface; and GUI designers are responsible for more graphical and aesthetic implementation of the UI design. AUI designers consult a product designer (and usually a hardware engineer and a representative of the product planning team) if they use any sound or voice in the product and if they can change or add speakers and amplifiers. AUI designers also identify auditory user interface logic with UI designers and decide when, where, and how the product should generate sounds. Along with GUI designers, AUI designers figure out how to conceptualize the most proper brand image and how to synchronize visual and auditory scenes.

Additionally, AUI designers mainly work with two types of engineers: hardware engineers and software engineers. With them, AUI designers discuss the specifications and location of buzzers, speakers, and amplifiers. After completing sound implementation, AUI designers have to tune the sounds in the real product because sounds are likely distorted due to software and hardware issues.

A job description for AUI design naturally includes MIDI (Musical Instrument Digital Interface) sequencing, composing and arranging music, mixing and mastering, and making sound files. They generally work with more than one partner company (i.e., sound company), but it is usually recommended for AUI designers to be able to deal with any kinds of sounds for themselves.

For analysis and evaluation of the sounds and interfaces, AUI designers need methodological frameworks. To this end, knowledge of Human Factors, Engineering Psychology, Human Computer Interaction, and some statistics can be important assets. AUI designers have to effectively report and present their work and are expected to participate in ICAD, as well as other relevant conferences such as APSCOM (Asia-Pacific Society for the Cognitive Sciences of Music) and ICMPC (International Conference on Music Perception and Cognition).

2. THEME A: AUI GUIDELINE [9]

The goal of this project was to develop a cognitive and affective AUI guideline according to product groups of household appliances so that auditory signals from the product could be intuitively mapped to their functions. This project was needed because users' mental model or expectancy has not been satisfied with arbitrarily matched sounds, which were inconsistent within and between products, and resulted in users' annoyance. Additionally, it was an attempt to overcome the GUI centered interface design and provide users with enriched multimodal user experience. The initial sound generation specification of this project was limited to a buzzer.

2.1. Overall Process

This project specifically focused on mappings between parameters of auditory signals and the functions of the products. To make these guidelines, we traced numerous considerations in creating auditory signals and extracted both general and specific guidelines. The overall process is as follows.

1) Gather general AUI guidelines
2) Analyze AUI needs based on use scenarios
3) Conduct Focus Group Interviews (FGI) and a survey on the use of AUIs in household electronic appliances
4) Conduct a Function Analysis according to product groups
5) Analyze parameters of auditory signals and create sound pools
6) Apply music rules and extract sound samples for experiments
7) Construct cognitive and affective dimensions of auditory signals
8) Develop prospective sound samples mapped to a guideline
9) Evaluate appropriateness of auditory signals and develop a final guideline

2.2. Details and Implications

In each step of the procedure, we obtained a number of basic results and specific user data. In the first step, we investigated related literature and gained general information and some guidelines on the use of sounds (limited to non-speech sounds, but beyond the specific application for household appliances), so that we could use them to improve users' understanding of the system and enhance users' performance and satisfaction. In the next step, we created some plausible scenarios about the use of various electronic devices. From those user scenarios, we identified users' mental model and particular situations that require auditory signals. Additionally, we listened to 22 household wives' notions (we call it VOC, "Voice Of Customers") about the current status of the AUIs in their household appliances using FGIs and a simple survey. They provided us with what is preferred and what is needed to improve for the auditory signals of those products. As a result of FGI, we found that household wives considered washing machines and microwaves as the most important products, which should use sounds cautiously. Meanwhile, we analyzed functions of six household electronic appliances including refrigerators, Kim-Chi refrigerators, air conditioners, washing machines, dish washers, and microwaves. A Function Analysis [10] provided us with a very useful taxonomy for application of not only AUIs but also VUIs (Voice User Interfaces). Since the application of different sounds for each of the functions might be auditory pollution, we intended to categorize the functions as meta-function groups for the application of minimal and adequate sounds. See Table 1 for the details.

We chose several attributes of sounds within the limitations of the physical specification of the buzzers used in the products. Finally, the number of notes, frequency and frequency range, melody pattern (including polarity), duration of the entire sound, and tempo of the sound were considered. Timbre, which is a critical factor in mapping data to sounds [11], was excluded because this study was limited to the use of buzzers which can generate only a single timbre. Based on expert consultation and literature analysis, 45 sounds were finally created for the experiment.

Sound parameters of these samples were mapped to the functional hierarchy within the product. It is very similar to the application of earcons for hierarchical menus (e.g., [12-15]) except that sound was mapped to functional hierarchy instead of menu hierarchy. Moreover, it included more specific guidelines for musical parameters than previous works.

To match sounds with functions, participants rated the appropriateness of every single sound with every function. The results indicated that the signals devised as the specific functions were rated high for the intended functions. This demonstrated that the rules which we applied for the sound samples were valid for users' mental model and expectancy for AUIs in the household products.

In addition to this direct mapping, we tried to match sounds with functions using affective words as a medium between them. The results were also positive. For example, the function, 'Power on' was linked with the words 'alive' and 'vivid'; and increasing sounds. 'Error' was correlated with 'embarrassed' and 'nervous'; and repetitive sound patterns.

Table 1: Functional Grouping of Household Products for AUI

Functions	Descriptions	Examples
Power on/off	Turn on/off the products	Most of the electronic products
Horizontal shift	Change the mode between the same levels	Select the type of food in the microwave
Vertical shift	Change the level up and down	Decrease temperature in the air conditioner
Function on/off	Start/pause the general functions	Start the washing machine after the option settings
Inform	Inform the end of the function /ask a user's action	Finish the washing cycle/ microwave
Warn	Warn the system's or user's errors	Door is open in the refrigerator
Special functions	Play/(pause) the special functions	Brand-specific wind in the air conditioner

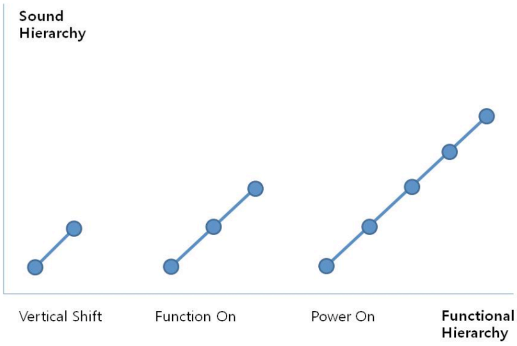


Figure 1: A conceptual figure of sound hierarchy mapped to functional hierarchy. Multiple sound attributes were differentiated due to functional hierarchy including the number of notes, duration, and the range of frequency.

To compare newer sounds with the existing sounds in the product usage context, participants evaluated two sets of sounds (new and existing) on the information architecture of the real product image using the low fidelity software-based prototypes (MS PowerPoint). The results showed that the new sounds obtained higher scores on the appropriateness, preference, and overall satisfaction scales.

Finally, professional composers created several sound sets fitting the guideline details, the brand identity of LG Electronics, and each product group image.

This project provided a meta-guideline for the AUI of LG Electronics household products. The process consisted of several aspects of interaction design such as participatory design of the target user group and user evaluation. Nonetheless, it still had several shortcomings. Since these were buzzer applications, we excluded factors such as timbre, loudness, and harmony of the sounds. Additionally, the last evaluation was conducted using low fidelity prototypes. In the next project, these points were improved to attain higher validity.

2.3. Cadenza

1) Several considerations are needed to decide frequency range: masking, buzzers' prominent frequency, less sensitivity to high frequency of old adults.
2) Buzzer-generated frequency is sometimes not same as musical frequency (e.g., 1045Hz vs. 1046.5Hz for C6). Thus, AUI designers have to tune in the real product.
3) Consider looping (repetition) of the sounds when making sound patterns.
4) In general, we use sounds shorter than 2 seconds.

3. THEME A': VUI GUIDELINE [16]

Theme A' is a variation of the Theme A. While Theme A deals with a meta-guideline for auditory user interface, Theme A' deals with a meta-guideline for voice user interface. Note that this VUI guideline includes a speech recognition part, but focuses more on voice feedback and voice guidance. Indeed, speech recognition is also a critical area that an AUI designer has to cover at an electronic company. Speech recognition is more often addressed for automotive user interface than other interfaces. We made a separate guideline for automotive user interface (e.g., [17]), but see the Volkswagen group's recent research [18] for a more accessible reference which discusses practical issues relevant to speech recognition in the vehicle. Voice feedback once disappeared from the household appliances market, but has reappeared since early 2000 (e.g., LG Electronics and Samsung Electronics) due to improvements in the quality of Text-to-Speech (TTS) and speech recognition techniques.

The use of speech is the most basic method to convey information or feedback auditorily in user interfaces. Therefore, guiding usage procedures of the product via voice user interface can enhance usability and increase product value [19, 20]. Moreover, since speech delivers more specific meaning than non-speech sounds does, it can make user interfaces more efficient and satisfactory [21]. We found some research on voice user interface, related to command vocabulary and command syntax [22] or combining human speech and TTS [23]. However, there has been little research on detailed voice feedback for overall functions of specific household electronic appliances. In this study, again, 6 groups of household products were analyzed (only the microwave was replaced by an oven range from the AUI project of Theme A).

3.1. Overall Process

The procedure was very similar to the AUI guideline project except that we devised a higher fidelity touch screen prototype of a voice user interface for each product. It is one of the most critical improvements from the AUI guideline project, which allowed for more external validity.

1) Conduct a Function Analysis of 6 products
2) Conduct a Task Analysis of respective use cycles
3) Extract usability issues (FGI)
4) Create prototypes of 6 products with VUI
5) Conduct an experiment
6) Finalize the VUI Guideline

In this study, in addition to a Function Analysis, we did a Hierarchical Task Analysis [24] to gain more detailed task procedures for each product. As a result of analysis, we divided products into two groups. Washing machines, dish washers, and oven ranges belonged to the product type that has a start and an end of the task in the product cycle. In this product category, users have to wait for the end of the system operation after their final input. For (Kim-Chi) refrigerators and air conditioners, if users manipulate the operating status, it is reflected immediately by the system. Guidelines should be different for these two types of products.

Based on design experience and Focus Group Interview results, we extracted major usability issues for voice user interface of electronic goods as follows: physical attributes of voice (e.g., gender, tone, and intonation), honorific expression according to usage context (mainly in Korean, thus excluded from this paper), simplicity of VUI (e.g., duration and in setting value guidance), and type of informing/warning. Including each issue in four sessions, we conducted an experiment with 34 house wives. We measured preference, appropriateness, task errors, and reaction time. The experiment was composed of pre-analysis and practice session, experiment, post-interview, and survey. In a practice trial, participants could get familiar with the touch screen prototype. After completing the practice trial and basic demographic questionnaire, participants entered the Usability Testing room. During the experiment, one camera recorded overall behavior of the participant and the other camera captured the touch screen. The questionnaire for each condition was composed of a seven point Likert scale.

3.2. Details and Implications

3.2.1. Physical Attributes of Voice

In voice user interfaces, physical attributes of voice could be critical factors because cognitive efficiency and affective satisfaction with the interface might be different depending on them. Participants rated female voices higher than male voices on preference and appropriateness scales. Moreover, notions of participants about male voices were negative. They noted that a male voice was unfamiliar and awkward for electronic goods. This tendency to favor female voices is consistent with previous research [23, 25].

For voice tone, we obtained unexpected results. Participants preferred the usual tone of the female voice over the high tone of the female voice. Usually, in a voice response system, a female voice of high tone is considered more vivid and kind, and thus preferred. However, in this study, people favored the common tone because they considered it as more reliable and comfortable for the electronic appliances. We might infer that since the household appliances are used every day at home, the result might be different from the case of a voice response system, which is used infrequently by the same user.

Intonation including inflection and stress is one of the most important factors for expressiveness of the speech [26] and could be implemented in various ways. In this study, it was limited to two types: dynamic and general. We found that if it was not too exaggerated, there was little difference in preference between the two types of voices.

3.2.2. Simplicity of Voice Expression

As mentioned earlier, while the use of speech is the clearest way to convey meaning aurally, it requires some time. If user's operations are repetitive, the long duration of the same voice feedback gets even worse. From this perspective, we attempted to identify the most proper speech expression in repetitive functions such as mode change or value adjustment.

We implemented and compared three types of sounds: value + predicate (e.g., in the air conditioner, "25 degree"), value-only (e.g., “25”), and directional non-speech sounds (e.g., two notes such as ‘Sol’, ‘La’ for increasing value). Participants liked the second type, value-only the most. They reported that value + predicate was too long and annoying. Non-speech sounds were reported as not distinct.

In addition to this issue, we examined one more issue related to simplicity. In some household appliances including washing machines and ovens, there are several steps in which the user has to set up values before the system’s operation. For these procedures, we can apply voice feedback after either every single setting or all of the settings. To test this, we implemented three conditions, again: voice feedback after every setting value; only final voice feedback wrapping up all setting values; non-speech sounds after every setting and final voice feedback once wrapping up all setting values. As might be expected, participants preferred the last condition, a combination of non-speech sounds and voice.

3.2.3. Informing/Warning Voice

The existing household products generated only non-speech sound signals when they required the user’s action or warned the user that there is an error. Therefore, users have to approach and check the control panel of the product. Based on FGI, the use of voice in informing/warning situations was expected to be very helpful. Accordingly, we devised three conditions for that scenario as follows: non-speech sounds only, voice only, and non-speech sounds + voice. During the experiment, while participants were watching TV, informing/warning sounds were unexpectedly generated. Then, participants had to come to the prototype and select the proper action among the options. Reaction time of non-speech sounds only was the longest, followed by non-speech sounds + voice, and voice only. However, in the appropriateness and preference scale, participants rated non-speech sounds + voice as the highest.

3.3. Cadenza

1) We used a combination of human speech for the important or repetitive parts (e.g., greetings when booting) and TTS for the less important parts (e.g., respective menu items).
2) People preferred a dubbing artist’s voice over a child’s voice for a GUI animation character (e.g., the penguin) because of the clarity of the voice and reliability of the product.
3) One of the downsides of the voice is that users have to listen to the same length-voice whenever they manipulate the same control even after they become familiar with it. To compensate for this, irrespective of dials or buttons, we apply non-speech sounds + silent interval + voice. Therefore, when users are accustomed to the interface and use the control rapidly, they hear only short non-speech sounds signal.
4) Results of the analysis with regard to age showed that older adults preferred a voice user interface more than young adults. It might be because older adults are less familiar with the electronic goods and they need more guidance of using the products. Gradual loss of vision might be one of the reasons for this.

4. THEME B: EMOTIONAL PALETTE [27]

Whereas Theme A and A' focus on the functional mapping of sound and voice, Theme B and B' focus more on the emotional (or affective) mapping of sound even though they still have functions as well.

Emotional design is becoming more and more important. However, the systematic approach to integration of emotional user experience elements in the product design has rarely been tried. The goal of the Emotional Palette project was 1) to make a common design identity (or language) in order to foster communication among designers of each design section at the company, and with outside partners and 2) to find optimized combinations of basic elements through user study and to build design guidelines for affective design elements. Designers have their own feelings and standards about design elements. To express ‘energetic’, a designer may want to choose ‘red’, but another may choose ‘black’. Even if they all want to use ‘red’, each ‘red’ might be differently imagined and expressed (Think about one of the important issues of Cognitive Science, ‘Qualia’, see e.g., [28]). This variance is good for the creativity of the work, but sometimes hinders the desirable congruency to represent corporate identity. There have been several studies about design elements and emotional factors [29-32], but none is comparable with this study in scale or methodology.

The methodology of this study mainly relied on the Kansei engineering [33], or the Sensibility Ergonomics [34], which is pervasive in academic and industrial community in Asia (but not limited to Asia, see, e.g., [35]). In this type of research, researchers usually adopt affective words as a medium between physical elements and emotion (or affection) about a certain domain.

In this study, first, we extracted affective keywords fitting corporate design directions based on design trend analysis. Then, user experience elements were created and matched with affective words. Finally, a prototype system was made to guide the design of affective factors in electronic products. In this study, user experience elements were defined as color, material & finishing, and sound.

4.1. Overall Process

1) Extract 31 affective keywords through various document analysis and trend analysis
2) Create user experience element stimuli sets (each set of color, material & finishing, and sound)
3) Measure appropriateness for participant’s self-image and for various product groups’ image, and preference depending on each of user segmentation groups.
4) Construct respective sensibility dimensions of each product group that contain mapping between keywords and design elements according to user segmentation groups.

5) Develop a prototype system based on the results of the research above.

4.2. Details and Implications

4.2.1. Extracting Trend Keywords

First, 120 affective words were extracted through trend analysis by experts group (*Nelly Rodi* in France). Additionally, we collected 342 trend adjectives by means of literature and previous research analyses, and designers' free association reports. Then, user experience designers—two color experts, two material & finishing experts, and two sound experts—extracted 50 keywords by rating the validity of collected sensibility adjectives. The standards for the validity were the degree of reflection of design trend and the appropriateness for each design element. We finally selected 31 design trend keywords eliminating simple description or technical words.

4.2.2. Designing and Matching Emotional Design Elements

Each expert group composed appropriate stimuli for the 31 affective words. Color was shown as color bars and material & finishing displayed as representative images. A few stimuli sets were chosen and elaborated by domain experts before the experiment.

For the mapping experiment between design elements and keywords, a total of 320 participants were recruited according to each of product group's market segmentation. Participants were provided with sets of design elements, and asked to answer several questions about preference and fit for self-image and for each product group. In the case of sounds, they could listen to one stimulus repeatedly. As a result of the experiment, we could obtain a mapping between each design stimulus, affective words, and user segmentation using Multi-Dimensional Scaling and Correspondent Analysis.

4.3.3. Developing Emotional Design Guideline System

Based on the results above, a prototype system was developed to guide designers to properly combine affective elements with the design concepts (see Figure 2). Above all, designers can see 31 adjectives on the 2-Dimensional coordinate system. When a designer chooses a trend keyword, the proper color bar and the material & finishing images are recommended on the right side of the system with a representative image on the bottom left. Two or three sounds that were rated highly on the concept can be heard via a sound tab. Moreover, designers can check the appropriateness of design elements according to users' segmentation. Finally, diverse combinations of all of these elements can be presented at once. It is worth noting that these elements may function as recommended samples to construct a common language between hundreds of designers for corporate design identity, but may not function as a predefined material for real products.

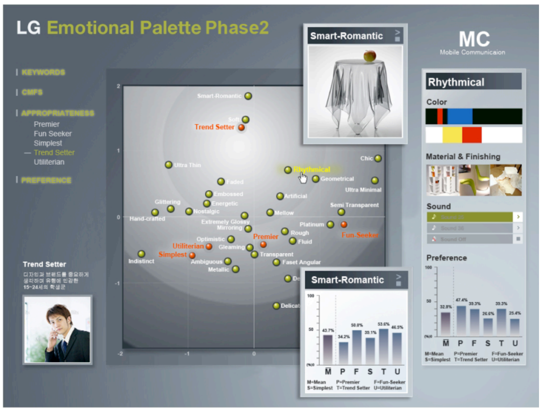


Figure 2: A screen shot of Emotional Palette prototype system for mobile phones. Design elements mapped to keywords and segmented user groups.

5. THEME B': SONIC LOGO

Even though the preceding three projects are all related to improving product identity, the most relevant project to corporate brand identity for AUI designers is creating a sonic logo, (also called sonic brand or jingle, imagine Intel's sound). For this type of project, AUI designers need to collaborate with other parts of the company beyond designers, including marketing and product-planning, etc. Among many of the sounds used in electronic products, customers and users are likely to remember the sound generated when they turn on or off the system. For example, you may be easily able to recall the Microsoft Windows' opening or closing sound. Therefore, our strategy was to create a power sound as a sonic logo and develop and apply it to other areas such as demos and advertisements on TVs, radios, and web sites. Even though users' vision is occupied with other tasks, this sound can remind them of a unique brand image. Therefore, the sonic logo should be matched with the image of the corporate identity.

5.1. Overall Process

We attempted to develop a sonic logo following general procedure, but it may vary on a case by case basis.

1) Position the image of our own company brand
2) Position the image of the existing jingles of other companies
3) Position the image of the product groups
4) Create and select sound samples
5) Develop theme variations and music

The overall process might look simple, but each step includes complex collaborations and considerable reports and decision making. This is what we consider the 'iterative design and evaluation process' itself.

5.2. Details and Implications

If the company already has a brand image positioning map, the project can be easily started. It is because brand image positioning is 1) the first step of this project; 2) the most difficult part, having to draw an agreement from each part of the company; and 3) outside the designers' job. Meanwhile, we collected and analyzed competitors' jingle sounds. Analyses include the number of notes, duration, key, rhythm, type of timbre, use of voice, chord progression, type of coda, usage scenes, and overall impression. Results of analyses helped us make our sound distinct from others while keeping it comparable to others in quality.

One of the easiest ways to differentiate feelings between product groups is to apply a different timbre for each sound (this is impossible for the buzzer system, though). For example, we might create one simple melody contour for all the household appliances and vary it with different timbres: Woodwinds for air conditioners; bell sounds for refrigerators; and water drop sounds for washing machines.

System booting sounds are a type of sonic logos. When designing the system booting sound, various collaborations with other designers occur. For the case of one of the premium TVs, product designers requested a specific scenario of 'power on' for the TV (like the sun rise from the Milki-Way). Thus, we created several samples of corresponding sound for that scene. In another case of a blue-ray player, GUI scenes such as the beginning of the movie were first conceptualized and were followed by AUI alternatives. One interesting case was the design of the booting sound of a Personal Navigation Device (PND). In our PND models, there was no booting sound. However, in that specific model, the system booting lasted too long due to a hardware issue. In a development meeting, we came up with an idea of using sound to compensate for this boring time. This can be a good example of an application of Cognitive Psychology to AUI design. According to the Cognitive-Attention Theory, doing a concurrent task while monitoring time passage leads to temporal estimates that are shorter than without the task [36]. We expected that adding sounds might make users humming and be less annoyed while waiting for the system to boot up. Consequently, in this particular model, AUI was created first and was followed by appropriate GUI designs.

5.3. Cadenza

1) Most of the companies uses major keys, but some (e.g., 'National' in Japan) uses minor keys in their jingle.
2) Sonic logos can be made without melody contours (e.g., Macintosh), but in this case, it is generally difficult to apply the sound to a buzzer system.

6. CODA

This paper presented a short review of AUI design at an electronic company. Since Pythagoras and Plato, there have been attempts to create the corresponding relations between the physical phenomena of sound and psychological responses to them. Along the same line, a semiotic approach to today's sound design [37] looks valid because it is the science of mapping auditory signal to meaning (e.g., function, hierarchy, image, and emotion). Even though music or sound is very subjective and depending on emotional state, to be more scientific we have adopted language (e.g., functional words and adjective words) as a logical linkage between sound and mind because most people have expertise in their language and use it as an effective communication tool in everyday lives.

The methods and processes used in this paper are not all we have used and may not be ideal. Notably, the guidelines presented here are for specific electronic products of the particular company. Moreover, guidelines sometimes do not work well when we design for a real, specific product because there are always unexpected issues. Therefore, they may not function for other products or in other contexts. Indeed, the ICAD community has not developed the standard sounds, methods, tools, and evaluation techniques with respect to auditory user interface design yet. This type of case study might contribute to laying a brick for that foundation.

Auditory display is an interdisciplinary science. The author has studied sociology, cognitive science, engineering psychology, and film scoring. Other areas that current AUI designers majored in include computer science, classical music, computer music, and user interface design. However, irrespective of majors, whoever likes or is interested in music (auditory), mind (user), and machine (interface) can tackle the challenge.

7. ACKNOWLEDGEMENT

I very much appreciate Jeff Lindsay and Yarden Moskovitch for proofreading.

8. REFERENCES

[1] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and icons: Their structure and common design principles," *Human-Computer Interaction,* vol. 4, pp. 11-44, 1989.

[2] W. W. Gaver, "Auditory icons: Using sound in computer interfaces," *Human-Computer Interaction,* vol. vol.2, pp. 167-177, 1986.

[3] P. Kenealy, "Validation of a music mood induction procedure: Some preliminary findings," *Cognition and Emotion,* vol. 2, pp. 41-48, 1988.

[4] J. D. Mayer, J. P. Allen, and K. Beauregard, "Mood inductions for four specific moods: A procedure employing guided imagery vignettes with music," *Journal of Mental Imagery,* vol. 19, pp. 133-150, 1995.

[5] J. Panksepp, and G. Bernatzky, "Emotional sounds and the brain: The neuro-affective foundations of musical appreciation," *Behavioural Processes,* vol. 60, pp. 133-155, 2002.

[6] V. N. Stratton, and A. H. Zalanowski, "The

effects of music and cognition on mood," *Psychology of Music,* vol. 19, pp. 121-127, 1991.

[7] B. N. Walker, and G. Kramer, "Auditory displays, alarms, and auditory interfaces," in W. Karwowski (Ed.), *International Encyclopedia of Ergonomics and Human Factors*, (2nd ed.) (pp. 1021-1025). New York: CRC Press, 2006.

[8] T. Hermann, "Taxonomy and definitions for sonification and auditory display," in 14th Internatinal Conference on Auditory Display (ICAD2008), Paris, France, 2008.

[9] J. Lee, M. Jeon, and K. Han, "Developing the design guideline of auditory user interface of digital appliances," *Korean Journal of the Science of Emotion & Sensibility,* vol. 10, no. 3, pp. 307-320, 2007.

[10] C. D. Wickens, S. E. Gordon, and Y. Liu, *An introduction to human factors engineering*, New York: Addision Wesley Longman, 1998.

[11] B. N. Walker, and G. Kramer, "Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making," *Ecological psychoacoustics*, J. Neuhoff, ed., New York: Academic Press, 2004.

[12] S. A. Brewster, P. C. Wright, and A. D. N. Edwards, "A detailed investigation into the effectiveness of earcons," in the 1st International Conference on Auditory Display, Santa Fe, USA, 1992, pp. 471-478.

[13] S. A. Brewster, V. P. Raty, and A. Kortekangas, "Earcons as a method of providing navigational cues in a menu hierarchy," in *HCI'96*, 1996, pp. 167-183.

[14] S. A. Brewster, "Using Non-speech sounds to provide navigation cues," *ACM Transactions on Computer-Human Interaction,* vol. 5, no. 3, pp. 224-259, 1998.

[15] G. Leplatre, and S. A. Brewster, "Designing non-speech sounds to support navigation in mobile phone menus," in the 6th International Conference on Auditory Display, Atlanta, GA, 2000, pp. 190-199.

[16] H. Chae, J. Hong, M. Jeon *et al.*, "A study on voice user interface for domestic appliance," *Korean Journal of the Science of Emotion & Sensibility,* vol. 10, no. 1, pp. 55-68, 2007.

[17] S. Kim, J. Choe, E. Jung, S. *et al.*, "A study on the navigation menu structure with size screen," in the Korean Conference on Human Computer Interaction, Kangwon, Korea, 2008, pp. 380-385.

[18] J. Chang, A. Lien, B. Lathrop *et al.*, "Usability evaluation of a Volkswagen group in-vehicle speech system," in the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2009), Essesn, Germany, 2009, pp. 137-143.

[19] R. M. Mcleod, "Human factors issues in speech recognition," *Contemporary Erogonomics*, E. J. Lovesey, ed., London: Taylor and Francis, 1994.

[20] H. C. Michael, P. G. James, and B. Jennifer, *Voice user interface design*, Boston: Addison-Wesley, 2004.

[21] C. Nass, and L. Gong, "Speech interfaces from an evolutionary perspective," *Communications of the ACM,* vol. 43, no. 9, pp. 36-43, 2000.

[22] D. Johnes, K. Hapeshi, and C. Frankish, "Design guidelines for speech recognition interfaces," *Applied Ergonomics,* vol. 20, no. 1, pp. 47-52, 1989.

[23] L. Gong, and J. Lai, "Shall we mix synthetic speech and human speech? Impact on users' performance, perception, and attitude," in the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA, USA, 2001.

[24] E. Hollnagel, "Task analysis: Why, what, and how," *Handbook of human factors and ergonomics*, G. Salvendy, ed., New Jersey: John Wiley & Sons, Inc, 2006.

[25] M. K. Margulies, "Male-female differences in speaker intelligibility: Normal versus hearing impaired listeners," *Journal of Acoustical Society of America,* vol. 65, no. S1, pp. S99-S99, 1979.

[26] T. V. Raman, *Auditory user interfaces: Toward the speaking computer*, Boston: Kluwer Academic Publishers, 1997.

[27] M. Jeon, U. Heo, J. Ahn *et al.*, "Emotional Palette: Affective user eXperience elements for product design accoring to user segmentation," in the International Conference on Cognitive Science, Seoul, Korea, 2008.

[28] G. Harman, "Some philosophical issues in Cognitive Science: Qualia, intentionality, and the mind-body problem," *Foundations of Cognitive Science*, M. I. Posner, ed., pp. 831-848, Boston: MIT Press, 1998.

[29] L. M. Brown, and J. J. Kaye, "Eggs-ploring the influence of material properties on haptic experience," in the 2nd International Workshop on Haptic and Audio Interaction Design, Seoul, Korea, 2007, pp. 1-2.

[30] J. Lee, M. Jeon, Y. Kim *et al.*, "The analysis of sound attributes on sensibility dimensions," in the 18th International Congress on Acoustics, Kyoto, Japan, 2004, pp. 1795-1798.

[31] D. Payling, S. Mills, and T. Howle, "Hue music-creating timbral soundscapes from coloured pictures," in the International Conference on Auditory Display, Quebec, Canada, 2007, pp. 91-97.

[32] R. Bresin, "What is the color of that music performance," in the International Computer Music Conference (ICMC2005), Barcelona, 2005, pp. 367-370.

[33] M. Nagamachi, "Kansei Engineering as a powerful consumer-oriented technology for product development," *Applied Ergonomics,* vol. 33, pp. 289-294, 2002.

[34] K. Lee, "Sensibility Ergonomics: Needs, concepts, methods, and applications," *Korean Journal of Ergonomics,* vol. 17, no. 1, pp. 91-102, 1998.

[35] J. A. Kleiss, "Characterizing and differentiating the semantic qualities of auditory tones for products," in the Human Factors and Ergonomics Society 52nd Annual Meeting, 2008, pp. 1692-1696.

[36] N. S. Hemmes, B. L. Brown, and C. N. Kladopoulos, "Time perception with and without a concurrent nontemporal task," *Attention, Perception, & Psychophysics,* vol. 66, pp. 328-341, 2004.

[37] A. Pirhonen, K. Tuuri, M. Mustonen *et al.*, "Beyond clicks and beeps: In pursuit of an effective sound design methodology," *HAID 2007, LNCS 4813*, I. Oakley and S. A. Brewster, eds., pp. 133-144, 2007

A METAPHORIC SONIFICATION METHOD - TOWARDS THE ACOUSTIC STANDARD MODEL OF PARTICLE PHYSICS

Katharina Vogt, Robert Höldrich

Institute for Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria
vogt@iem.at, hoeldrich@iem.at

ABSTRACT

The sound of a sonification has, like any sound, a metaphoric content. Ideally, the sound is designed in a way that it fits the metaphors of the final users. This paper suggests a metaphoric sonification method in order to explore the most intuitive mapping choices with the right polarities. The method is based on recorded interviews, asking experts in a field what they expect data properties to sound like. Language metaphors and sounds of the recordings are then interpreted by the sonification designer. The method has been used for developing an 'Acoustic Standard Model of particle physics' with physicists at CERN.

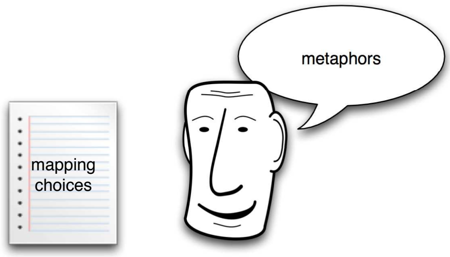


Figure 1: Metaphor sonification method. A questionnaire on sound metaphors and possible mapping choices.

1. MOTIVATION

Conceptual metaphors have been discussed, e.g. by G. Lakoff and M. Johnson [1]. Metaphors help us understanding an idea of a target domain by citing another one in a source domain. Even more fundamental, they shape our perception of reality. Also science builds on existing experiences: *"So-called purely intellectual concepts, e.g., the concepts in a scientific theory, are often – perhaps always – based on metaphors that have a physical and/or cultural basis. The high in 'high-energy particles' is based on* more is up. *[...] The intuitive appeal of a scientific theory has to do with how well its metaphors fit one's experience."* [1, p. 19]

For a *good* sonification design, it would thus be enough to know about the underlying metaphors of a scientific theory *and* the metaphors for sound of these basic experiences. By mapping, e.g., higher energies to what people in our culture perceive as *higher* in sound, a completely intuitive sonification could be created. B. Walker and G. Kramer [2] questioned already in 1995 if there is something like best auditory mappings for certain data properties and what they could be. They tested different mappings which they had assessed as *good* or *bad*, and were surprised by the actual outcome of the test, as the 'bad' mappings actually led to best results. The same authors point out that *"interface designers have usually implemented what sounds 'good' to them"* and conclude that testing with the final users is crucial. An effective mapping cannot be predicted a priori and also the polarity of mapping has to be taken into account. The results are also interesting in the specific context of our data, as they found for instance *"that increasing mass is generally best represented by decreasing pitch"*.

B. Walker conducted several studies in this direction [3, 4]. He implemented magnitude estimations between sound attributes and conceptual data dimensions. Magnitude estimation is a standard psycho-acoustical procedure for studying the dependancy of an acoustic variable on its perceptual correlate (e.g., frequency and pitch). Walker extended the method to conceptual data variables. For data-to-display pairs he found positive or negative polarities (the increase in a data dimension is reflected by the increase or decrease of the sound attribute), and scaling functions, giving also the slope of the dependency. In extensive experiments he showed that polarity and scaling functions matter for the quality of AD, and a priori predictions about the best choice are often difficult but can be determined empirically. For some mappings, the analysis showed unanimous polarities, as, e.g., for velocity to frequency. For most mappings, the positive polarity was dominant. While these results are highly valuable for sonification design, a complete analysis of the sound metaphors of any scientific theory is beyond the scope of creating an AD.

S. Barrass argues, that sonifications should be done in the 'world of sound' that the end-users know. In a physics' related context, e.g., the sound of a Geiger counter is one that can easily be understood, even if the data has nothing to do with radiation at all. *"The Geiger-counter schema also seemed to reduce the amount of time it took naïve users to learn to manipulate the [...] data, and provided a context for interpreting the sounds in terms of the geological application domain."* [5, p. 405] Also in the experiments cited above, different listener groups (e.g., blind and sighted people) chose different polarities as best data display. Walker concludes that *"sonification must match listener expectance about representing data with sound"* and that it *"is also important to consider the perceptual reactions from a more diverse group of listeners"* [4]. While the latter argument meant mainly individuals differing in listening expertise, we argue that also differences in the conceptual understanding of data dimensions play a role. Energy in the context of macroscopic objects might mean something completely different for engineers than in the microscopic view for particle physicists. In accordance to Walker, we assume that general metaphors that are valid in any context can never be achieved. There will not be a general table that a sound designer can simply read-out for any sonification problem. *"As with any performance data that are used to drive interface guidelines, care must always be taken to avoid the treating the numbers as components of a design recipe." [4, p. 596]*

Motivated by these assumptions, we developed a metaphoric sonification method, *metaphor*, Fig. 1. The basic idea is to question scientists in the field about the sounds or metaphors they use or what they expect special data properties to sound like. The method is a sensible starting point for sonification design, that allows informed parameter mapping choices for the designer. It can also be used for event-based methods as earcons or even for model-based sonification, where at least parameter tuning can be adjusted to fit the sound results to the intuition of the domain scientists. The method does not deliver a ready-made sonification design, but rather leaves creative space for the specialist who –by questioning the domain experts– gains insight into their possible 'world of sound'.

An existing approach to support the sonification design process is EarBenders, a database of stories on everyday listening experiences by S. Barrass [6]. He suggested this method in analogy to classical case-based design from human computer interaction, because the sonification community still lacks a considerable amount of case studies of ADs. The database can be accessed, when a new sonification design is demanded for a field, where the designer has no previous experience. One method of searching the database is a metaphorical one, as also Barrass argues that a *"metaphoric design can help when sounds are not a natural part of the design scenario, which is often the case in computer-based applications."* [6, p.51] But even with a large data base, a search for a new sonification problem often does not deliver exact matches.

There have many been different approaches to design guidelines in AD. For an overview, see [7]. Three conceptually different approaches shall be mentioned: the Task and Data analysis by Barrass [6], realized as a systematic questionnaire; the sonification design space map (SDSM) by de Campo [8], a map of quantitative data characteristics; and *paco* (pattern design in the context space), an iteratively evolving data base of design patterns [7].

The power of metaphors

A comment should be given on the human nature of sensorial metaphors. Mappings of conceptual data variables and auditory percepts are rarely homogeneous, i.e. judged similarly by different people, which may partly be a result to learning. But, it may also be intuitive in the sense that cross-modal metaphors are found in common language (e.g., a *tone* color). Martino and Marks [9] suggest this as a form of weak synesthesia as compared to strong synesthesia, where associations between an inducer in one modality cause induced percepts in another (e.g., seeing absolute colors when hearing corresponding tones). While correspondences in strong synestehsia are systematic and absolute, in weak synesthesia they are defined by context. The authors suggest a *'semantic-coding hypothesis'*: high-level sensory mechanisms are involved, which are developed from experience with percepts and language. Thus also language can cause percepts, and these are rather homogeneous within a group of people of the same cultural background.

2. A METAPHORICAL SONIFICATION METHOD

Our metaphoric guideline on sonification design is a similar approach as EarBenders, but for the case that no a priori sound examples exist. It allows the sonification designer to gain insights into the field from a meta-level point of view. The method is based on asking potential sonification users about which sounds they would expect or associate to the data and task. Different kinds of metaphors in the answers are then re-interpreted to the sound domain. The procedure can be generalized as M–ET–APH–OR:

Material: Become acquainted with the data. Define which features should be covered by the sonification. A TaDa (see [6]) may help in this task. Set-up a questionnaire, which may give you cues for the most important metaphors of the domain science. It should have a free, associative part, but also suggest mapping choices including the polarity. Define number and (the professional/ personal) background of the interviewees.

Enregister Talks: Interview domain scientists face-to-face and record the interviews.

Analyse PHrasing: Take notes on the questionnaire, extract and describe the sounds of the recordings. For instance, intra-personal fits or misfits between language metaphors and the produced sounds can be interesting. Collect the sonification ideas that have come up during the interviews. If there is enough data material, do some statistical analysis. Find common (interpersonal) metaphors. List also differing metaphors or cases where, e.g., the polarity of the mapping seems unclear.

Operate with Results: Based on the results of the questionnaire, decide on the best mapping choice and implement it.

The main finding of this procedure is knowledge about the specific metaphors and associations of scientists (or others) in their specific field. As a side effect, ideas for the basic sonification design can come up during the interviews – more, than a single sonification designer would have thought of. Also, if a domain scientist contributes to a sonification in this way, s/he spent already time with it and will be curious about the outcome. Thus the sonifications may be more wide-spread.

The recording of the questionnaire is important, as it is hard to speak about sounds, especially for people who have never done so before. Firstly, the recording allows the interviewees to *make* sounds rather than describing them. Secondly, misunderstandings can be avoided, especially when the interviewee and/or the sonification expert are not native speakers of the same language. It has to be taken into account that most test persons can think of more sounds than they can actually produce. The personal interview is very important, because it helps questioning the outcome of the sounds and interpret the metaphors behind. Finally, the recordings of the discussions can re-assure the sonification designer.

A disadvantage of the *metaphor* procedure is the additional effort. Also for a sonification design in a predominantly exploratory focus, the metaphors collected in the interviews do not help much, as new, unknown data features are searched for. But for any sonification with at least some known structures involved, this method helps for a good mapping choice and more acceptance in the domain community.

In developing the method, a study was conducted following the *metaphor* concept defined above. It is discussed in Sec. 3.

3. TOWARDS AN INTUITIVE PARTICLE DISPLAY

In particle collision experiments, e.g. at CERN, the European Organization for Nuclear Research, different kinds of particles are measured. The most common visual display shows colored tracks of particles that have been produced by a collision, sometimes as a movie. In a short term project in autumn 2009, I conducted a questionnaire on data from CERN, that supported the design decisions for an 'acoustic standard model' of particle physics. The description follows the *metaphor* procedure described above, even if the experiences from the survey were used to create the method.

3.1. Material

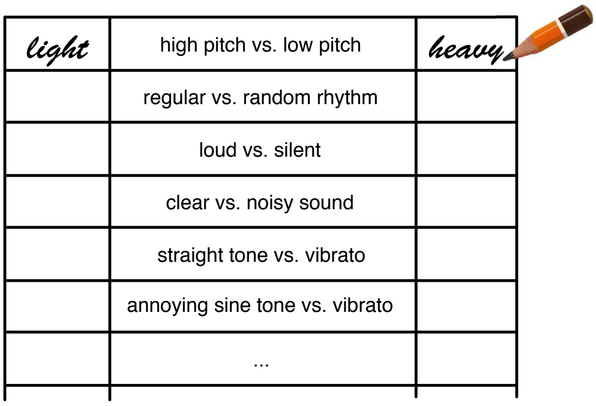
The Standard Model of particle physics describes a framework of three of the four known interaction types and the elementary particles that interact with each other. All visible matter in the universe is constituted by these particles. It is not within the scope of this paper to give a complete overview over the properties of these particles, but a schematic plot is shown in Fig. 5. There are 6 quarks and 6 leptons both for matter and (antiquarks and antileptons) for antimatter. The stable parts of everyday matter is built up by the *up* and *down* quark (constituting the proton and neutron) and the electron and electron-neutrino. Quarks have a so-called color property, and cannot be observed freely: only color neutral objects, as baryons (a blue, a red plus a green quark) or mesons (color plus anti-color, in this example blue and yellow) are observed. Baryons and mesons are both hadrons, as opposed to leptons like the electron and bosonic force carriers like the Higgs-boson (the Higgs has not been observed and is thus a theoretical particle). There are hundreds of particles which are constituted by different quarks, therefore often referred to as a 'particle zoo'.

We elaborated a questionnaire on particles, containing a short introduction and 3 other parts. The participants were chosen from employees at CERN who have studied physics.

After a short introduction, free associations for eight different particles were being asked for: p: proton, p-: antiproton, e: electron, e+: positron, μ: muon, π: pion, $\kappa^{\pm}$: kaon, h: Higgs boson. The particles are the most common (in our data from CERN), and cover the most important features, like mass, matter (vs. antimatter), charge, and quark content (for hadrons). We included the Higgs' boson as the only imaginary particle, because it was a 'hot topic' at the time at CERN and in the media. This part of the questionnaire was recorded.

The second part of the questionnaire was only shown, after the free, associative part has been completed. A table of sound properties with pairs of extreme positions was given (see Fig. 2). We tried to phrase these properties in a general, rather musical wording, avoiding technical terms. The list was open ended and could be complemented by the interviewees having any other ideas.

Then, different particle properties were listed: mass (*heavy vs. light*), matter (*matter vs. anti-matter*), charge (*positively/ negatively charged vs. neutral*), quark content (*up, down, charm, strange, top, bottom*), particle type (*mesonic/*



light	high pitch vs. low pitch	heavy
	regular vs. random rhythm	
	loud vs. silent	
	clear vs. noisy sound	
	straight tone vs. vibrato	
	annoying sine tone vs. vibrato	
	...	

Figure 2: CERN questionnaire - I: Schematic plot of the sound properties' table with an exemplary mapping choice.

baryonic/ leptonic), and excitation, again in an open ended list. They could be chosen and filled into the left or right hand side of the sound properties' table, see Fig. 2. Properties not associated with any sounds were left out.

Finally, personal information including total years working in the field (including studying), specifying the field, years working at CERN, gender, and whether the persons ranked themselves as (partly) musicians, music lovers, or none of these, was collected.

3.2. Enregister talks

All interviews were conducted personally by mysef and had no time limit. In the open part, no additional information was given than a short introduction to the project. If the test persons were comfortable with this, they were asked to mimic sounds they imagined, or else to speak about their associations.

24 people ranging from a diploma student to a Nobel prize laureate have been interviewed (according to [4, p.596], this number is appropriate for such an experiment). Three participants were excluded from the analysis, as they had not studied physics which I only found out during the interview. One person did not want to be recorded or complete the questionnaire, but made some general remarks. One interviewee completed only the first, associative part, but not the fill-out part. Thus, 19 questionnaires were included in the analysis of which two were completed by females and the remaining 17 by males. Five interviewees ranked themselves as *(partly) musicians* and three as *none*, the rest as *music lovers*. The lengths of the interviews averaged around 15 minutes.

Reactions of the interviewees were very diverse. The task of thinking about the sound of particles, or even mimicking them, was too demanding for some: *"I am shocked"* clearly reflects that. Many people reacted in a way, that they were not the right person to ask: *"You know better than we do what to choose"*, or *"What you need is a synestethe!"*. Many participants established a relationship to their actual field of work. For instance, experimental detector physicists would say, *"I am thinking of layers because I am working with detectors and their layers"*. One even extended the notion of a particle detector to the human hear, and suggested to use very high sounds for particles which are hard to detect: *"I am already hard of hearing with high pitched tones"*. Those, who did try to mimic the sounds they thought of, experienced problems with the task. *"I hear my sound and I think - 'Ahh, that's not exactly what I meant'. I cannot produce all the sounds that I imagine."* One participant tried his sounds out several times in order to improve fitting his actual voicing to his imagination.

Nevertheless, 12 people did produce sounds and three participants even suggested specific sounds for all eight particles on the list. The recordings of the free questionnaire part for all particles are available at http://qcd-audio.at/tpc/quest.

Resulting mapping choices of the fill-out part are shown in Tab. 1.

3.3. Analyse Phrasing

For the analysis of the metaphoric sounds, the particle sounds were cut from the recordings and normalized. Also the spoken descriptions were collected, and general ideas for the sonification design extracted. The approaches in the recordings can be summarized as follows:

- Most people started systematizing even in the free, associative part – they are trained physicists. A clear majority suggested to map mass to pitch as a very first association.
- Phonetic or spearcon approaches following the particles' names were often applied. For instance the Higgs' sound was associated with a *"higgs"* or just *"igs"*, or proton became an *"ooo"* and the pion an *"iii"*.
- Many comparisons to the measurement were drawn. E.g., heavy particles crush *loudly*, or particles behave differently in various layers of the detector.
- Some suggestions were very concrete. (The examples cited here were taken into account in the display.)
  - Tone patterns, like J. S. Bach did with his famous b-a-c-h fugue, would allow recognizing particles. Simple particles, like protons, can become something like a bass line.
  - Each quark flavour can have a certain pitch assigned, meaning that hadrons are played as chords (thus baryons would sound as triads, for instance).

Pitch:	mass (18/18), favorit: mass
Amplitude:	mass (7/14), charge (4/14), matter (2/14), favorit: charge *(mass will be used for pitch, and does not need to be mapped twice, as pitch is a very strong mapping factor; charge was cited second most often)*
Rhythm:	lep/ had (3/12), mass(2/12), matter (2/12), individual suggestions (3/12), no clear favorite *in general, rhythm is more associated with the experiment, measurement or data*
Noise component:	lep/ mes/ bar (7/14), matter(3/14), quark content (2/14), favorit: lep/ mes/ bar *(but no clear mapping choice due to inconsistent polarities)*
Vibrato:	exc. (6/14), lep/ mes/ bar (4/14), matter (3/14), charge (2/14), favorit: excitation *(here the problem was different notions of excitation; we referred to ground state and excited states, but this is not reflected in measurements, and was thus often interpreted differently. Still, vibrato would be the favorite mapping for excitation.)*
Timbre:	matter (2/8), exc. (2/8), lep/ mes/ bar (2/8), no clear favorite: *and only few total number of suggestions (possibly, this is concept is too complex)*

Table 1: Mapping choices of the particle properties resulting from the MSM. The number of mentions vs. the whole number of all answers for this property is shown in brackets. Abbr.: lep=leptonic, had=hadroniv, mes=mesonic, bar=baryonic, exc=excitation.

- Matter is a normal sound and anti-matter its reversed playback.
- Particles sound like *cars passing by*, with their passing time and pitch variation depending on their speed.

Some statistical analysis was done, but as only 19 people were taken into account, no significant results have been found regarding different backgrounds. Fig. 3 shows how often particles were mimicked with sound or described (in words) in the associative part of the questionnaire. The Higgs' particle was treated most often, possibly because it is talked about a lot. The Higgs' sounds were often meant to be funny, e.g., a *"tadaa"*, like the theme of a feast, or a *"ka-boum"* for some ground breaking discovery. Neglecting the Higgs', the figure shows that well-known particles as electron and proton are cited most often. There are much fewer associations for rare particles.

Some particle properties were used much more often for mapping suggestions. Many test persons linked mass, the general particle type or matter (vs. anti-matter) to sound properties. Mass, for instance, has a macroscopic meaning that can easily be associated with sounds. The particle type (as hadronic or leptonic) is more abstract. For anti-matter, many explanatory metaphors exist - e.g., an anti-particle was described as its particle *"seen in a mirror"*. The quark content, at the end of the table, is an abstract property and was only cited five times summing all mentions of the 19 test persons together.

The most obvious mapping choice was pitch with mass, heavy mass meaning low pitch. *All* answers in the table were given accordingly (only the direction of the mapping was *once* given contrariwise, high mass being mapped to high pitch). These results are in line with experiments of Walker [4], where also a few (2 out of 19) participants chose

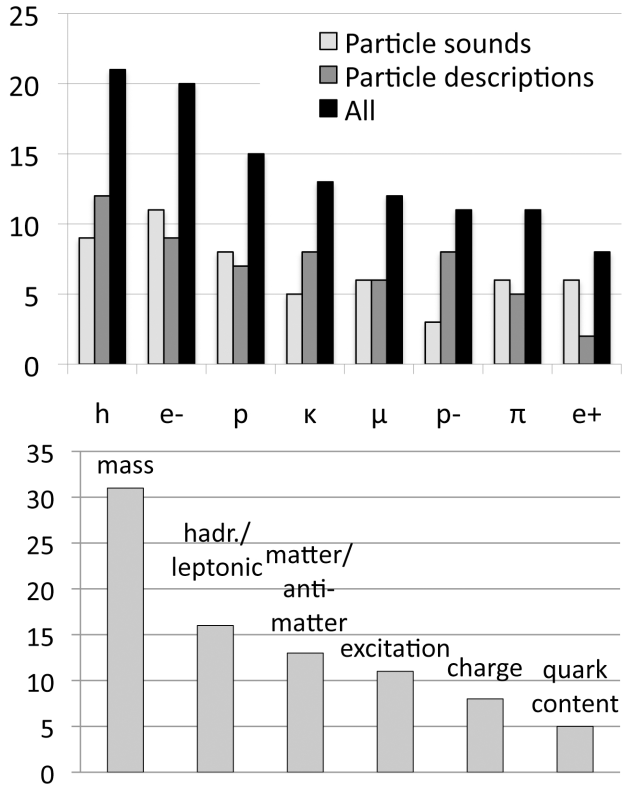


Figure 3: Quantitative results for the CERN questionnaire: *Upper figure:* Overall number of particle descriptions and sound associations, sorted by their sum. *Lower figure:* Number of entries of the particle property into the sound properties' table.

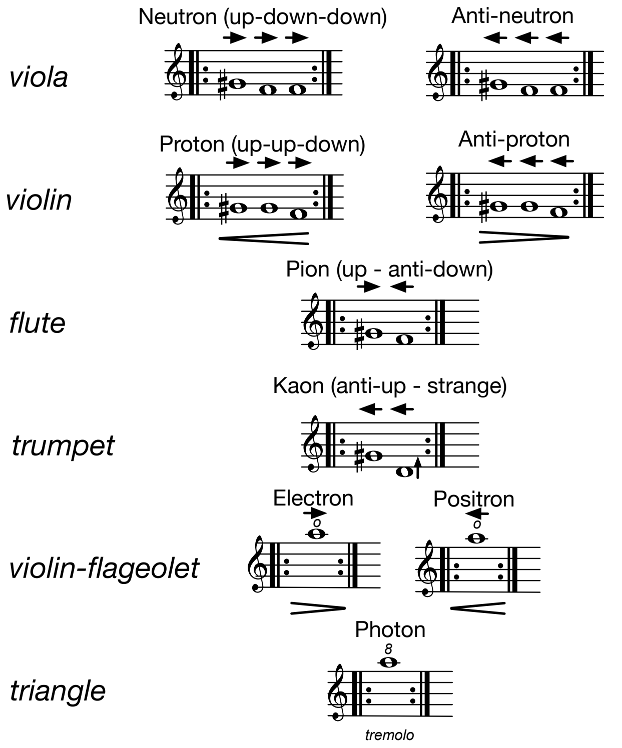


Figure 4: Example for the acoustical standard model. The forward and backward arrow denote the forward or backward playback time for each elementary sound.

opposing polarity for mass to frequency. In general, increasing sound frequency corresponds to decreasing mass.

Results of the sound property table are shown in Tab. 1. The most prominent choices were used for basic mapping decisions: Mass, as a central particle feature, clearly was linked to *pitch*, which is a salient auditory percept. Charge was suggested for *amplitude* second most often after pitch. In general, *rhythm* is more associated with the experiment, measurement or data. There was no clear mapping choice for *noise*, due to inconsistent polarities. Vibrato would be the favorite mapping for excitation.[1] *Timbre* had only few total number of suggestions, possibly because this is concept is too complex.

3.4. Operate with Results

In general, each particle shall be displayed as a recognizable sound of varying length, which is transformed under the dynamics dictated by an experiment. With all knowledge from above, we worked out the following sonification:

Mass is mapped to pitch, and every elementary particle (quarks/ leptons/ bosons) has an assigned pitch. First generation quarks (up and down) form a small, regular interval (a third). The strange quark is a *strange* mistuned fourth, and the charm is the *charming* octave, all in relation to the lightest and highest pitched up quark. Bottom and top quark follow each an octave lower. Perceptual grouping between different quark generations is difficult, but such composite particles are rarer anyway.

The leptons are separated in higher registers, and have a *light*, e.g. a flageolet sound. The according neutrinos follow as clear sine tones an octave above the leptons. The pitches vary slightly for every observable around these frequencies. In Fig. 4, some examples are shown.

Every sound has a clear attack and decay, and for antimatter, the sound is just reversed.

Hadrons are composites of 2 or 3 quarks - the according pitches are played successively as a tonal pattern, always starting at the highest pitch. Also the tone lengths of the quark sounds vary with mass, resulting in a polyrhythmic structure.

Charge is given by a crescendo (for positive) and a decrescendo (for negative charge) on the whole structure (the tonal pattern for hadrons or single sounds for the other particle types). A neutral particle is steady in amplitude.

Each observable (a hadron or a lepton) is played by one musical instrument. This assures the perceptual grouping of the single quark sounds to one coherent particle and allows a certain characteristic by its timbre. Surely, more hadrons exist, than perceptually distinguishable instrumental timbres are available, but they rarely all appear in a measurement together. A violin sound can be used for the often occurring proton, as it is the dominant instrument of the orchestra. A viola sound is chosen been as the more ‘neutral’ instrument in comparison to the proton-violin, representing the neutron.

The experiment dynamics can be implemented as spatialization and/or the Doppler effect, using the *‘car-passing-by’* association mentioned above. With this basic scheme, also other particle displays are possible: e.g., the sonification of ‘static’ Feynman graphs.[2]

4. DISCUSSION

The *metaphor* procedure proved to be helpful for our purpose, and the resulting sonification design is a coherent and possibly intuitive ‘Acoustic Standard Model of particle physics’. Though a free, associative approach is rather demanding for the test persons, they surprised me with many interesting sonification ideas and with the sounds they were ready to make.

[1] Though, it should be mentioned that there were ambiguities with the term ‘excitation’, which is referred to excited particle states vs. the ground state, as this cannot be seen directly in experiments.

[2] Feynman graphs are a complete schematic representation of equations describing for instance particle decays.

Some outcomes may not be surprising to those who have been studying intuitive mappings before. As cited above from [10], high mass is normally linked to low pitch, which also makes perfect sense from a macroscopic experience point of view. Still, we found it interesting to ask physicists about microcosmic structures, where high mass equals high energy, and could in principle be mapped to pitch with a different polarity (high energy to high pitch). The analyses showed, that the high mass - low pitch metaphor is so strong, that it also holds for microcosm and is even mentioned as a first association in open questions.

There is a trade-off between *open* and *concise* questions. While the sonification expert should not lay too much of her/his own ideas into the questions, this might also lead to some misunderstandings. Misinterpretations occurred probably with the sound parameters, as they were explained in 'non-technical' terms. This could be – and should be – solved by *playing* actual sound examples to the participants.

Some conclusions can be drawn on the particle data set and the participants. 'Everyday' properties, like mass, are cited much more often than abstract ones, like quantum numbers. Imagination is limited when the participants are only used to mathematical treatment, or, the metaphorical shift from mathematics to a perceptual quality is too demanding for a simple questionnaire. Analysis showed also, that the concepts of particles become clearer, the longer people work in the field. This can be a benefit as strong metaphors emerge from professional experience, but also a drawback, as there is a lack of flexibility with new modalities, as sound.

The method in general helps with basic design decisions, but also restricts it. While sonification of complex data is already very demanding, another condition has to be taken into account. The *metaphor* method is indeed easily applicable for parameter mapping. However, for model-based sonification, the possibilities for metaphoric sound design are rather limited. Metaphors can still be implemented in the model design (rather than the sound design).

An open question not directly covered by the proposed method is the evaluation of the sonification. This has to be achieved by other methods.

5. CONCLUSION

We described the metaphoric sonification method as a procedure to explore metaphors in a scientific field and use them for a sonification design. We questioned and analyzed 19 physicists at CERN about their expectations, and created an auditory particle display based on the result.

Listening examples

The recordings of the free questionnaire part for all particles are available at http://qcd-audio.at/tpc/quest.

Acknowledgments

We would like to thank all test persons who participated. Our research has been funded by the Austrian Science Fund FWF within the Translational Research Project QCD-audio, and the doctoral program "Hadrons in vacuum, nuclei and stars" of the University of Graz.

6. REFERENCES

[1] G. Lakoff and M. Johnson, *Metaphors we live by.* The University of Chicago Press, 1980.

[2] B. N. Walker and G. Kramer, "Sonification design and metaphors: Comments on walker and kramer, icad 1996," in *ACM Transactions on Applied Perception*, vol. 2, no. 4, October 2005.

[3] B. N. Walker, "Magnitude estimation of conceptual data dimensions for use in sonification," *Journal of Experimental Psychology: Applied*, vol. 8, no. 4, pp. 211–221, 2002.

[4] ——, "Consistency of magnitude estimations with conceptual data dimensions used for sonification," *Applied Cognitive Psychology*, vol. 21, pp. 579–599, 2007.

[5] S. Barrass, "A comprehensive framework for Auditory Display: Comments on Barrass, ICAD 1994," *ACM Transactions on Applied Perception (TAP)*, vol. 2, pp. 403–406, October 2005.

[6] ——, "Auditory information design," Ph.D. dissertation, The Australian National University, 1997.

[7] C. Frauenberger, "Auditory Display design. An investigation of a design pattern approach." Ph.D. dissertation, Queen Mary University of London, 2009.

[8] A. D. Campo, "Science By Ear. an interdisciplinary approach to sonfying scientific dara." Ph.D. dissertation, University of Music and Dramatic Arts Graz, 2009.

[9] G. Martino and L. E. Marks, "Synesthesia: Strong and weak," *Current Directions in Psychological Science*, vol. 10, no. 2, pp. 61–65, 2001.

[10] B. N. Walker and G. Kramer, "Mappings and metaphors in Auditory Displays: An experimental assessment," in *ACM Transactions on Applied Perception*, vol. 2, no. 4, October 2005, pp. 407–412.

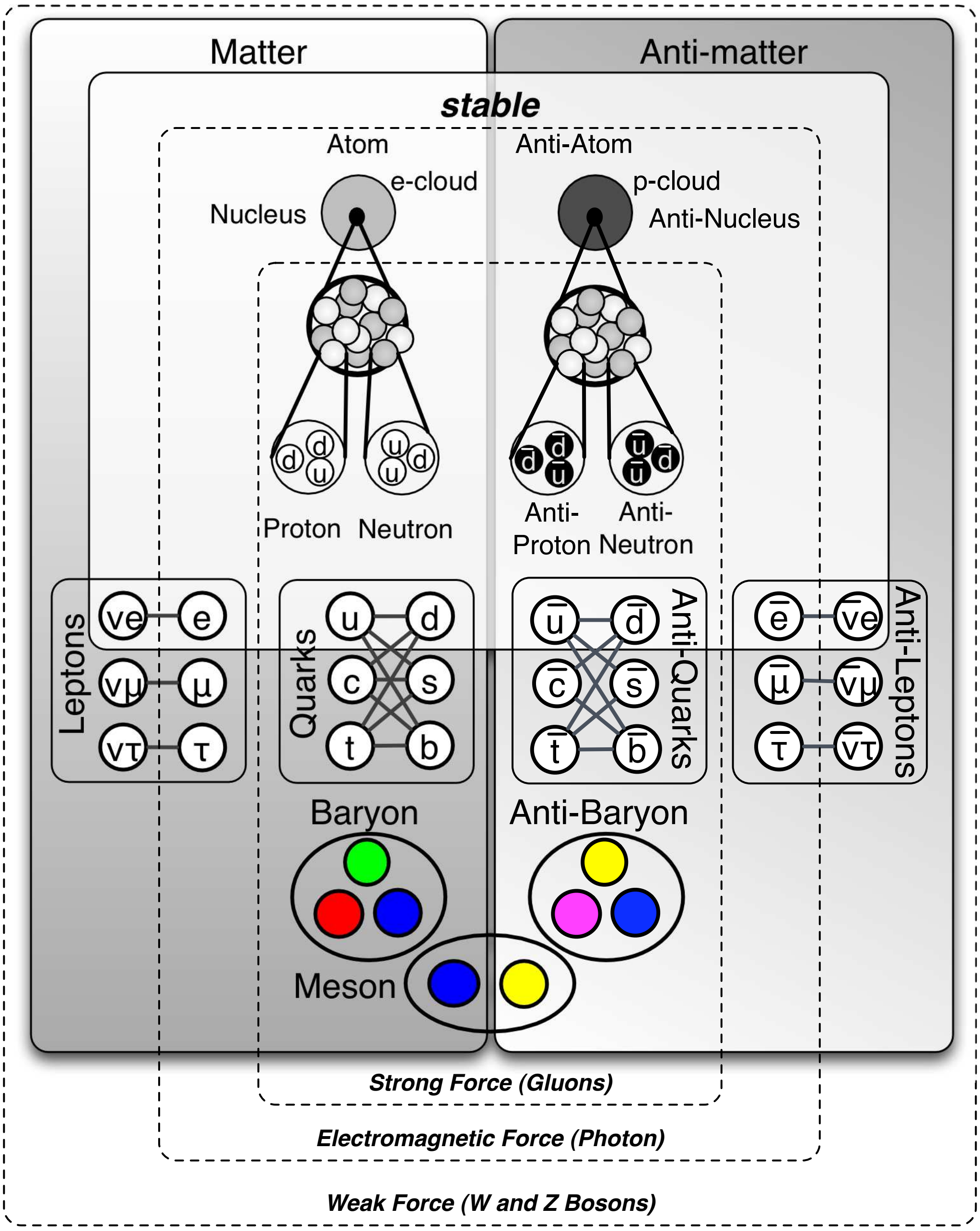


Figure 5: *Overview of the elementary particles in the Standard Model. The anti-particles are not shown but are completely analog to the matter-side. The following abbreviations are used: Leptons: ν_e - electron neutrino, ν_μ - muon neutrino, ν_τ - taon neutrino, e - electron, μ - muon, τ - taon; quarks: u - up, d - down, c - charm, s - strange, t - top, b - bottom. They are sorted in 3 generations with possible interactions indicated by lines between them.*

COMMUNICATIVE FUNCTIONS OF SOUNDS WHICH WE CALL ALARMS

Antti Pirhonen

Dept. of Comp. Science and Information Systems
University of Jyväskylä
P.O. Box 35, FI-40014, Finland
pianta@jyu.fi

Kai Tuuri

Dept. of Comp. Science and Information Systems
University of Jyväskylä
P.O. Box 35, FI-40014, Finland
krtuuri@jyu.fi

ABSTRACT

The design of alarm or warning sounds appears to be far from a trivial challenge. Even if the basic principles of creating an alarming quality for a sound have been widely accepted and applied, there seems to be a constant need for knowledge about what a ”good” alarm should sound like.

In this paper, we analyse the challenge of alarm sound design. The analysis is carried out in terms of an application context, which is an anaesthesia workstation in an operating room. We conclude that to result in satisfactory sounds, the design should not only concentrate on stereotypic qualities of expected alarms, like a strong psycho-physiological reaction but should also take more aspects into an account. It is proposed that these context dependent aspects, in turn, are extracted from the communicative functions of the sound's intended usage. For such a conceptual design of alarm sounds, a basic taxonomy of communicative functions in terms of alarm priority levels is proposed.

Even though this report concentrates on one application area, the approach would be applicable in several areas. Sound design for other safety critical applications, in particular, would benefit from our findings.

Keywords: warning sounds, anaesthesia, communicative functions

1. INTRODUCTION

What does a ”good” alarm sound like? When analysing alarm sounds or warning sounds, the design has actually started when the object of design is called an alarm or warning. The foremost communicative function of the sound to be designed has been embedded in that term.

In terms of communication, warning or alarming someone has a fairly self-evident function. When we warn, we wish someone to become aware of a danger or a risk. However, by performing a warning, e.g. by shouting or by hand gestures, we more or less inadvertently affect the person we are warning. For instance, we might startle or even frighten him or her. Or then, if the person to be warned feels that the warning was unnecessary, he or she might find the warning irritating or disturbing. Our warning may also be received by other people, for whom it is not at all relevant.

The description above can be applied to practically any context. In everyday life, we are used to false alarms or alarms which can easily be classified as irrelevant. In safety critical contexts, however, the requirements for warnings are much higher. While in some contexts, a strong and rapid reaction is all that counts, in safety critical environments the appropriateness of the reaction may be much more important than its strength of it [1]. Even though we talk about warning sounds, when hospital equipment is concerned, sometimes the primary function of these sounds is not necessarily to warn, but to inform [2]. However, this is not an either-or matter, since sounds can easily serve multiple functions. For example, vocal warnings vary in different situations, often telling us more than just that there is a danger, thus providing a basis for an appropriate reaction. We see that awareness of these varying communicative functions in regard to a given alarm condition, would enable the sound designer to formulate the design principles in a context-tailored manner.

Alarm sound studies traditionally concentrate on qualities like reaction time or perceived urgency. Indeed, these issues are important in the design of alarms. The studies in this domain have resulted in practical guidelines for alarm sound design (e.g. [3, 4]). The problem is that these guidelines mainly focus on communicating the urgency, thus acknowledging the alarming function of sound only. The second problem is that guidelines never cover all qualities of sound. They may provide details about intensity, frequency or rhythm, but many decisions about what the sound will actually sound like (qualitatively) are still left to the intuition of the individual designer.

In the current paper, we report a study in which we analysed the communicative functions of alarm sounds and some other non-speech sounds of an anaesthesia workstation in an operating room (OR) context. In the study, human expressions relating to each communicative function were used as a basis for the alarm sound design process.

2. DESIGN CASE: ANAESTHESIA WORKSTATION

The underlying purpose of this study was to find relevant and adequate information for the needs of designing a number of alarms in an anaesthesia workstation. Even though we focus on this particular case, the method and the underlying approach would be applicable in most applications of non-speech user-interface sounds, in particular in safety-critical contexts. For meet the brief of the current study, we needed to start by familiarising ourselves with the OR conditions, with particular regard to the soundscape. In an OR, there are numerous gadgets which all have their own repertoire of alarm sounds. A mayhem of different alarms is guaranteed when the room is equipped with technology made by different manufacturers, all of whom have their own product in mind alone when

This work is funded by Finnish Funding Agency for Technology and Innovation (TEKES), and the following partners: GE Healthcare Finland Ltd., Suunto Ltd., Sandvik Mining and Construction Ltd. and Bronto Skylift Ltd.

designing it. The hard acoustic properties of a typical OR, caused by the interior surfaces which have primarily been chosen for hygiene, do not make the situation any easier.

2.1. Method and procedure

Our research method was based on the so-called Rich Use Scenario (RUS) method, which has been created to understand the essence of the context of use for the needs of design [see 5, 6]. The application of RUS was modified for the current context.

The RUS method is based – as the name indicates – on use scenario. Typically, use scenarios are condensed descriptions of a use of an application [7]. They are used to reveal use-related issues, which would otherwise remain unnoticed. RUS differs from traditional use scenarios in that its focus is not on the observable details of use – like in overt behaviour – but on the experiences of the user, as a person. In RUS, the aim is to provide inspiration for designers. Therefore, RUSs are lively stories, which provide vivid imagery of the flow of using the application. The listener or, depending on the form of implementation, possibly, for example, the reader of the story, should be able to identify her or himself with the character(s) of the story. The technology to-be-designed is part of the environment which that person is living in and interacting with. The method has previously proved to be an effective way of immersing oneself in the context of use, from the user's point-of-view [6]. The story provides a common ground for a multidisciplinary team to reflect their design ideas. Programmers, graphic designers, interaction designers or other experts are thus able to justify their ideas within a common framework, which is understandable for the whole team: The team members ask themselves ”How would the character experience this or that idea”.

In the current case of an anaesthesia workstation, RUS took the form of a radio-play. The radio-play has proved to be an ideal form of implementing a use scenario for this kind of purposes, since:

- The whole design team is concentrating on each and the same point of the story at a time. In a written story, each member of a group would be at a different point of a story at a given point of time.
- The radio-play has been found to be an effective way of focusing the attention of group members [6], as well as evoking creative ideas (compared to video [8])
- As RUS provides a projection of the application use through the experiences and actions of the user, omitting visual elements in storytelling arguably facilitates the group participants' use of imagination and helps the ”enacting” of those experiences by themselves.
- As a form of presentation, audio is well suited for brainstorming sounds.

The alarm sounds, which were the actual object of interest, worked as sound effects in the radio play. The process, in brief, was as follows:

1. The manuscript for the radio-play was prepared in cooperation with the experts in the context (usability experts of the manufacturer).
2. The radio play was implemented. In the radio play, the sounds-to-be-designed appear as points of ”missing” sound effects, allowing them to be imagined.
3. Two design panel sessions were organised. The participants were six students of different subjects at the University of Jyväskylä. However, none of the participants had medical science as a major subject, i.e. the participants were amateurs in terms of the context. In the sessions, the participants planned and implemented appropriate sound effects at the given points of the radio-play (which were sounds from the anaesthesia workstation).
4. On the basis of the work of the non-expert design panels, draft sounds were implemented and embedded in the radio-play.
5. Two expert panel sessions, each made up of two anaesthesia nurses and one doctor, were conducted. In the sessions, the final radio-play, including the sound effects, was listened to and discussed.
6. A post-questionnaire was sent to all expert panel participants.
7. The discussions of the expert panels were transcribed and analysed.

The RUS manuscript and its radio-play implementation worked as input for the preparation of draft sounds in a non-expert design panel. The radio-play, with draft sounds included, provided a basis for the discussions of the expert panels. Draft sounds especially worked as effective ”triggers” of conversation.

The decision to ask non-expert panellists to produce the draft sounds was found successful. In the previous versions of RUS, the production of draft sounds was found to be problematic [6]. We conclude that our previous parallel between draft visual layouts and draft UI-sounds was not appropriate. In visual mock-ups, draft quality has been found to encourage the users to make suggestions. Possibly, the power of draft quality (especially hand drawn) is not in its coarseness *per se*, but in the ”human touch” – the panellist/designer can easily attune to the outcome and imagine having produced the draft by herself. The draft sounds we have previously used were produced with a computer, and they always contained some qualities so irritating that they did not provide a basis for constructive elaboration. However, while this time used human voice and real instruments, the sound idea was much better communicated.

The outcome of non-expert panels clearly illustrated the sound ideas of the panellists. In the case of alarm sounds, we took recurring ideas from various different draft sounds and arranged these features into one coherent sound set. All new draft sounds were re-articulated with the same instrument (metallophone), except sound #4, still trying to preserve the characteristics of the original draft sounds produced by the non-expert panellists.

2.2. Analysis

This report focuses on the analysis of the discussions of the two expert panels (phase 5 above). It has to be noted that the discussions were originally in Finnish, but we have tried to express the original nuances when translating the quotations in this report into English. When the discussions contain oral expressions of non-speech sounds, we use phonetic notation.

The current analysis concerns four different sounds in different alarm conditions, which represented different levels of warnings in the anaesthesia workstation. The radio-play used in panels dealt also with some non-alarm, non-speech sounds for the anaesthesia workstation, but they are not included in this analysis.

2.2.1. Sound #1: Medium priority alarm

The draft sound was produced with a metallophone. It consists of series of two damped hits at approximately 1 sec. intervals (D# tone, medium register).

The events causing the alarm condition in the scenario were:

- Blood pressure has exceeded the alarm level (patient based alarm)
- The entropy meter is badly connected (device based alarm).

General observations concerning design principles:

- In expert panel 1, the events of the scenario were found different in priority:

> ...but I think that if blood pressure has really been too high, it is quite different and requires different reaction than a badly connected entropy sensor – if it has not been pushed in tightly enough thus losing contact.

- Expert panel 2 wished medium-level alarms to be merely informing rather than alarming:

> ...It has to be noted, that 'aha', but not anything more severe, let's sign for it in a few minutes. But if you are busy with other, important tasks and that is tapping away all the time in the background, it would rile.

> ...it obviously depends on the scale – what is classified as important.

Opinions about the draft sound:

> ...Perhaps a bit too feisty...kind of loose...

> ...should not be that dense...

> ...I don't like that metallic tone, it's irritating...

> ...[should be] somehow softer...

> ...Were there two taps? Perhaps rather...well it depends on the qualities of the sound but perhaps one of that kind would be good.

> ...since there were two of them [taps], it made it kind of commanding, like 'hey, ...!!'

Features of the sound:

- Medium-level alarm should not be too loud, obtrusive nor frequent.

> ...perhaps high priority alarm should be something like this (tapping continuously) to grab attention, but these kind of sounds in which no immediate reaction is necessary, perhaps simple [ˈbø:b] would be adequate.

- On the other hand, it should be snappy and adequately startling.
- Soft, non-metallic timbre would be desirable.
- Single-tone structure (instead of two tones) and longer pause between repetitions (at 10-15 sec intervals) was proposed.

2.2.2. Sound #2: Low priority alarm

This draft sound was also played with a metallophone. It consists of single damped hits at approximately 2 sec. intervals (F tone, low register). The alarm condition in the scenario was due to the sensor for muscular activity becoming loose (device based alarm).

General observations concerning design principles:

- Low priority alarms should not be alarming at all:

> ...Well, you know, that when you hear those day after day...if a...I wouldn't say unnecessary but a less urgent issue causes extremely...is very strong, it drowns everything else.

- The conversations indicated that frequent alarms and the kind of alarms which are perceived as ”cry wolf”, should preferably be totally removed.
- To be meaningful, a low priority alarm should only inform, without demanding too much attention and rapid reaction. It can be repeated, not too frequently, but as a reminder.

Opinions about the draft sound:

> ...Not too bad. Quite sharp, though. But as a form of sound, not bad.

> ...Could be a bit softer...

> ...That the device has come loose, that is quite...working, not bad.

> ...Quite suitable pause between the sounds, so at least I didn't find it...

> ...At least there is no need for more frequent repetition, because it is not a question of something fatal...but you just pay attention, 'aha'

> ...I would draw the scale [of alarming] downwards.

> ...This sound, caused by such a minor issue frightens the patient needlessly...if I was there and heard such a sound, I would be astonished, asking if there is something badly wrong with me. And that is unnecessary from all points-of-view.

Features of the sound:

- Quite similar to the medium priority alarms, but low priority alarms especially should be soft reminders.
- Short, sharp sounds should be avoided. Since it is a question of an infrequently repeating sound, its duration may be longer with a soft onset.
- 5-10 sec was found to be a suitable interval between the repeating sounds, when different intervals were compared. (It has to be noted that the comparison was focused on the sound, and therefore very different from the real situation.) With a slower sound onset time, it was proposed that the interval could probably be even longer.

In other words, the underlying *action model* for sound design could be, for example, ”peaceful breathing” rather than ”hitting”.

2.2.3. *Sounds #3 and #4: High priority alarm*

For this case, two draft sounds were produced. The first one was intended as a patient based alarm. It was played with a metallophone, consisting of rapid bursts of two hits repeating at short intervals (G# tone, high register). The second draft sound was proposed as a device based alarm. It was implemented by stomping with feet in rapid three-hit bursts repeating frequently.

The events causing the alarm condition in the scenario were:

- False alarm / noise in EEG, caused by diathermia (patient based alarm)
- Battery running out (device based alarm)

General observations concerning design principles:

- The panellists found alarming qualities important in high priority alarms, when it is a question of a critical situation which requires immediate reaction. In such a case, the sound can and should be ”irritating”.
- There may be differences among well functioning alarm sounds in terms of their pleasantness. In the panel, the draft sound was found to be alarming but also more pleasant than the sounds of the existing product.
- The draft alarm sound for a low battery level was implemented differently from the previous one, thus illustrating the difference between patient and device based alarms.
- Some of the panellists also ideated a continuous, non-critical warning sound which could be on whenever the device is running on batteries. It illustrated the possibility of providing information about the condition *gently*, before anything wrong happens. Low pitch and the soft clicking tone can be used to avoid excessive obtrusiveness, which would shift attention from critical issues.

> ...Our nitrogen servo always comes to mind ...namely that when you inspect it, you will need to unplug it from the mains. So when it is running on batteries, it continuously makes a sort of low, ‘clicking’sound [ˈnɑkˈnɑkˈnɑk] – so you are bound to notice that it is running on batteries, without mains.

Opinions about the draft sounds:

- Patient based alarm (#3):

> ...If there really is a tachycardia or asystole, this sound would turn the head, ‘what’s going on’. But for a fault caused by noise, that is...no thanks!

> ...I wouldn’t call it bad at all. When urgent warning qualities are needed, this works...

> ...I found this quite good.

> ...I think that this one had the most pleasant tone of these, I don’t know what makes is pleasant though.

- Device based alarm (#4):

> ...I found that good. It was so different from the other sounds. Quite...ok.

> ...There are so many tonal sounds in use. Once when we were urgently trying to find which meter was screaming or whose device was beeping..., it was revealed that the sound source was actually a refrigerator, indicating that it is freezing something...So that kind of distinguishable sound is really relevant.

> ...I found it good that there was a short interval.

> ...It came through like [ˈkɒpsˈkɒps] screaming that ‘there is a failure, there is a failure’

> ...And it was different, so that you will react differently, go to the device to see what’s wrong.

Features of the sound:

- Expert panellists approved both sounds as very good alarm signals for a critical situation that needs immediate care.

> ...I find the distance between two pairs [bursts] most important. It has to be short...

> ...those two hits...they should not be too far away either, to make the sound almost continuous

> ...This was good in both ways, both the distance between the pairs and between the single sounds of a pair.

> ...I think that there wouldn’t be anything wrong with either of these high priority alarms. They could be adopted just as they are now.

- Even though the sound for patient based alarm was produced with the same instrument as the previous alarms (whose timbre was not found pleasing), metallophone was now found appropriate and even pleasant.

> ...In my ears, that did not sound like the same sound being played more frequently. Rather, there was something more pleasant and it was getting attention quite well still.

> ...In my own mind, based on previous experiences, I associated this with an alarm of a train which is soon departing, meaning that you have to hurry now.

2.3. Summary of findings

2.3.1. *Features of draft sounds relating to the alarm priority categories*

In general, in the sound ideas of the non-expert panellists, the intended alarming quality was produced with a repeating series of two or three beats. These ideas came up spontaneously, from several panellists, even though there were no external cues (e.g. in the instructions) for this kind of structure. Another common feature among the proposed series of beats was that they did not constitute any melodic structure, but were repeated at a constant pitch or were percussive. It could thus be interpreted that a melodic alarm or warning (suggested in IEC/ISO 60601-1-8, annex F) would not be very intuitive. This conclusion is also supported by the observation that the same panellists proposed melodic features for non-alarming sounds, as well as previous studies [9].

The absolute pitch of the proposed alarm sounds varied, as well as the timbre and sound source (vocal sound, metallophone, xylophone, foot stomping), while the basic idea of an alarming quality remained the same. As explained above, for the needs of the expert panels, the ideas of the non-expert panellists were summed and reproduced in a newly articulated set of draft sounds, which was in accordance with the alarm priority scale.

In high priority alarms, the series of two or three beats/notes appeared in dense bursts (onset distance approx. 80-180 ms). In addition, the bursts were densely repeated (interburst interval approx. 300-700 ms). Within an articulation, these gaps between beats and bursts remained consistently the same. In medium level alarms, the beats or notes (one or two) were articulated more calmly (onset distance approx 400 ms) and with less intensity. In addition, the frequency of series was lower (interburst interval approx. 1,2 s) and generally produced with lower pitch than high priority alarm sounds. Low priority alarms were characterised by low frequency and a softly articulated structure of one beat or note (interburst interval approx. 2 s).

It has to be noted that the way in which the alarm priority levels appear in the ideas of the non-expert panellists' draft sounds, corresponds amazingly well with the IEC/ISO 60601-1-8 standard (see especially tables 3 and 4 on page 35), which is mainly based on rhythm. The clearest difference between the draft sounds and the standard was in the distance between bursts, which are defined as much longer in the standard than the panellists proposed. An exception to this is the high priority alarm given in the standard, in which one burst consists of four rapid sub-bursts. So the alarm defined in the standard consists of very rapidly repeating sub-bursts, but only in periods of two.

In terms of the rhythm, it can be summed up by saying that the difference between the draft sounds of the panels and the ISO standard was that the draft sounds were more alarming by nature. One likely reason for this was the non-expert panellists' lack of experience of the context of use. Another possible reason is that the straightforward naming of sounds as "alarms" – even in the case of low priority – might have stressed the need for alarming communicative function. However, generally speaking, the professionals appeared to prefer the spontaneously produced draft sounds of the panellists to the existing sounds of the workstation, even though the latter ones follow the standard.

2.3.2. *Expert panels' assessments of the alarm priority levels and draft alarm sounds*

As a rule, the panellists found that there are too many alarms in an OR or they are seen as irrelevant. On the other hand, it was admitted that most of the alarms are necessary. The key issue is the way in which different alarm conditions are classified into the three alarm categories. It was claimed that only in conditions which are really urgent should be alarming by nature. It was wished that medium and low level alarms would be merely informing or reminding. Since the conditions in those cases do not require immediate reaction, too alarming a sound would disturb the ongoing work.

Our interpretation is that the biggest challenge in alarm sound design is not the design of high priority sounds, in which even extremely alarming sounds are found as appropriate. On the contrary, the challenge is in the design of low and medium level alarms, which are usually experienced as too frequent and too strong in terms of the alarm conditions.

> ...Those are always the ones we try to switch off, because there is the risk that you become deadened to the constant alarm, and don't pay attention to a real one. So we try to adjust the alarm threshold values and other things so that – when there is a real change...If the patient's blood pressure is high, say, 200, we raise the alarm threshold value so that only if the pressure gets even higher, there is an alarm. But not constant alarming, as told, because then you might not react.

> ...All the kind where the patient's life is under immediate threat, then of course, not when the importance and message is that we anyway act in five seconds, so it doesn't have time to irritate. The issue is fixed or switched off...it overrides everything else...and it won't have time to irritate.

> ...In quite many cases it is a question of a situation like 'oh no, that started to alarm and interrupted what we were just doing...'

> ...it would be good to have an alarm if it is detached from the hoses , of course, but anyway, you won't die in a second if it is detached like this; but those real alarms should engage only when they are definitely needed.

> ...[medium and low priority] issue and message is heard and noted and reacted to as soon as it is possible. However, when that 'as soon as possible'moment is there, the alarm should not hinder action by irritating and by preventing concentration.

So it can be argued that the communicative function of low and medium level alarms should be informing and reminding, rather than alarming. They differ, however, from information signals (as referred in ISO standard) in that they anyway relate to alarm conditions and should be interpreted in that context. Low and medium level alarm signals were seen in quite a similar way in similar functions: soft informers and, when needed, reminders of an alarm condition. In a post-questionnaire, all 6 panellists agreed that three alarm priority levels are appropriate, but most of them (4) added that existing alarm sounds have more alarming qualities than necessary. According to the panellists, this is mainly due to inappropriate illustration of the intended priority level in the sounds, rather than inappropriate prioritisation of alarms. Both of these – the intended prioritisation and the one perceived – should be critically considered.

Even though it could be argued that the draft sounds which we used in the expert panels are more alarming than what the ISO standard suggests, they worked quite well as a part of a contextual, radio-play format scenario. High priority alarms were found appropriate and even pleasant. Their division into patient and device based alarms received positive feedback. In low and medium level alarms there was a wish for a different character. Apparently, using the metallophone in all sounds was not a very good idea, since the sharp and metallic tone was found appropriate in high priority alarms only. In contrast, in low and medium level alarms, panellists wished the "scale of alarmness" to be downgraded. There was a desire for softness in tone and in onsets and offsets of sounds.

Medium level alarms were also found too ”dominant” and the sequence of bursts too frequent. In the low level alarm the sound was also found to be repeating too frequently. When different intervals between the repetitions were tried out, 5-10 seconds seemed appropriate, but 10-20 seconds was found too long (according to the ISO standard, in a low priority alarm the interval should be 15 seconds). It was mentioned, however, that if the sound were longer with soft rather than sharp onset and offset slopes, the interval could be longer. The panellists never argued that even the low priority alarms were completely unnecessary. However, expert panel 1 wished the regular measurements of blood pressure and the sounds of the pulseoximeter to be removed:

> . . . it does not need to tell us every five minutes that now blood pressure has been measured. . .

As mentioned, alarms were considered to be necessary, but negative attitudes towards the existing alarms were expressed many times. What is so irritating in the existing sounds? One possible feature is that they are felt to be artificial and machine made:

> . . . I have to say that the current sounds, they are clearly kind of mechanical, mechanically created sounds, so they are not. . . and these [draft sounds of the scenario] have been produced somehow with natural instruments.

> . . . Or stomping with feet or something. . . these sound somehow more pleasant than those kind of. . .

> . . . [the current alarm sound] sounds kind of stuffy imitation, while that [draft sound] is clean and clear. . .

In the post-questionnaire, most of the panellists (4 out of 6) wanted primarily to change the timbre of the existing alarm sounds. The other two panellists, respectively wished to change the characteristics or repetitions of the alarm sound. Obviously, recurrent exposure to alarms and the above-mentioned discrepancy between alarm priorities and alarm signals also has an effect on irritation.

In the current alarm sounds, the alarming quality has been designed by defining the technical parameters of sounds. ISO standard's guidelines encourage mechanical production of sounds in which certain parameters fall in the recommended range. However, the standard does not handle a sound *per se* as a meaningful and intentional object. We thus propose that the level of alarmness should be seen as a communicative function, mediated by the alarm signal, rather than just technically scaled features of sound. Communication is mediated in the amateur panellists' and sound designers' draft sounds in a natural way – through the spontaneous articulation of intention. More attention should also be paid to the acoustic characteristics of sounds and various connotations and affective reactions evoked by them. Perhaps the metallophone, which was used in the articulation of draft sounds, has not enough expressive power since it is not possible to make many adjustments to its timbre, duration or the internal dynamics of the sound.

2.3.3. Division into patient and device based alarms

As previously mentioned in this report, the expert panels participants found the division into patient and device based alarms a good idea. In panel discussions, the panellists encouraged the making of this division in forthcoming alarm sounds. It was found important that the sound itself (e.g., its timbre) should indicate the source of alarm. This appeared to also a priority or emergency issue. The post questionnaire showed that the panellists were not concerned about the growth in the number of different kinds of alarm sounds resulting from this division, but they believed that the division would make the interpretation of different sounds easier.

A device based alarm was seen as lower in terms of priority than a patient based one. Since both of the alarm types need to be interpreted with the same priority level scale (low, medium, high), the alarms within each category need to resemble each other to some extent. For instance, the rhythmic structure could define the priority level category, while tone could be used to make a difference between patient and device based alarms; device based alarms tone could be less alarming (soft, damped) and patient based, in turn, clearer and sharper. This difference could be seen in the high priority draft sounds, in which the device based alarm was produced by stomping feet and the patient based with a metallophone.

3. CONCLUSIONS AND DISCUSSION

Since the categorisation of alarm conditions and related alarm sounds into three priority levels has been found appropriate, the focus of sound design should be on the design of appropriate sounds for each level. The clinicians who participated in this study clearly expressed that they need sounds which would correspond better with their referents than the existing sounds of their current devices. A general principle in sound design could be that none of the sounds should feel unnecessary in its context (i.e., could be removed) nor unnecessarily strong or uncomfortable (i.e., should be removed).

According to the current standard of alarm sounds, the primary function of alarm sounds is to make the operator shift attention to the cause of the alarm (IEC 6060-1-8, p. 77). The communicative functions of alarm sounds thus mainly relate to appropriate focus of attention and indication of urgency (priority levels). We find it important to notice that the communicative functions of these sounds are not only for alarming. Even if the sounds which we dealt with in the case study have been classified as alarm sounds, the analysis of the discussions of the practitioners made us suspicious about the appropriateness of the term ”alarm”, at least in an OR context. When going through the use scenario and the related sounds one by one, we found that ”alarming” or ”commanding” was the foremost function of the sound only for high priority alarm conditions. Still, our clinician panellists acknowledged that sounds for lower priority conditions are needed mostly because of the information they provide. This suggests that their primary communicative function is informing-related (see also [2] and Table 1).

Only in truly urgent events is it justified to use alarming features in sound. In the design of lower priority level sounds, the mere mechanical reduction of alarming features of a high priority sound is not an adequate guideline. In addition to alarming and urging, other functions and means to express them should be considered in order to prevent inappropriate reaction or inconsistent associations in the given context (taking all modes of listening into account [10]). In other words, the designer needs to consider subtle ways of getting attention when it is a question of low or medium level alarms. What is subtle and what is not should be assessed in terms of the operator's experiences.

It was interesting that the draft sounds, ideated by amateurs for given situations, corresponded amazingly well with the alarm sound standard. Even if they were even more alarming than the existing sounds (which strictly follow the standard), they were found more pleasant. According to our observations, the key issues in this phenomenon were ”human touch” and communicative intention conveyed in the articulation of the sound. In music performance, articulation (i.e., how notes are expressed as sounds) is a central factor in the mediation of emotional and intentional states of mind. Likewise, we would recommend striving towards communicative expression and concentrating on the articulative nature of sounds when designing sounds for hospital technology.

The finding above fits in well with the notion of embodied cognition [11]. Since human experience is ultimately based on bodily reflections, it is more natural to interact with sounds which can effortlessly be related to corporeal events than with sounds which feel artificial [12, 13].

The design case presented in this paper provides general principles and detailed guidelines for the needs of sound design in this particular case and context. The principles, in particular, would be beneficial in other contexts as well. In terms of priority levels, the central observations could be summed up as follows:

- High priority alarms:
  - Sounds can clearly be even more alarming – in terms of perceived urgency – than the sounds which follow the standard.
  - Even if it can be assumed that there is an immediate reaction to the alarm, continuous alarm signals are not recommended. For instance, in the structure of dense bursts, there should be a pauses (e.g. as follows: 5 dense bursts – 3 seconds’ pause – and so on).
  - Melodic structures do not seem to be perceived as alarming; possibly quite the contrary.
  - Percussive sounds appear to be favoured, at least by our panels, i.e., the underlying mental model or action model is to warn by beating, stamping, knocking etc.
- Medium priority alarms:
  - The sound needs to get attention, but excessive commanding or sharp quality should be avoided.
  - One single soft sound object, repeating at about 10 second intervals, could be adequate. The alarm sound ISO standard provides an appropriate guideline for the frequency of repetitions.
  - Sound objects could be constructed in terms of the burst definition in the ISO standard. The strictly defined structure of burst should be broken up, though. Sound objects do not need to be mechanical, beeping pulses either.
  - Attention should be paid to the timbre and internal dynamics of sounds. Crucial factors in the perceived softness are onset (attack) and offset (release) phases of a sound object and the *legato* between separate objects.
  - Percussive sounds did not work in the panels. The related action model should be softer than beating, e.g. arched swing, circular movement or waves.
  - Even though it is a question of an alarm, the dominating communicative function should not be commanding or alarming, but something more subtle.
- Low priority alarms:
  - The sound should be noted, but all commanding or sharp qualities should be avoided.
  - Close resemblance to a medium-priority alarm, but more subtle, soft, ”round”, simpler in structure and more peaceful.
  - Alarming qualities – in the traditional manner – should not be included at all.
  - One single, very subtle sound object, repeating at 15-30 second intervals. The standard is a good basis for defining the interval.
  - A burst consisting of two melodic sounds, suggested by the standard, appears too obtrusive for low priority alarms.
  - Communicative intention should be to guide attention or inform with subtle, pleasant means. A commanding or alarming quality is not at all appropriate.

Table 1: Classification of certain communicative functions in terms of priority. Proposed primary functions are highlighted. Because of the sensitivity of low priority alarms, expressive functions are proposed for primary role.

	Directive functions (e.g. alarming, prompting)	**Assertive functions** (e.g. informing)	**Expressive functions** (e.g., expressing arousal or calmness)
High priority	**Urging to act right now, demanding attention**	Asserting immediate threat. Optionally informing about the cause (e.g. patient or device)	Expressing high levels of arousal
Medium priority	Asking for attention, suggesting an operator action	**Informing or giving feedback about alarm condition** (state of the patient or the device)	Expressing calmly, but sensitively implying slight arousal
Low priority	Guiding attention, implying a potential need for action	Informing or giving feedback about alarm condition (state of the patient or the device)	**Expressing gently**

As we suggest conceiving alarm sounds as communication, let us examine the communicative functions of alarm sounds within the context of speech act theory [14]. According to the theory, a speech act is directive when its intention is to get the hearer to undertake an action. Alarming and attention getting sounds thus primarily serve *directive* functions. Similarly, an *assertive* speech act, which presents some state of affairs in the world, corresponds with the communicative functions of informing. An *expressive* speech act is also very relevant to alarm sounds, as it refers to expressing the affective state being involved in the act of communication. These mentioned points of speech acts (directive, assertive and expressive) bear the closest relevance to the functions of alarm sounds, and we propose that they can be used as general top-level categories in conceptualising the spectrum of different communicative functions and their relative ”weights” in each case of alarm sound design. Table 1 demonstrates how these categories can be applied in formulating design principles for alarm sounds of high, medium and low alarm condition priorities.

We argue that the naming of sounds directs the orientation of the sound designers. Judging by the experiences of the practitioners, too much alarming quality is usually included in medium and low priority alarm sounds. We argue that the key issue on the way to more acceptable sounds would be to analyse the communicative functions of each sound to be designed at the very beginning of the design process. It might be a good idea to call each sound according to its primary communicative function – calling a sound an alarm would at least sub-consciously perhaps orientate a designer to look for ways of communicating through sound that are too obtrusive.

4. REFERENCES

[1] J. Edworthy & E. Hellier, ”Fewer but better auditory alarms will improve patient safety”, Quality and Safety in Health Care vol.14, no.3, 212–215, 2005.

[2] F. J. Seagull, Y. Xiao, C. F. Mackenzie, & C. D. Wickens, ”Auditory alarms: from alerting to informing”, In Proceedings of International Ergonomics Association 2000/Human Factors and Ergonomics Society 2000 Congress, Vol. 1, pp. 223–226, 2000.

[3] R. D. Patterson, ”Guidelines for auditory warning systems on civil aircraft”, CAA Paper No. 82017, Civil Aviation Authority, London. 1982.

[4] J. Edworthy, S. Loxley & I. Dennis, ”Improving auditory warning design: relationship between warning sound parameters and perceived urgency,” Human factors, vol. 33, no. 2, p. 205, 1991.

[5] A. Pirhonen, K. Tuuri, M. Mustonen, & E. Murphy, ”Beyond clicks and beeps: In pursuit of an effective sound design methodology”, In: Oakley, I. & Brewster, S. (Eds.) Haptic and Audio Interaction Design. Proceedings of Second International Workshop, HAID 2007. Seoul, Korea, November 29-30, 2007. Lecture Notes in Computer Science 4813, Berlin/Heidelberg: Springer Verlag, pp. 133–144, 2007.

[6] E. Murphy, A. Pirhonen, G. McAllister, & W. Yu, ”A semiotic approach to the design of non-speech sounds”, In: McGookin, D. & Brewster, S. (Eds.) Proceedings of First International Workshop, HAID 2006, LNCS 4129, Berlin/Heidelberg: Springer Verlag, pp. 121–132, 2006.

[7] M. J. Carroll, Making use: Scenario based design of human-computer interactions. Cambridge, MA: MIT Press, 2000.

[8] P. Greenfield, D. Farrar, & J. Beagless-Roos, ”Is the medium the message? An experimental comparison of the effects of radio and television on imagination”, Journal of applied developmental psychology, vol. 7, no. 3, 1986, 201–218.

[9] P. Sanderson, A. Wee, E. Seah, & P. Lacherez, ”Auditory alarms, medical standards, and urgency”, In: Stockman, T., Nickerson, L. & Frauenberger, C. (Eds.) Proceedings of International Conference on Auditory Display ICAD 2006, Queen Mary University of London, June 20-23, 2006, CD-ROM format.

[10] K. Tuuri, M. Mustonen & A. Pirhonen, ”Same sound – different meanings: A Novel Scheme for Modes of Listening,” In Proceedings of Audio Mostly 2007 Ilmenau, Germany: Fraunhofer Institute for Digital Media Technology IDMT, 13–18, 2007.

[11] M. Leman, Embodied Music Cognition and Mediation Technology, Cambridge, MA: MIT Press, 2008.

[12] R. I. Godøy, ”Gestural-sonorous objects: embodied extensions of Schaeffer’s conceptual apparatus,” Organised Sound, vol.11, no. 2, 149–157, 2006.

[13] K. Tuuri, ”Gestural attributions as semantics in user interface sound design,” In: Kopp, S., Wachsmuth, I. (Eds.), Gesture in Embodied Communication and Human-Computer Interaction, No. 5934 in LNAI, Springer-Verlag, 257–268, 2010.

[14] J. Searle, ”Expression and meaning: Studies in the theory of speech acts,” Cambridge, UK: Cambridge University Press, 1979.

THE SONIFICATION METAPHOR IN INSTRUMENTAL MUSIC AND SONIFICATION'S ROMANTIC IMPLICATIONS

Volker Straebel

Technische Universität Berlin
Fachgebiet Audiokommunikation
Sekr. EN8, Einsteinufer 17c, 10758 Berlin, Germany
volker.straebel@tu-berlin.de

ABSTRACT

The sonification metaphor is not limited to electronic sound synthesis and computer music, but can be applied to instrumental music as well. The relation of sonification to program and experimental music is discussed and works by Iannis Xenakis, Karlheinz Stockhausen, John Cage and Alvin Lucier are briefly introduced. The paper leads to a discussion of the connection between sonification and romanticism, where the desire is to directly evoke an understanding of natural phenomena.

1. INTRODUCTION

When, after initial explorations in the 1870s [1], the concept of data sonification was established in the 1980s and further developed in the following decade [2], [3], two assumptions came to be taken for granted: First, sonification is considered a human/computer interface and hence the means of sound production are electro-acoustic, and second, sonification reveals some information about the matter represented by the sonified data. Weinberg and Thatcher even describe the latter aspect of data exploration as "immersive" and claim "a direct and intimate connection to the information" [4][1]. From a musicologist's point of view, the concept of data sonification appeals for two reasons. First, sonification as an idea released from its ties to computer applications can act as a metaphor for non-electronic compositions that are strictly representational in nature. I am referring here not so much to instances of program music that communicate a narrative, but to works in the tradition of experimental music that map extra-musical data to musical parameters. Second, the basic assumption of sonification researchers, that their technique provides a means to gain an immediate understanding of the matter represented in sound, seems to be derived from certain concepts suggested by Early Romanticism. The idea of a *Natursprache*, a poetic language in which one could directly experience nature as the embodiment of a divine being, is prominently expressed in the writings of Novalis [5][2], [6]. While there is no proof that those involved in sonification research read the late 18th century German philosophers, the proximity of their ideas might suggest a connection through common cultural knowledge.

2. PROGRAM MUSIC AND THE SONIFICATION METAPHOR

To claim that absolute music, that is, instrumental music without reference to extra-musical entities, is the paradigm of music per se, is an assertion of early 19th century music aesthetics. Vocal music obviously refers to the themes expressed in the lyrics, and instrumental music had always served social or ritual purposes. By means of tone painting, instrumental music can imitate the sounds around us, like birds, water or thunderstorms. Beethoven's statement that his Sixth Symphony, the *Pastoral Symphony*, was "mehr Ausdruck der Empfindung als Malerey" – "more the expression of feeling than painting" – marks the beginning of a conception of program music where the music does not merely convey a literary narrative through musical imitation of characteristic acoustic objects (think of Smetana's *Moldau*, for instance), but instead creates an imaginary drama or represents a poetic idea. That the extra-musical program does not need to be known to the listener is shown by Tschaikowsky's Sixth Symphony, *Pathetiqué*, where the composer preferred to keep the program to himself. This establishes an interesting double bind, since the listener knows s/he is not meant to take the work for absolute music, yet the program remains a secret. The listener is supposed to experience a meaning beyond the music, just as someone listening to sonification signals is supposed to interpret information communicated through sound.

[1] p. 9

[2] p. 147, n30

Starting in the 1950s, composers began to refer to extra-musical entities, particularly scientific data, on a different level. They no longer *imitated* sounds or expressed certain feelings or poetic ideas, but incorporated algorithmic or conceptual procedures in their compositional process. The sonification metaphor was used, before its concept was established, to map the shapes of stones or the panorama of the Alps to melodic lines or sound spectrums. This way, the representational aspect was internalized into the composition itself.

2.1. Iannis Xenakis: *Pithoprakta*

In 1955/56, Iannis Xenakis used stochastic calculations in the composition of his orchestra piece *Pithoprakta*. The speeds of the glissandi of 46 separately scored string instruments were determined by a formula describing the Brownian motion of gas particles. For a section of 18.5 sec. duration (measures 52-60), Xenakis calculated 1148 speeds, which he distributed into 58 values according to Gauss's law. A graph illustrates the movements of the pitches (fig. 1).

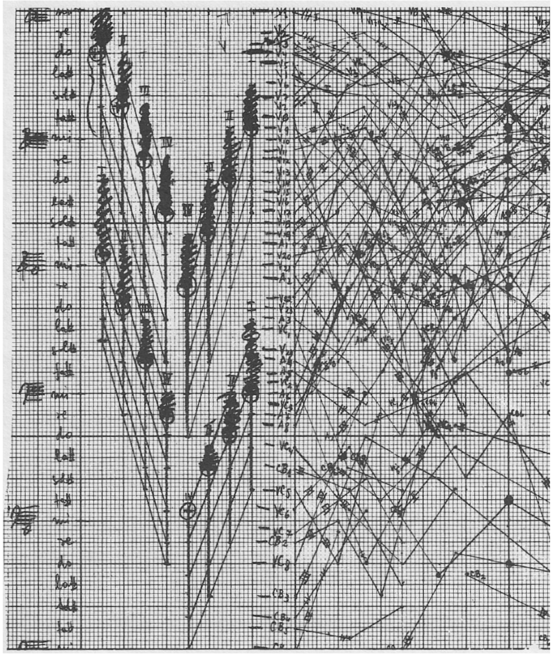


Figure 1: Iannis Xenakis, sketch for *Pithoprakta* [7][1]. Copyright © 1967 by Boosey & Hawkes Music Publishers Limited. Used by kind permission.

In each part, the durations of the glissandi remain constant, occasionally skipping one beat. The durations are 3, 4 or 5 beats per measure with 26MM (beats per minute) a measure. The three tempi superimposed create a rather complex resulting rhythm [7][2].

Since the durations of the glissandi of each instrument remain constant, the change of speed needs to be expressed by the change of interval that gets spanned within the constant duration (fig. 2). The sonified speed of the gas particles is musically represented by the glissando's differential. (The use of pizzicato, however, makes it rather difficult to actually hear the glissandi, since the attack emphasizes the pitches from which the sliding tones start).

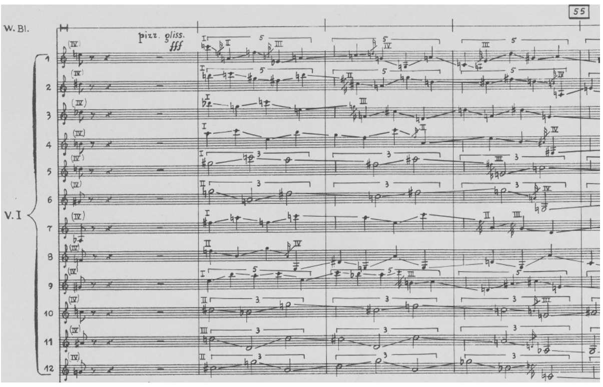


Figure 2: Iannis Xenakis, *Pithoprakta.* Violins I, measures 51-54 [8]. Copyright © 1967 by Boosey & Hawkes Music Publishers Limited. Used by kind permission.

In *Pithoprakta*, Xenakis did not actually sonify measured or otherwise observed data, but merely illustrated the mathematical description of the physical phenomenon. He called his approach "one of those 'logical poems' which the human intelligence creates in order to trap the superficial incoherencies of physical phenomena, and which can serve, on the rebound, as a point of departure for building abstract entities, and then incarnations of these entities in sound or light" [7][3]. Here, the composer limits the role of sonification to a point of departure for his inspiration. He is not so much interested in the concept of translating scientific data into musical parameters, but rather in emphasizing the connection between the arts of music and mathematics, as established by Ancient Greek philosophers and the theorists of the Middle Ages.

2.2. Karlheinz Stockhausen: *Gruppen*

In a similarly abstract way, Karlheinz Stockhausen made reference to the mountains around Paspels, Switzerland in his *Gruppen* for three orchestras. Begun in the little village in the summer of 1955 and completed two years

[1] p. 18

[2] p. 15

[3] p. 13

later, *Gruppen* demonstrates Stockhausen's serial conception of translating rhythm into timbre and vice versa by increasing and decreasing speed – an idea obviously derived from Stockhausen's experiments with tape manipulation in the electronic music studio.

In his seminal essay *...wie die Zeit vergeht...* ("how time passes" [9]), Stockhausen discussed his approach to achieving aesthetic unity by subjugating micro- and macro-time, that is timbre and rhythm, to the same compositional principles. Here, he presents a graph of a so-called group spectrum, i.e. the relation of superimposed tempi, in a very distinct shape (fig. 3). In an interview, Stockhausen revealed (almost 20 years later) that many envelopes of structural sections of *Gruppen* are precise representations of the mountain panorama he viewed from his window in Paspels [10][1].

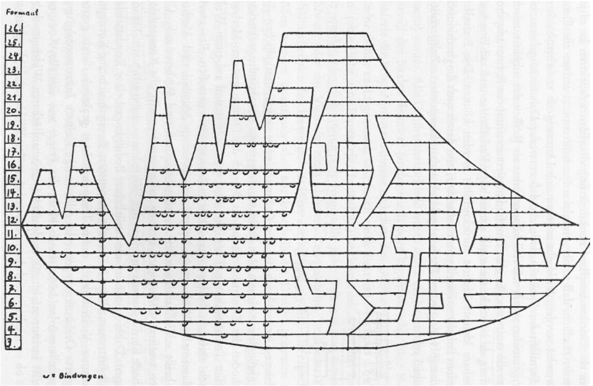


Figure 3: Karlheinz Stockhausen, sketch for *Gruppen* [9][2]. Copyright © Archiv der Stockhausen-Stiftung für Musik, Kürten (www.stockhausen.org). Used by kind permission.

Obviously, these shapes cannot be perceived by the listener. The graph controls the music on a highly abstract level and was certainly never meant to be directly experienced. Nevertheless, considering Stockhausen's metaphysical yearning for unity, and given the humor of the 27-year-old who had acted against his serial principle when he inserted a metaphorical "thunderclap" in his tape piece *Study I* (1953) on the occasion of the birth of his daughter [11][3], we witness the well-known setting of a composer establishing a hidden subtext. We might have pushed the sonification metaphor to an extreme by connecting it with a parameter mapping that resists decoding. But the composer's interest might lie not so much in communicating what s/he already knows as in creating an aesthetic situation that is open to the unknown.

3. SONIFICATION AND EXPERIMENTAL MUSIC

John Cage, one of the most prominent exponents of American experimental music, described his aesthetic concept of composition, performance and listening as being fundamentally experimental in nature, i.e. open to an unpredictable outcome or experience: "New music: new listening. Not an attempt to understand something that is being said, for, if something were being said, the sounds would be given the shapes of words" [12][4]. Here is no place for an author who expresses emotions or attitudes in music with the intention of communicating them to the listener. Instead, the composer "may give up the desire to control sound, clear his mind of music, and set about discovering means to let the sounds be themselves rather than vehicles for man-made theories or expressions of human sentiments" [12][5].

That these means are to be *discovered* is characteristic for Cage's understanding of the creative act. Discovery and experiment are scientific procedures which Cage unhesitatingly employed in the realm of composition. To remove personal preferences, he utilized various chance techniques, most notably the Chinese oracle *I-Ching*, but also the observation of imperfections in music paper (in his *Music for Piano*, 1952-56) or the "placing of transparent templates on the pages of an astronomical atlas and transcribing the positions of stars" [13][6] (in *Atlas Eclipticalis*, 86 instrumental parts to be played in whole or part, 1961/62). What was already obvious in the use of stars in *Atlas Eclipticalis*, the notion of translating meaningful data into music and thereby establishing a programmatic subtext, became more prominent in Cage's open music theater piece *Song Books* (1970). Here, the performers are asked to map the lines of a portrait of Henry David Thoreau (*Solo for Voice 5*), the profile of Marcel Duchamp (*Solo for Voice 65*) or a certain route on the map of Concord, Mass. (*Solo for Voice 3*) to a melodic line.

Cage also used electronic means to translate physical data derived from light sensors and capacitance antennas into musical parameters that would influence a complex live-electronic sound system, as in *Variations V* (1965). There, Cage finally had available the facilities to "transform our contemporary awareness of nature's manner of operation into art" [12][7]. That same year, Alvin Lucier premiered his *Music for Solo Performer*, where enormously amplified brain waves stimulate percussion instruments, and ten years later the idea of using biofeedback in the arts was prominent enough to establish a project at the Aesthetic Research Center of Canada [14]. The rise of live-electronic music and more

[1] p. 141
[2] p. 123
[3] p. 94
[4] p. 10
[5] p. 10
[6] p. 62
[7] p. 9

easily available sensor technology in the 1960s, as well as the improvement of computer performance for algorithmic composition and sound synthesis, led to a considerable increase in the number of electronic music compositions influenced by the sonification metaphor [15], [16]. To supplement research undertaken in the field of electro-acoustic music, I will here discuss two works of instrumental music that are inspired by the sonification metaphor.

3.1. John Cage: *Ryoanji*

In 1983-85, John Cage composed *Ryoanji* in five parts for flute, oboe, trombone, voice, and double bass, to be performed solo or in any combination, but always together with a part for percussion (or orchestra in unison). The scores are graphic, consisting of curved lines that indicate glissandi with time equaling space on the horizontal and pitch equaling space on the vertical axis (fig. 4). The title *Ryoanji* refers to the Ryoanji Zen garden in Kyoto, Japan, where 15 large stones are placed in 5 groups (of 5, 2, 3, 2, and 3 stones from east to west) on a slim rectangle of raked sand. Cage created the graphs in the score by placing stones from a collection of 15 at chance-determined positions on paper and tracing parts of their perimeters. Per double page, 15 to 30 stones were used, and sometimes up to four lines overlap in one instrument, so that parts need to be pre-recorded and played back during live performance [17][1], [19].

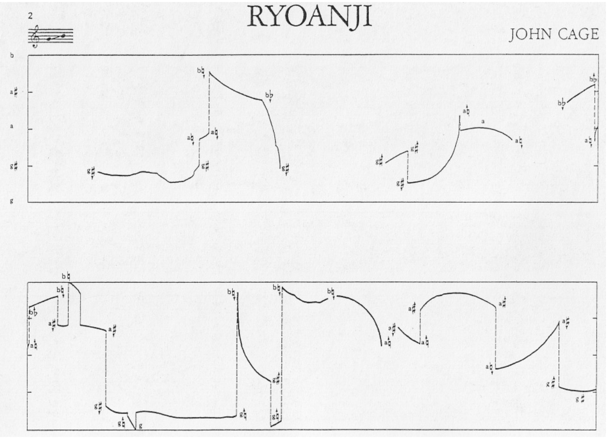


Figure 4: John Cage, *Ryoanji*. Flute [21]. Copyright © 1984 by Henmar Press Inc. Used by kind permission. All rights reserved.

(To be precise: in preparing the score, Cage actually did not draw around stones; he used paper templates that resembled the shapes of the 15 stones. This was not so much to facilitate the composition process, but to ensure that repetitions would occur: "I obviously couldn't write music with stones, because when you draw around a stone you don't necessarily draw the same way each time" [20][2].)

An important aspect in data mapping is the question of scale. The horizontal axis of the *Ryoanji* score equals time, but only the part for voice contains a tempo indication, two minutes per double page that constitute one section (or "garden"). From what we know about Cage's practice when scoring and rehearsing the first performances, we can assume the same tempo was requested for all the parts. In contrast, the scaling of the vertical axis is indicated in the scores since it changes chance-determined for every section, varying from one semitone (in the flute, pp. 6/7) to one octave and a fourth (in the voice, pp. 18/19). Obviously, the scaling factor of the pitch axis greatly impacts the sounding result. It determines whether the changes in pitch are microtonal and subtle, or large intervals are spanned in high tempo. So it comes as no surprise that Cage used this factor as a dimension of composition. By the way, musically this situation is very similar to Xenakis's transformation of particle speeds into glissandi by controlling the intervals spanned by these glissandi.

According to Cage's performance instructions, "[t]he glissandi are to be played smoothly and as much as is possible like sound events in nature rather than sounds in music" [21]. In other words, the sonification of natural objects is supposed to sound like nature, not music. But Cage's composition is not so much representational of stones he traced but of the Ryoanji garden as a whole. In his comments, Cage claims "the staves are actually the area of the garden" [20][3], and "for the accompaniment [i.e. the percussion part] I turned my attention to the raked sand" [17][4]. In summary, it may be argued that the composition *Ryoanji* is a conceptual artistic representation of the garden, incorporating elements of sonification.

Inspired by his work with magnetic tape, Cage had begun utilizing propositional notation where time equals space in the early 1950s. As he explained, "with propositional notation, you automatically produce a picture of what you hear" [20][5]. This connection of music music and visual representation can easily be turned around, so that the music follows what you see (e.g. the aforementioned music from star maps and drawings). Basically, this means nothing other than the interchangeability of visual and auditive representation, which is one of the fundamental assumptions of sonification research.

[1] pp. 134-136; also, with illustrations: [18]

[2] p. 280

[3] p. 242

[4] p. 135

[5] p. 243

3.2. Alvin Lucier: *Panorama*

In his composition *Panorama* for Trombone and Piano (1993), Alvin Lucier mapped the panorama of the Swiss Alps to the pitches of a slide trombone. He worked from a reproduction of a panorama drawing by Fritz Morach after a landscape photo by Hermann Vögeli [22]. The print (98 x 11.5 cm) indicates the mountain peaks with vertical lines that label their name and height (fig. 5).

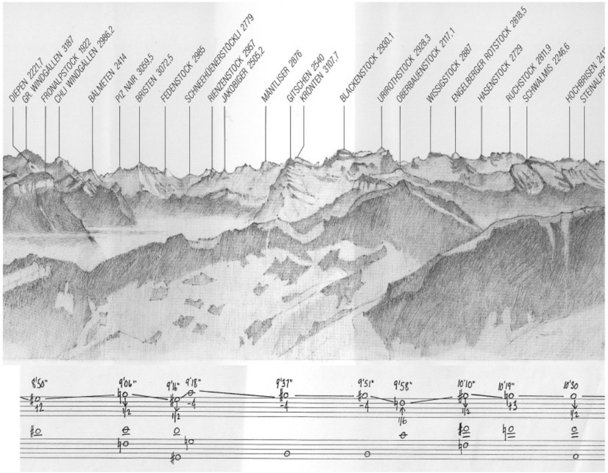


Figure 5: Alvin Lucier, Panorama. Excerpt from drawing [22] and score. Copyright © Alvin Lucier. Used by kind permission.

In his composition, Lucier kept the scaling of both dimensions fixed: The mountain height in meters divided by 8 results in the frequency in Hz, and the distance between two mountain peaks in mm is interpreted as time in seconds, leading to a total duration of 16 minutes. Since Lucier did not use a longitudinal section of the mountain range but worked from a panorama view, the distances he sonifies do not correspond to the actual distances between the peaks, a fact that is reflected in the work's title.

The piano part indicates the moments when the trombone reaches a mountain peak with a single tone or a two-tone chord of adjacent pitch classes. Since the trombone freely slides through the continuum of frequencies, while the piano is bound to pitches in equal temperament, beatings and difference tones occur. This is a typical technique Lucier has incorporated in his compositions since the 1980s.

Besides his affinity for translation processes in music and sound art, Alvin Lucier often draws his inspiration from scientific experiments. His reference to sonification satisfies both interests. It also stands for the composer's belief that his aesthetic research may reveal the beauty and charm of the world around us. He once explained that "in imitating the natural, the way the natural world works, you find out about it, and you also connect to it in a beautiful way" [23][1]. And disclosing the metaphysical implications of his aesthetic approach, Lucier stated that his works were "perhaps closer in spirit to alchemy, whose purpose was to transform base metals into pure gold" [24][2].

4. SONIFICATION'S ROMANTIC IMPLICATIONS

John Cage and many experimental music composers after him are known to have been influenced by the writings of the American Transcendentalist Henry David Thoreau. In *Walden*, an essay that reflects his experience of living alone in the woods from 1845 to 1847, Thoreau interprets the quality of sounds heard from a long distance and echoed in the valleys as "a vibration of the universal lyre" [25][3]. In 1851 Thoreau witnessed a telegraph line being erected [26][4]. Despite his doubts about the usefulness of this invention ("We are in great haste to construct a magnetic telegraph from Maine to Texas; but Maine and Texas, it may be, have nothing important to communicate." [25][5]) Thoreau enjoyed and attached importance to the sound of the "telegraph wire vibrating like an Æolian Harp" [26][6]. To Thoreau, the telegraph became "[t]he first strain of the American lyre" [27][7] and and "the divine humming of the telegraph" [27][8] revealed revealed the "spirit [that] sweeps the string of the telegraph harp – and strains of music are drawn out endlessly like the wire itself. We have no need to refer music and poetry to Greece for an origin now. […] The world is young & music is its infant voice" [26][9]. Finally, Thoreau states, the "wire […] always brings a special & general message to me from the highest" [28][10] – "the wind which was conveying a message to me from heaven dropt it on the wire of the telegraph which it vibrated as it past [sic]" [28][11]. This phrasing is indeed close to the often quoted definition of sonification as the "use of non-speech audio to convey information" [29][12].

Thoreau, however, favors the language metaphor, when he records in his journal the way he sees himself: "A writer a man writing is the scribe of all nature – he is the corn & the grass & the atmosphere writing" [26][13].

[1] p. 348

[2] p. 11

[3] p. 123 (chapter *Sounds*)

[4] p. 16 (Aug. 28, 1851)

[5] p. 52 (chapter *Economy*)

[6] p. 75 (Sept. 12, 1851)

[7] p. 3 (Feb. 13, 1854)

[8] p. 4 (Feb. 13, 1854)

[9] p. 238 (Jan. 3, 1852)

[10] p. 437 (Jan. 9, 1853)

[11] p. 76 (Sept. 12, 1851)

[12] chapter 1, *Executive Summary*

[13] p. 28 (Sept. 2, 1851)

He also claims we were "in danger of forgetting the language which all things and events speak without metaphor" [25][1].

The idea that nature implies a language that, if only understood, could reveal metaphysical entities otherwise inaccessible to mankind is prominent in German Early Romanticism of the late 18th century. Novalis, in his *Die Lehrlinge zu Saïs* ("The Novices of Sais"), compares the routes men take to wondrous figures, which seem to belong to the script of ciphers that one can behold everywhere, on wings and egg shells, in clouds, in snow, in crystals and geological formations, in filings drawn to a magnet, finally in figures created by sand on vibrating sheets [30] [2]. The latter obviously refers to the experiments of Ernst Chladni, who covered metal sheets with a thin layer of sand before he made them vibrate by means of a violin bow. Depending on the vibrations the sand would move on the surface and establish geometrical figures (fig. 6).

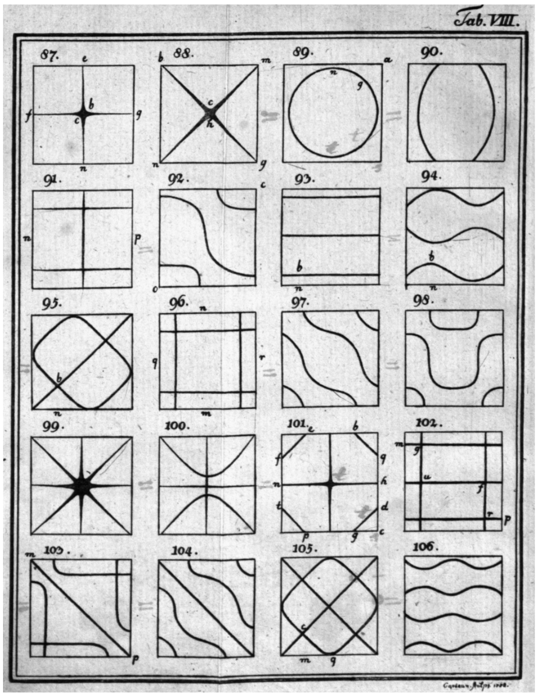


Figure 6: Chladni-Figures [31][3]

Just as Thoreau would half a century later, Novalis likened nature to an "Æolian harp, a musical instrument whose sounds are again keys to higher strings in ourselves" – a setting he calls *Ideenassociation* [association of ideas] [32] [4]. In his notes toward encyclopedism, Novalis finally asked himself whether all sculptural form, from crystal to man, could not be described acoustically, as inhibited movement. As chemical acoustics [32][5].

In their research on sonification, scholars and artists alike borrow from the romanticists' yearning to break the spell and make the world understandable by translating natural phenomena to our senses for immediate perception. Gregory Kramer opens his essential *Introduction to Auditory Display* [3] with a quote by Sufi teacher Hazrat Inayat Khan (1882-1927) from his *Mysticism of Sound and Music*: "[I]n the realm of music the wise can interpret the secret and nature of the working of the whole universe" [33][6]. In his epigraph, however, Kramer changed "music" to "sound" and thereby extended the source of inspiration from a man-made art to a physical quality as such. Similarly, Andrea Polli quotes Walt Whitman's nature poem *Proud music of the storm* when introducing her works that sonify meteorological data [34], and Chris Hayward titled a paper on the sonification of seismological data poetically – and somewhat euphemistically – *Listening to the Earth Sing* [35].

Without typecasting these and many other authors as hidden romanticists, I would like to emphasize the unspoken implication of metaphysical assumptions and romantic motives in sonification research. The recognition of these implications may not only further an understanding of the intellectual fascination with which sonification projects are received, but may also facilitate exchange between scientists and composers. The latter seems to me of prime importance if we are to improve the aesthetic and artistic quality of sonification applications and artworks.

5. ACKNOWLEDGMENTS

I would like to thank Alvin Lucier for most generously providing me access to the sketches of his *Panorama*. Thanks are also due to one of the reviewers for much appreciated detailed and constructive criticism.

[1] p. 110 (chapter *Sounds*)

[2] p. 79 (the very beginning)

[3] appendix, table VIII

[4] p. 212 (966) – "Die Natur ist eine Aeolsharfe – Sie ist ein musikal[isches] Instrument – dessen Töne wieder Tasten höherer Sayten in uns sind."

[5] p. 68 (376) – "Sollte alle plastische Bildung, vom Krystall bis auf den Menschen, nicht *acustisch*, durch gehemte Beweg[ung] zu erklären seyn. Chemische Acustik."

[6] p. 16

6. DISCOGRAPHY

J. Cage, *Ryoanji*. Therwil: hatART, CD6183, 1996. CD (Robert Black, doublebass; Eberhard Blum, flute; Iven Hausmann, trombone; Gudrun Reschke, oboe; John Patrick Thomas, voice; Jan Williams, percussion).

A. Lucier, “Panorama”. *Alvin Lucier - Panorama*. New York: Lovely Music LCD1012, 1997. CD (Roland Dahinden, trombone and Hildegard Kleeb, piano).

I. Xenakis, “Pithoprakta” [p1965]. *Xenakis - Eonta, Metastasis, Pithoprakta*. [France]: Le Chant du Monde, LDC 278 368, [1986]. CD (Maurice Le Roux, conductor; Orchestre National de l’ORTF).

7. REFERENCES

[1] F. Dombois, “The ‘Muscle Telephone’: The Undiscovered Start of Audification in the 1870s.” In *Sounds of Science - Schall im Labor (1800 - 1930)*, ed. J. Kursell, Workshop Sounds of Science, Max-Planck-Institut für Wissenschaftsgeschichte Berlin, 2006, pp. 41-45. Berlin: Max-Planck-Institut für Wissenschaftsgeschichte, 2008.

[2] S. Bly, “Presenting Information in Sound.” *Conference on Human Factors in Computing Systems*, Gaithersburg, Maryland March 15-17, 1982.

[3] G. Kramer, “An Introduction to Auditory Display.” In *Auditory Display. Sonification, Audification, and Auditory Interfaces*, ed. G. Kramer, Sante Fe Institute Studies in the Sciences of Complexity. Proceedings 18, pp. 1-77. Reading, MA: Addison-Wesley, 1994.

[4] G. Weinberg and T. Thatcher, “Interactive Sonification: Aesthetics, Functionality and Performance.” *Leonardo Music Journal*, vol. 16, pp. 9-12, 2006.

[5] A. Stone, “Being, Knowledge, and Nature in Novalis.” *Journal of the History of Philosophy*, vol. 46, no. 1, pp. 141-163, Jan. 2008.

[6] A. Goodbody, *Natursprache. Ein dichtungstheoretisches Konzept der Romantik und seine Wiederaufnahme in der modernen Naturlyrik (Novalis - Eichendorff - Lehmann - Erich)*, Kieler Studien zur deutschen Literaturgeschichte 17; Ph.D. Univ. Kiel 1983. Neumünster: Wachholtz, 1983.

[7] I. Xenakis, *Formalized Music. Thought and Mathematics in Music*. Ed. S. Kanach. Stuyvesant, NY: Pendragon, 1992.

[8] I. Xenakis, *Pithoprakta* [1956]. Score. [Bonn]: Boosey & Hawkes rental library, no year.

[9] K. Stockhausen, “...wie die Zeit vergeht...” [1956]. In *Texte zur elektronischen und instrumentalen Musik. Aufsätze 1952-1962 zur Theorie des Komponierens*, ed. D. Schnebel, Texte 1, pp. 99-139. Köln: DuMont, 1963.

[10] J. Cott, *Stockhausen. Conversations with the composer*. New York: Simon & Schuster, 1973.

[11] M. Kurtz, *Stockhausen. Eine Biographie*. Kassel - Basel: Bärenreiter, 1988.

[12] J. Cage, “Experimental Music” [1957]. In *Silence. Lectures and Writings*, pp. 7-12. Hanover, NH: Wesleyan University Press, 1961.

[13] J. Cage, “Notes on Compositions II.” In *Writer. Previously uncollected pieces*, ed. R. Kostelanetz, pp. 51-62. New York: Limelight Editions, 1993.

[14] D. Rosenboom (ed.), *Biofeedback and the Arts: Results of Early Experiments*. Vancouver: Aesthetic Research Center of Canada, 1976.

[15] A. Schoon, and F. Dombois, “Sonification in Music.” *15th International Conference on Auditory Display, Copenhagen, May 18-22, 2009*, Copenhagen 2009. www.ICAD.org (accessed Jan. 6, 2010).

[16] D. Minciacchi, “Translation from neurobiological data to music parameters.” In *The Neurosciences and Music*, ed. G. Avanzini, Annals of the New York Academy of Sciences 999, pp. 282-301. New York: New York Academy of Sciences, 2003.

[17] J. Cage, “Notes on Compositions IV.” In *Writer. Previously uncollected pieces*, ed. R. Kostelanetz, pp. 133-142. New York: Limelight Editions, 1993.

[18] J. Cage, “Ryoanji: Solos for Oboe, Flute, Contrabass, Voice, Trombone with Percussion or Orchestral Obbligato (1983–85).” *PAJ: A Journal of Performance and Art*, vol. 31, no. 3, pp. 57-64, Sept. 2009.

[19] C. Thierolf, “Plötzliche Bilder. Die Ryoanji-Zeichnungen von John Cage.” In *Hanne Darboven / John Cage. Staatsgalerie moderner Kunst*, ed. Bayrische Staatsgemäldesammlungen München, Kunstwerke 4, pp. 42-76. Ostfildern: Gerd Hatje, 1997.

[20] J. Cage, *Musicage. Cage muses on words, art, music*. Ed. J. Retallack. Hanover, NH: Wesleyan University Press, 1996.

[21] J. Cage, *Ryoanji for Flute with percussion or orchestral obbligato and ad libitum with other pieces of the same title*. Score. New York: Henmar Press, 1984.

[22] H. Vögeli and F. Morach, *Waldspitzpanorama*. Zug, Schweiz: Zuger Kantonalbank [promotional print], no year.

[23] A. Lucier and W. Zimmermann, "[Conversation]." In Zimmermann, *Insel Musik*, pp. 347-350. Köln: Beginner Press, 1981. Originally published Zimmermann, *Desert Plants*, Vancouver: A.R.C. (Aesthetic Research Centre), 1976.

[24] A. Lucier, "Origins of a Form. Acoustical Exploration, Science, and Incessancy." *Leonardo Music Journal*, vol. 8, pp. 5-44, 2005.

[25] H. D. Thoreau, *Walden* [1854]. Ed. J. L. Shanley, The Writings of Henry D. Thoreau. Princeton, NJ: Princeton University Press, 1971.

[26] H. D. Thoreau, *Journal. Volume 4: 1851-52*. Ed. L. N. Neufeldt and N. C. Simmons, The Writings of Henry D. Thoreau. Princeton, NJ: Princeton University Press, 1992.

[27] H. D. Thoreau, *Journal. Volume 8: 1854*. Ed. S. H. Petrulionis, The Writings of Henry D. Thoreau. Princeton, NJ: Princeton University Press, 2002.

[28] H. D. Thoreau, *Journal. Volume 5: 1852-53*. Ed. P. F. O'Connell, The Writings of Henry D. Thoreau. Princeton, NJ: Princeton University Press, 1997.

[29] Kramer, Gregory, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, and J. Neuhoff, *Sonification Report: Status of the Field and Research Agenda. Prepared for the National Science Foundation by members of the International Community for Auditory Display*: International Community for Auditory Display, 1997. http://www.icad.org/websiteV2.0/References/nsf.html (accessed Aug 31, 2008).

[30] Novalis, *Schriften*. Vol. 1: Das Dichterische Werk. Ed. P. Kluckhohn and R. Samuel. Stuttgart: Kohlhammer, 3rd ed. 1977.

[31] E. F. F. Chladni, *Entdeckungen über die Theorie des Klanges*. Leipzig: Weidmanns Erben und Reich, 1787.

[32] Novalis, *Das Allgemeine Brouillon. Materialien zur Enyzyklopädistik 1798/99*. Ed. H.-J. Mähl. Hamburg: Meiner, 1993.

[33] H. I. Khan, *The Mysticism of Sound and Music*. Boston: Shambhala, 1996.

[34] A. Polli, "Atmospheric/Weather Works: Artistic Sonification of Meteorological Data." In *Ylem. Artists Using Science & Technology*, vol. 24, no. 8, pp. 9-13, July/Aug. 2004. www.ylem.org (accessed Jan 16, 2010).

[35] C. Hayward, "Listening to the Earth Sing." In *Auditory Display. Sonification, Audification, and Auditory Interfaces*, ed. G. Kramer, Sante Fe Institute Studies in the Sciences of Complexity. Proceedings 18, pp. 369-404. Reading, MA: Addison-Wesley, 1994.

A SOCIAL PLATFORM FOR INFORMATION SONIFICATION: MANY-EARS.COM

Anton Schertenleib

University of Canberra,
Information Sciences & Communication,
ACT 2601, Australia
anton.schertenleib@canberra.edu.au

Stephen Barrass

University of Canberra,
Faculty Of Arts & Design,
ACT 2601, Australia
stephen.barrass@canberra.edu.au

ABSTRACT

In this paper we describe the Many Ears project that will develop the first example of a social site for a community of practice in data sonification. This site will be modeled on the Many Eyes site for “shared visualization and discovery” that combines facilities of a social site with online tools for graphing data. Anyone can upload a dataset, describe it and make it available for others to visualize or download. The ease of use of the tools and the social features on Many Eyes have attracted a broad general audience who have produced unexpected political, recreational, cultural and spiritual applications that differ markedly from conventional data analysis. The Many Ears project seeks to find out what will happen when data sonification is made more available as a mass medium? What new audiences will listen to sonifications? Who will create sonifications and for whom? What unexpected purposes will sonification be put to?

1. INTRODUCTION

From the beginnings of email in the 1970’s, the internet has always been a social medium. The online newsgroups in the 1980’s allowed people from all over the world to chat about computers, science, recreational activities, social issues and an ever increasingly diverse array of alternative .alt directions. The invention of the HTML browser in the 1990’s allowed people to produce graphical pages with distinct URL addresses, and the world wide web bloomed with individually authored sites. Search engines made it possible to surf the web using keyword topics and phrases. In the 2000’s Web2.0 technologies such as the Wiki provided a framework for the distributed authorship of a site such as Wikipedia[1]. The hugely popular Facebook[2] site is composed of more than 350 million pages produced as a consequence of self representation and social interaction. Online Content Management Systems, such as Drupal[3], enabled communities to develop sites such as the ICAD[4] site in a bottom up manner. Content from different sites can be recombined in a Mashup constructed from material that has been aggregated and recomposed using RSS feeds. Sites such as Flickr[5] and Google Maps[6] provide Application Programming Interfaces (API’s) that allow other sites to access content from their online databases. Government departments, bureaus and agencies, museums, galleries and historical archives are similarly making data collections available online and accessible through API’s. Beyond access, online service providers provide data processing such as a text to speech synthesizer that can be embedded into another site.

Many Eyes[7] is a site that combines social facilities with online services to enable “shared visualization and discovery”. The social facilities of the site include personal member pages, discussion topics, comments, ratings, watches and content sharing from the site. Participants can upload a dataset, describe it and make it available for others to visualize or download. The data can be visualized in 15 different ways that include line graphs, scatter plots, column/bar, pie, scatter and bubble charts. Data from different data sets can be mashed together to explore new questions. The extension of data visualization as a mass medium to a general audience has resulted in unexpected political, recreational, cultural and spiritual applications that differ markedly from the scientific analysis of data[8].

Music is a mass medium that has been an important cultural force throughout the 20th century in movements such as Folk, Rock and Tropicalismo. The social power of music combined with the social extension of Many Eyes raise the question of what would happen if data sonification was to become a mass medium too? What new audiences would sonification reach, and what new purposes would it be put to?

This paper describes the Many Ears[9] project that seeks to answer these questions. This project is based on the development of the first social site for data sonification. The following sections begin with background on the motivation for data sonification. We then overview existing tools for data sonification and the techniques they provide. The next section then provides an overview of the facilities provided by online sites for social visualization. The main section then presents a specification and plan for the development of the Many Ears site that integrates the key aspects of social visualization sites with key sonification techniques. The final section briefly describes future work on data collection that will be used to answer questions about the audience and applications of the site.

2. BACKGROUND

2.1. Sonification

Sonification is generally defined as the design of non-verbal sounds to convey useful information[10]. The advantages of non-speech audio for conveying information include:

- Sensitivity to temporal relations and rapid changes
- Multi-sensory perception of multi-parameter datasets
- Accessibility for the visually impaired,
- Explore datasets in frequency rather than spatial dimensions
- Identify new phenomena that current display techniques miss
- Find otherwise hidden correlations and patterns masked in visual displays
- Monitor data while looking at something else (background event-finding)
- Complement existing visual displays (since the ear is sensitive to different frequency bands and patterns than the eye)

The invention of the MIDI protocol in the early 1980's allowed scientists to connect computers to pop-music synthesisers and listen to data-sets from their experiments. In the 1990's sound cards for computer gaming enabled a broader range of sonifications that were not constrained by the MIDI palette of musical instruments. In parallel with the developments in audio hardware there have also been developments in software tools for sound synthesis. Many sonification researchers today use sound tools designed for computer music like those listed in Table 1. These tools are standalone applications that are compiled for different platforms. They provide a general range of synthesis techniques that typically include additive, subtractive, FM, formants and granular algorithms.

	Csound [11]	**Max/MSP [12]**	**SuperCollider [13]**	**Pure Data[14]**
User interface type	graphical	graphical	document	graphical
API / interface	shell script, Python, TCL, Java	C, Java, Python	shell script	shell script, Java, Python
Development status	mature	mature	stable	mature
Cost	free	495$	free	free

Table 1: Tools for computer music and sound synthesis.

However sonification involves a concern with data that is peripheral in computer music. This concern entails a focus on data formats, selection, statistics, signal processing, mapping and representation. Sonification researchers have developed strategies for mapping data into sounds that include audification, parameter mapping, model-based and stream-based approaches. Sonification tools have been developed that focus on importing data and various sonification strategies and techniques, as shown in Table 2. These are also desktop applications for individual use.

	SoniPy[15]	**Sonifyer [16]**	**Sonification Sandbox[17]**	**xSonify[18]**
Development status	prototype	stable	stable	prototype
Import data	TXT, binary, SQL	EEG[1] format, TXT	MS Excel, CSV	TXT, Web Service Interface
Technology	Python	Cocoa	Java	Java
Sound synthesis	n/a	FM Synthesis	MIDI-fication[19]	MIDI-fication[19]
Sonification type	n/a	parameter mapping, audification	parameter mapping	parameter mapping
Sound rendering	Audio file	Audio file	MIDI file	MIDI file
Cost	free	free	free	free

Table 2: Sonification tools.

2.2. Collaborative Visualizations

A summary of the most popular collaborative visualization projects currently available in the internet is shown in Table 3. All projects are free of charge but Swivel also has an optional monthly fee for access to additional functionalities related to group collaborations.

	DEVise [23]	**Data360 [24]**	**Swivel[25]**	**Many Eyes[7]**
Data types	numerical	numerical	numerical	textual, numerical
Visualization techniques	2	-	7	15
Upload own data	-	✓	✓	✓
Online data manipulation	-	-	✓	✓
Collaboration/forum	-	✓	✓	✓
Registered users	-	n/a	16.445	37.847
Visualizations	n/a	n/a	n/a	50.009
Uploaded datasets	-	6.831	n/a	96.736
Audience	research	industries, research, politics, general public, education	industries, research, politics, general public, education	industries, research, politics, general public, education
Costs	free	free	starting at 12$/month	free

Table 3: Online data visualization projects.

[1] EEG: Electroencephalography

The idea of collaborative visualization began in the late 1990's with the DEVise project that consists of a Java desktop application and a Java Applet[22]. The visualization was prepared offline with a desktop application that allowed you to import data, define the data schema description and the visualization parameters, and then visualize the data. The Java Applet called DEVise JavaScreen then made it possible to share the visualization over the internet. The initiator of a sharing session can control the application and views of the current visualization, and other people can join the session as viewers. However the shared interface did not support interpersonal communications which had to be done through other channels such as email, phone or chat.

A web2.0 site called Data360 site came online in 2004 as a "Wiki for Data" that provides facilities for uploading datasets and visualizations. Users can create a platform and start uploading data sets they want to visualize. If a platform is declared public other data360 users can start to subscribe and access the data sets and visualizations as viewers. They can start discussions about the data and even create reports based on the data.

Swivel and Many Eyes launched at almost the same time around the end of 2006. Many Eyes now has over 37,000 registered users, more than double the number in the Swivel user base. Many Eyes offers 15 kinds of visualization (including word clouds) while Swivel offers 7. Data from different data sets can be combined to explore new questions.

2.3. Sonification Community

The communal website of the International Community for Auditory Display[4] has been build with a Web2.0 tool called Drupal. Anyone who is interested in Auditory Display may register, and then contribute to the site through the various community forums and discussion lists. The navigation bar on the site has links to pages titled About, Conferences, Awards, ICAD Board, Knowledge Base, Press, Community Area, Audio, News Aggregator, and Contacts.

The Conferences page provides links to previous and current International Conferences on Auditory Display beginning from 1992 in Santa Fe. The Awards page shows awards made at those conferences. The Board page shows the current Board members and their affiliations. The Knowledge Base includes 10 Papers to Start, 10 Audio Examples, a ToDo list, the NSF Whitepaper on Auditory Display, a Bibliography of all papers published in the ICAD conference proceedings, and many more resources. The Media page shows Stories about ICAD on TV, radio, in newspapers or other media.

The Community Area lists the registered Members of the site, Editors who can change pages at the request of Members, Forums, Mailing Lists, Polls, and Treasure Hunt 2008. The Audio page provides a place for sonification audio files. Currently there is only one sonification there, with 2541 downloads and 850 plays. The page does not provide an upload mechanism so presumably a Member must send their sonification file to an Editor for it to be uploaded. The News aggregator has RSS feeds from relevant and interesting articles on the internet. The Contact page provides email links to the site Editors.

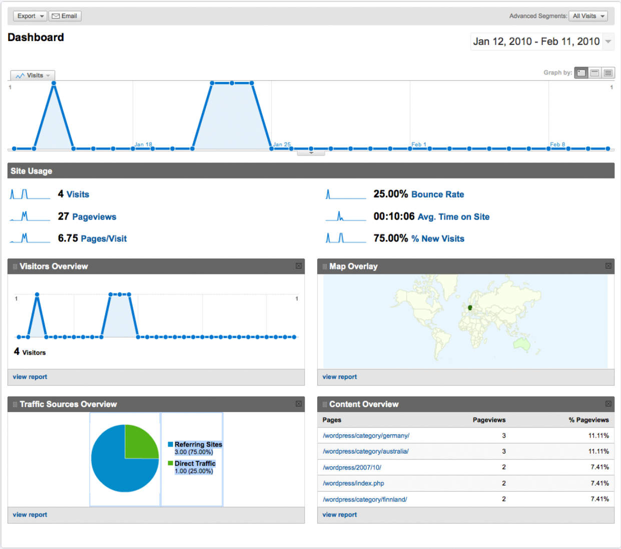
The ICAD community has also been developing a more general public awareness and appreciation of sonification through a series of concerts that began at ICAD 2004 in Sydney. This first concert titled Listening to the Mind Listening was staged at the Sydney Opera House and attracted an audience of more than 350 [20]. The concert consisted of 10 sonifications of EEG brain data recorded from someone listening to a piece of music. These sonifications were selected from thirty submissions. The public concert of sonifications was repeated at ICAD 2006 at the ICA in London in which there were 8 sonifications of data from the CIA factbook on world population and resources [21]. A social web2.0 style website was developed for the sonification competition at ICAD 2009 in Copenhagen. This website provided online access to a dataset of DNA from yeast. Participants were able to download the dataset, and upload a sonification with a description and credits. There were 50 sonifications submitted in categories of musical and scientific. Only three sonifications were submitted in the scientific category. Anyone could listen to and rate the sonifications on a around 10 descriptive scales such as boring, fascinating, confusing and others. The submissions typically received more than 100 plays over a period of two months. However the site is no longer visible and the sonifications and related data about engagement with them are unfortunately no longer accessible.

3. PROJECT PLAN

3.1. Phase 1: Initial site

Step 1 was to register the URL www.many-ears.com as a reference to the concept of social visualization established by Many Eyes. Next we needed to find a technology that could enable the implementation of a site with social features as well as tools for online sonification. After some feasibility studies we chose to build the back end of the web application with Java J2EE[26] and associated web development frameworks. For persistent data storage we chose MYSQL[27] for the database as well as the option to store files directly in a file store structure referenced by a unique identifier. This backend provides a login and file upload functionality. The front end is HTML[28] based with Adobe Flash[29] elements embedded e.g. an audio player.

The initial site consists of a short introductory animation that introduces the Many Ears project followed by a mock up of the user interface in Flash. The establishment of the site provides the basis for the ongoing collection of data about usage. The screenshot from Google Analytics[30] in Figure 1. gives an overview of traffic to the site during the month from 10 January to 9 February 2010. This shows there were at total of 15 visits on 5 days, and that the largest proportion of the traffic came from referring sites. Google Analytics will be used to monitor the activity on the site throughout the project, and to analyse the effect of the introduction of the social features and online tools in each phase.



Dashboard

Jan 12, 2010 - Feb 11, 2010

Site Usage

4 Visits
27 Pageviews
6.75 Pages/Visit
25.00% Bounce Rate
00:10:06 Avg. Time on Site
75.00% % New Visits

4 Visitors

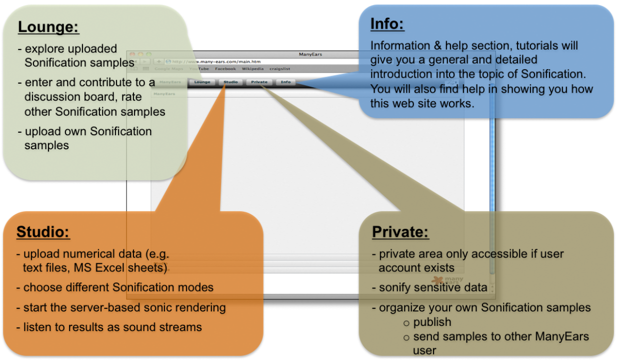
Content Overview

Pages	Pageviews	% Pageviews
/wordpress/category/germany/	3	11.11%
/wordpress/category/australia/	3	11.11%
/wordpress/2007/10/	2	7.41%
/wordpress/index.php	2	7.41%
/wordpress/category/finnland/	2	7.41%

Figure 1. Google Analytics for www.many-ears.com

3.2. Phase 2: Social features

Many Ears is a social site in which the content is generated by the community of practice. Registered participants have a personal page similar to Facebook where they can provide their personal profile, upload datasets, and upload sonifications.
The personal profile consists of a text ID, a visual ID, a sonic ID, a tag list, and a text box for describing contacts, affiliations, interests, external links or anything else.
The data area allows the upload of datasets, subject to a file size limit. Data sets each have a tag-list, a text description box, and a checkbox for public or private access. Public data sets have an additional comments box, and a recommendation icon. Registered participants can leave comments and recommendations, and can download public data-sets.
The audio area allows the upload of audio files in WAV or MP3 format, subject to a file size limit. Audio files each have a tag-list, a text description box, a checkbox for public or private access and an audio player. Public audio files have an additional comments box, and a recommendation icon. Registered participants can leave comments and recommendations, and can download public audio files.
Visitors to the site can browse the personal pages in the Lounge area, shown in Figure 2.



Lounge:
- explore uploaded Sonification samples
- enter and contribute to a discussion board, rate other Sonification samples
- upload own Sonification samples

Info:
Information & help section, tutorials will give you a general and detailed introduction into the topic of Sonification. You will also find help in showing you how this web site works.

Studio:
- upload numerical data (e.g. text files, MS Excel sheets)
- choose different Sonification modes
- start the server-based sonic rendering
- listen to results as sound streams

Private:
- private area only accessible if user account exists
- sonify sensitive data
- organize your own Sonification samples
    - publish
    - send samples to other ManyEars user

Figure 2. Many Ears with function descriptions

The social feature is currently designed to support online submissions to the ICAD 2010 concert event [31]. For this we would like to provide basic functionalities for people to upload their compositions, adding meta information like title, descriptions etc. All the compositions can be displayed in a list which could be ordered by different categories. Registered users can also rate and evaluate the sound samples. Commenting functionalities are also planned to encourage discussion of the submissions. This will also provide a starting point for collection of data about the effect of social features. Quantitative data about traffic on the site will be collected with Google Analytics. Qualitative information about the higher level socialisations will be analysed from comments, ratings and recommendations.

3.3. Phase 3: Online Sonification

Phase three will introduce online tools for data sonification into the Many Ears site. This will consist of the addition of a new area where participants can select a tool to apply to a dataset in their data area to produce a sonification in their audio area. A Sonification Template comprises a list of settings and parameters which will be handed in the next step to the back end sound render software. The number of templates is scalable and can be extended. In future it may also be possible for participants to design and upload their own sonification templates into the site.

The initial templates will be based on the sonification algorithms that were implemented previously in the xSonify[18] tool based on Java MIDI API:

- MIDI Notenumber parameter mapping - each data value is mapped to the notenumber of a MIDI instrument at constant volume
- MIDI Loudness mapping – each data value is mapped to the volume of a MIDI instrument at constant pitch.
- The average value over a certain number data values is mapped to the repetition rate of a MIDI instrument at a constant pitch and loudness.

Once these templates have been established based on existing Java technology we will progress to more sophisticated sonification techniques that will include audifications, multi-parameter mappings, redundant mappings, model-based and stream-based, and metaphors such as a Geiger-counter. These sonifications will be implemented by linking Many Ears to sound synthesis engines such as Pure Data[14] or SuperCollider[13] that offer application programming interfaces (API) based on Java technologies.

Once the general online synthesis capability has been implemented it will be made accessible through an online API to create an online sonification service that can be embedded in any website. This enables 3rd party applications/systems to use the Many Ears platform as external sonification rendering tool. The purpose of this feature is to enable data visualization sites to use sonification as an alternative to their already existing visualizations.
From the technical side we are planning to use Java Servlet Technologies[32]. In Many Ears registered 3rd party

applications/systems upload the data attached with sonification parameters and receive a unique identifier in return. This unique identifier can be used to download the rendered sound sequence by invoking a download servlet or if not finished the progress information of the rendering taking place on Many Ears. An example of a similar internet based service provider is a text to speech application from Vozme[33].

3.4. Phase 3: Data collection and analysis

After the deployment of the sonification tools in Phase 2 we will measure if there is an increase in users, traffic, comments, number of datasets being uploaded, social interactions and actual sonification activities. The goal is to develop a better understanding of the relevance sonification might have on the general public. For this we would like to gain insights in the acceptance and behaviors of people from different backgrounds using sonification and also to attract new audience to this field.

For the collection of quantitative information from the participants we suggest two categories of data acquisition:

3.4.1. Passive data collection: Logging information in the background

Data acquisition where the user doesn't have to play an active role will be Many Ears' logging mechanisms running in the background. We will use this data to answer the following questions:

- Where does the user come from? Which country and even city?
- The visitors loyalty. If and how often is a user returning to Many Ears?
- Which functionalities of Many Ears seem to be the most interesting?
- Which sonification tools/methods seem to be the most popular in terms of certain kinds of data, applications or particular audiences?
- Which association is there between the users discipline and their sonification preferences?
- Number of registered users, amount uploaded datasets and sonification samples?
- Number of contributions to discussions?
- Numbers of downloaded sonification compositions?

3.4.2. Active data collection: Surveys and forum

Active data collection is where the participants are encouraged to discuss their experiences with other users. All discussion are stored in Many Ears database and accessible to all other Many Ears users. Going through all the written comments will be the most challenging and laborious part of the data analysis at the end of the project, but we expect it will be invaluable and very exciting.

The rating option, also mentioned above, will also provide information about the participants preferences for sonification compositions along with information of the participants personal profile. Other qualitative comments about the site will be collected by online surveys that will appear as a dialog when entering the website.

4. SUMMARY

Many Ears is the first example of a social site for online sonification. The social features of the site will allow the participants to generate the content in a manner similar to Facebook and other web2.0 sites. These social features will include a personal page where participants can provide a profile and upload datasets and sonified audio files. Visitors to the site will be able to browse these pages to discuss the data sets and sonifications and provide give ratings. The site will also be the first example of an online sonification rendering engine. Participants will be able to apply a sonification technique to a data set to produce an audio rendering online in their page. The range of sonification techniques is scalable and it is intended to allow participants to add new techniques. The rendering engine will be made available to other websites through and online API. The effects of the site will be monitored and analysed through quantitative analysis of site traffic, and quantitative analysis of social discourse on the site. We hypothesise that the introduction of a social site with online tools will enable sonification to reach new audiences and extend it towards a mass-medium that will have new and unexpected applications.

5. FUTURE WORK

We are planning a long term collaboration with the Shirohisa Ikeda Project[35] in Puerto Rico. This collaboration is aiming on the introduction of sonification to high school students. The teaching methods in schools as we know it are laid out visually. Many Ears could act as an alternative way of working with data.

6. ACKNOWLEDGEMENT

We are pleased to acknowledge Mr. Robert M. Candey of the NASA Goddard Space Flight Centre for his helpful input and discussions on this paper and his expertise as an external advisor on Anton Schertenleib's Ph.D. Supervisory panel.

7. REFERENCES

[1] Wikipedia, Online Encyclopedia - http://wikipedia.org, retrieved 4 February 2010

[2] Facebook, Social Networking - http://www.facebook.com, retrieved 4 February 2010

[3] Drupal, PHP based Content Management System - http://www.drupal.org, retrieved 4 February 2010

[4] ICAD, International Community for Auditory Display - http://icad.org

[5] Flickr, Online photo sharing platform - http://flickr.com, retrieved 4 February 2010

[6] Google Maps, Online street map platform - http://maps.google.com, retrieved 4 February 2010

[7] Many Eyes Project, data visualization web site - http://www.many-eyes.com, retrieved 4 February 2010

[8] Fernanda B. Viégas, Martin Wattenberg, Matt McKeon, Frank van Ham, Jesse Kriss, *Harry Potter and the Meat-Filled Freezer: A Case Study of Spontaneous Usage of Visualization Tools,* IBM T. J. Watson Research Center, 2008

[9] Many Ears Project, acoustical data visualization web site - http://www.many-ears.com, retrieved 4 February 2010

[10] Gregory Kramer, Bruce Walker, Terri Bonebright, Perry Cook, John Flowers, Nadine Miner, John Neuhoff, *Sonification Report: Status of the Field and Research Agenda,* National Science Foundation, 1998

[11] Csound, Sound synthesis software - http://www.csounds.com, retrieved 4 February 2010

[12] Max/MSP, Sound synthesizer software - http://cycling74.com, retrieved 4 February 2010

[13] SuperCollider, Sound synthesizer software - http://supercollider.sourceforge.net, retrieved 4 February 2010

[14] Pure Data, Sound synthesizer software - http://puredata.info, retrieved 4 February 2010

[15] SoniPy, Python sonification tool - http://www.sonification.com.au/sonipy/index.html, retrieved 4 February 2010

[16] Sonifyer, Sonification tool - http://www.sonifyer.org, retrieved 4 February 2010

[17] Sonification Sandbox - Sonification tool - http://sonify.psych.gatech.edu/research/sonification_sandbox/index.html, retrieved 4 February 2010

[18] xSonify - NASA sonification project - http://spdf.gsfc.nasa.gov/research/sonification/sonification.html, retrieved 4 February 2010

[19] Schaffert, N., Mattes, K., Barrass, S., Effenberg, A.O., *Exploring function and esthetics in sonifications for elite sports,* 2nd Int. Conference on Music Communication Science, Sydney, 2009

[20] Barrass, S., Whitelaw, M., Bailes, F. A. Listening to the Mind Listening: analysis of reviews, sonifications and designs Leonardo Music Journal vol16: 13-19, 2006

[21] ICAD 2006, ICAD conference in London - http://www.dcs.qmul.ac.uk/research/imc/icad2006, retrieved 4 February 2010

[22] Java Applet Technology - http://java.sun.com/applets, retrieved 4 February 2010

[23] Devise, data visualization web site - http://pages.cs.wisc.edu/~devise/index.html, retrieved 4 February 2010

[24] data360, data visualization web site - http://www.data360.org, retrieved 4 February 2010

[25] Swivel, data visualization web site - http://www.swivel.com, retrieved 4 February 2010

[26] Java J2EE Technology -

[27] MYSQL, Relational DBMS - http://www.mysql.org , retrieved 4 February 2010

[28] HTML, Hyper Text Markup Language - http://www.w3.org/MarkUp, retrieved 4 February 2010

[29] Adobe Flash - http://www.adobe.com/products/flash, retrieved 4 February 2010

[30] Google Analytics, Web traffic analysis platform - http://www..google.com/analytics, retrieved 4 February 2010

[31] ICAD 2010, ICAD conference in Washington D.C. - http://www.icad.org/icad2010, retrieved 4 February 2010

[32] Java Applet Technology - http://java.sun.com/products/servlet, retrieved 4 February 2010

[33] Vozme, Online text to speech service - http://www.vozme.com, retrieved 4 February 2010

[34] Shirohisa Ikeda Project - http://www.shirohisa-ikeda.org, retrieved 4 February 2010

SONIC TRIPTYCHON OF THE HUMAN BRAIN

Thomas Hermann

Ambient Intelligence Group, CITEC
Bielefeld University, Bielefeld, Germany
thermann@techfak.uni-bielefeld.de

Gerold Baier

Manchester Interdisciplinary Biocentre
University of Manchester, Manchester, UK
Gerold.Baier@manchester.ac.uk

ABSTRACT

This paper describes the motivation, data and sonification technique for three sound examples on the auditory display of human brain activity, selected and formatted as contribution to the ICAD aural submission category. The human brain generates complex temporal and spatial signal patterns whose dynamics correspond to normal (e.g. cognitive) processes and as well as abnormal conditions, i.e. disease. Our sonification technique *Event-based Sonification* allows to render multi-channel representations of the multivariate data so that temporal, spectral and spatial patterns can be discerned. Being a scientific approach, the sonifications are reproducible, systematic and the mapping is made transparent. Control parameters help to increase the saliency of specific features in the auditory display. This is demonstrated using data with sleep spindles, a photic response and epileptic discharges. Since all 'sonic pictures' are rendered with the same technique, a variety of dynamic phenomena related to different brain states are demonstrated as auditory Gestalts. Sonification of the EEG offers a meaningful complement of the prevailing visual displays.

1. INTRODUCTION

The human brain is the most complex organ/device in operation and a strongly investigated target in many disciplines, from medicine over cognitive neuroscience to psychology and computer science. Although the basic principles of neural activation are known, it is still unclear how the brain implements its fantastic processing capabilities at a larger scale, or how exactly malfunctions correspond to pattern changes. In particular, there is a need for novel techniques to elucidate how normal activity and disease-related abnormalities influence cerebral dynamics,

Sonification is a promising approach to better understand the brain since (a) cerebral activity forms complex spotio-temporal patterns of rhythms and synchronization dependencies. Listening is therefore ideal to pick up such structures and their changes over time; (b) brain activity is prominently organized in the spectral domain where rhythmic patterns at specific frequencies are known to correspond to brain activity – this naturally matches the built-in spectral decomposition of the listener's ears; and (c) brain activity is spatially organized over the cortex – which fits to our ability to interpret spatial sound patterns in terms of localised sources.

EEG audification (i.e. direct playback of the signal) is a standard technique in neurophysiology to observe the firing of individual neurons, yet for large-scale multivariate recordings audification is not the technique of choice. We present in this paper sonifications created by our approach of *Event-based Sonification* and show that this is a successful means to better understand the complex dynamics of the working and diseased brain.

2. EEG: RECORDING AND ANALYSIS

The electroencephalogram (EEG) is a standard non-invasive method to continuously record human cerebral activity at high temporal resolution. The EEG measures the electric potential on the scalp at pre-specified electrode positions with respect to a reference signal. The signals are typically depicted as multivariate time series as shown in Fig. 1. Neurologists try to detect characteristic patterns (e.g. related to epileptic conditions) from such visualizations, and are required to analyze hundreds of patient EEG recordings to develop this skill. Nevertheless, due to the overwhelming complexity of human brain dynamics important problems remain. We suggest that human auditory signal processing offers a complementary mode of perception of such data with specific benefits in a clinical setting [1].

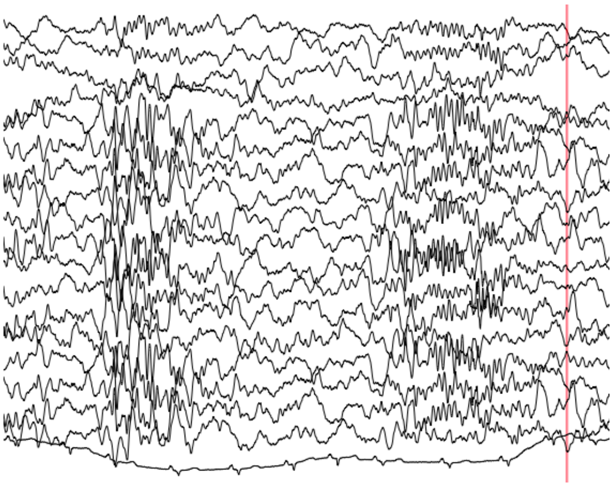


Figure 1: Typical signal plot of EEG data, here during sleep. Characteristic changes in the left and right half of the figure are caused by the so-called sleep spindles.

3. EVENT-BASED SONIFICATION

Many normal (e.g. sleep spindles, see Fig. 1) and abnormal (e.g. epileptic, Part 3, below) rhythms occur in the range of 0.1-30 Hz, corresponding to the phenomenon of rhythm rather than pitch in a real-time display. While it is easy to see global changes in rhythm as in Fig. 1 it is often hard to specify and interpret complex rhythmic relationships between multiple channels. A large number of sonification techniques have been proposed and tested, yet, from our current experience, Event-based Sonification seems to be the most promising candidate.

Event-based Sonification is a real-time method to transform the multivariate data stream (typically, between 19 and 30 channels, sampling rate between 100 and 1000 Hz) into a set of events which are then represented by sonic events whose superposition constitutes the sonification. As event we define local maxima in the data series, but we allow flipping of the signal polarity if necessary, for instance to better align the patterns of spike-wave discharges. For each event, the level difference to the previous minimum and the time difference to the previous maximum are extracted as important features for later use in the sonification.

Since even a single data series is a noisy mixture of several frequencies, there is the risk that by taking only local maxima relevant lower-frequency events are missed. This is coped with by defining a procedure to extract *hierarchic maxima*, which means to detect also the maxima of the maxima event series as maxima of 2nd order and so on until a lower bound (e.g. 1 Hz) is reached. From the time to the previous maximum at the order of analysis we can estimate the rhythm frequency, and accordingly we create sound events of different pitch so that faster rhythms are represented by rhythmical sound events at higher pitch. The detailed technique is more complex and explained in [1, 2].

There are two important parameters to filter the massive stream of sonic events according to a window of analysis: (i) we can set a center of interest on the frequency scale, causing nearby events to be pronounced in level and brilliance, (ii) we can set a threshold of rhythmical persistence, filtering of all events that occurred less regularly than defined by this limit, which eliminates sporadic or random events.

For example, for epileptic rhythms at 3 Hz (typical absence seizures), a corresponding center of interest frequency of 3 Hz and a high threshold leads to sonifications where spontaneous EEG is almost suppressed yet clearly rhythmic sound events stand out during epileptic episodes.

4. SONIC TRIPTYCHON - THE IDEA

For the current contribution we have selected three data segments recorded from different subjects/patients during different conditions. Each sonification lasts several minutes so that the listener can tune in to the typical soundscape of the signal. Despite the fact that the overall brain activation is largely unaffected by specific conditions (e.g. differences in extremely different emotional states will have essentially no impact on the EEG), many other conditions produce characteristic rhythmic changes. We have attempted to optimize the sonifications such that the most characteristic differences can be perceived from the audio stream. The three sonifications with more than 10 minutes duration give ample occasion to the listener to get accustomed to the stationary background patterns and slowly learn to differentiate normal from abnormal structure in the sound stream and thereby in the EEG data. The three 'sonic brain images' give three different perspectives[1] of cerebral activity. While the examples are rendered here for stereo presentation, they can be perceived as 16 channels spatial audio in our laboratory, creating the impression of an immersion in the brain. We emphasize that our sonification technique is strictly systematic and will render reproducible sonifications with the same data, so that the sound is a valid medium to interpret the underlying structure.

As a visual prelude to the sonification pieces, an animation is played that provides relevant background information for the experimental condition and patterns. Then the sound is played in a darkened room to reduce distraction from pure listening, as done previously in [3]. Media examples for the present contribution are provided at our website [2]

5. SONIC TRIPTYCHON PART 1: SLEEP EEG

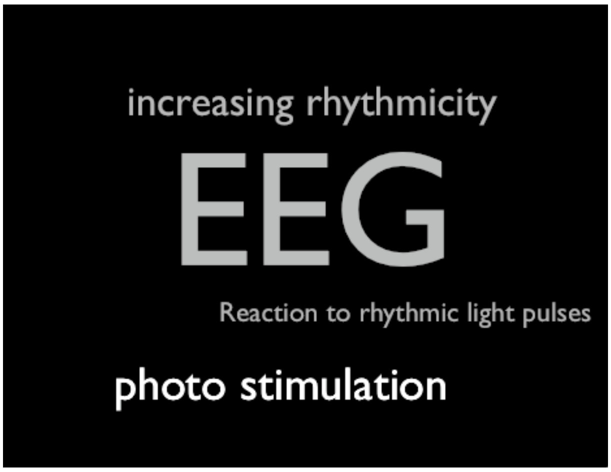
The first EEG data set is recorded during sleep. A characteristic pattern in stage 2 sleep is the so-called sleep spindle, episodes of activity in the lower beta band around 12-14 Hz. These occur in bursts interrupted by periods of reduced rhythmic activity. We use the center of interest frequency filter to improve contrast of activity in this region, leading to clearly audible event trains. Occurrence, duration, and rhythmic arrangement of sleep spindles can be perceived. In addition, the repetitive occurrence allows an estimation of mean duration and rhythmic variability. Since rhythm frequency maps to the pitch of the events, change of pitch is directly related to change of spindle frequency.

6. SONIC TRIPTYCHON PART 2: PHOTOSTIMULATION

The second EEG data set was recorded during photostimulation (PS), a technique to examine the influence of repetitive light flashes at certain frequencies on brain activity. Photosensitive subjects show a change of EEG background pattern in response to such stimuli and in epilepsy patients this can lead to a number of so-called photoparoxysmic responses, exceptionally even to the start of a seizure. The present data were recorded from a patient where the photoresponse was classified as type 4, i.e. abnormal activity was found during PS over most of the scalp. In the sonification we hear that during the stimulation a more and more pronounced rhythm at about 3 Hz is established which continues to the end of the recording. Note that the stimulus is not represented in the this sonification.

[1] or "peraudects", to use a newly invented sonic analogue term to spectare

[2] see our website `http://www.techfak.uni-bielefeld.de/ags/ami/publications/HB2010-STO`



increasing rhythmicity

EEG

Reaction to rhythmic light pulses

photo stimulation

7. SONIC TRIPTYCHON PART 3: EPILEPSY

The final EEG data set is from a patient with epilepsy. Following normal background activity for about one minute, the first seizure starts apparently spontaneously and spreads quickly over the entire cortex with a characteristic rhythm of about 3 Hz. During the seizure there is a stronger synchronization between channels (coincidences of events) and less rhythmic variability (events are not filtered due to higher regularity in repetition). Finally, epileptic activity dissolves and the dynamics returns to normal background EEG. There are two seizure episodes in the sound example. It is an interesting question whether there are any precursors of the seizures in the EEG data that could be detected in the sonification which but were overlooked in the plots. Devices optimised for this task could help to develop warning methods but such a goal was not pursued in the present context.

epileptic

EEG

clear rhythm

background

generalized seizure

8. DISCUSSION AND CONCLUSION

From the EEG sonifications it can be understood that there is continued activity in the brain at a broad range of frequencies. Cognitive and emotional phenomena that are perceived as dramatic subjectively influence this permanent neural activity only to a very limited extend in single recordings. In the case of sleep EEG, the sleep spindles are very characteristic and quickly constitute an auditory gestalt that can be picked up, remembered and rediscovered, for example during the next phase of stage 2 sleep or even in sonifications of stage 2 sleep in another subject. Disorders that are associated with disturbances of sleep rhythms like depression could benefit from the auditory approach.

For the sonification of EEG during PS we have emphasized general rhythmic properties of the whole scalp. As the corresponding EEG responses are in general less regular one would define new features to make specific abnormal responses more salient for clinical applications. Of particular use is the possibility to work on-line, as e.g. the detection and risk management of photo-induced responses could be accelerated with the auditory mode of data display.

The epileptic EEG discharges are heard as rhythmic spatially distributed sound events with clear rhythmic relationships. In the auditory mode, the temporal evolution of the multivariate rhythmic patterns can be both intuitively understood and analytically studied, e.g. the slow frequency change during the seizure, the grouping of events in certain channels during some times but not in others and so on. The epilepsies have been named as one example of a dynamical disease [4] a class of disorders that is particularly characterized by qualitative changes in rhythmic patterns. We propose that all dynamical diseases are potential candidates for promising future developments in sonification research.

Concerning what to do next for the establishment of the technique in clinical practice we seem to face a ‘chicken or the egg’ dilemma: On the one hand, it is not yet clear what the best sonification technique will be for specific problems. On the other hand, who is willing to invest time to learn the ’sonic language’ without knowing that this is going to be the eventually most successful method? Our hope is that by starting with very concrete pathologies (e.g. epilepsy) we can optimize the sonification method such that already known patterns are displayed in both modes and will lead to the recognition and acceptance of auditory displays. This provides the opportunity for professionals to see the benefits to become familiar beyond this threshold of acceptance, and discover patterns which are beyond the current limit of understanding.

9. ACKNOWLEDGEMENT

We thank Prof. Dr. Ulrich Stephani, Dr. Hiltrud Muhle und Dr. Gert Wiegand from the Neuropediatric Clinic, University of Kiel, Germany, for the cooperation on data analysis and advice with questions concerning normal and abnormal brain activity.

10. REFERENCES

[1] G. Baier, T. Hermann, and U. Stephani, “Event-based sonification of EEG rhythms in real time,” *Clinical Neurophysiology*, vol. 118, no. 6, pp. 1377–1386, 06 2007.

[2] ——, “Multi-channel sonification of human EEG,” in *Proceedings of the 13th International Conference on Auditory Display*, B. Martens, Ed., International Community for Auditory Display (ICAD). Montreal, Canada: ICAD, 06 2007, pp. 491–496.

[3] T. Hermann and G. Baier, “Die Sonifikation des menschlichen EEG,” in *Katalog: Wien Modern 2008*, B. O. Polzer, Ed. Wien: Verein Wien Modern, 11 2008, pp. 25–27.

[4] L. Glass and M. C. Mackey, “Pathological conditions resulting from instabilities in physiological control systems,” *Ann N Y Acad Sci*, vol. 316, pp. 214–35, 1979.

RE-SONIFICATION OF GEOGRAPHIC SOUND ACTIVITY USING ACOUSTIC, SEMANTIC, AND SOCIAL INFORMATION

Alex Fink, Brandon Mechtley, Gordon Wichern, Jinru Liu, Harvey Thornburg, Andreas Spanias, and Grisha Coleman

School of Arts, Media and Engineering, SenSIP Center,
& School of Electrical, Computer and Energy Engineering
Arizona State University
alex.fink@asu.edu, bmechtley@asu.edu

ABSTRACT

Sonic representations of spaces have emerged as a means to capture and present the activity that conventional representations, such as maps, do not encapsulate. Therefore, to convey the activity information of regions, both large and small, we use sounds and information provided by regional communities in the automated design of soundscapes to re-sonify geographic sound activity. To quantify this community knowledge, we have developed an ontological framework to determine the importance of sound and concepts to one another using acoustic, semantic, and social information. This framework is then used in the automated design of a generative soundscape model purposed to identify and re-sonify sounds that impart relevant information about a geographic region. Furthermore, we are developing a social networking website to facilitate the collection and re-sonification of sounds and data.

1. INTRODUCTION

The ability to understand the activity local to specific geographic regions is limited when presented through maps, directories, and other conventional representations. Even novel interactive representations, such as Photosynth [1] and Google Street View [2], present community information in the form of artifacts (in this instance, images). Exploring geography through sonification, however, presents a method of experiencing a location through sonic events. This concept has been explored in a number of systems that primarily focus on displaying where sounds are recorded and allowing them to be played as recorded [3, 4, 5]. As a primary carrier of information about activity, sound can project information about how people relate to their surrounding environments and what these environments mean in terms of their daily lives. This link between geography and activity is often explored in the context of soundscapes. As acoustic experiences are considered to play a significant role in human ecology, soundscapes may supplement our comprehension of activity in physical environments and geographic spaces as well as our understanding of cultural and anthropological issues [6, 7, 8]. Whether real or imagined, soundscapes have been used to enhance immersive experiences in real and virtual worlds for purposes including music [9], audio-visual production [10], geographic exploration [11, 12, 13], and community understanding [14]. Many previous innovative works in soundscape synthesis address community meaning and aesthetics in interactive systems through the knowledge of a composer, often gained through community presence, interaction, and/or interviews [11, 12, 13, 15]. We seek, however, to create scalable, automated methods of soundscape design, where meaning is defined by communities themselves, drawing on their provision of both sound recordings and community knowledge.

Meaningful re-sonification of activity in a geographic region can be difficult when recorded sounds from that region are either 1) abundant or 2) scarce. Where recordings in a region are few in number, re-sonification itself may be sparse or highly repetitive without the inclusion of relevant sounds from other locations. Conversely, if recordings from a region are plentiful, many sounds may be redundant or uninformative about the area's activity. Both situations may be addressed by classifying and using those sounds that are relevant and important to an area. Traditional classification of sounds within a soundscape (keynote, signal, and soundmark) is primarily focused on their perceptual role to listeners [6, 7]. This classification is area-specific, depending on the perception of sounds as dictated by the meaning and prevalence of sounds in a community. While the identification of important sounds to an area does not provide this classification, it is able to distinguish which sounds convey the relevant activity of a region, a relevance perhaps best determined by that region's own community.

The concept of community-defined importance of sounds has long been held in the auditory field; in [6], Schafer states,

> Acoustic design should never become design control from above. It is rather a matter of the retrieval of a *significant aural culture*, and that is a task for everyone.

This idea also extends beyond the auditory domain; Google's PageRank technology, for example, determines the importance of web pages by considering the number and relative importance of other pages that link to them [16]. The relevance of such pages is then defined by the internet community's own activity. Similarly, the acoustic knowledge and the actions of a community can help to reveal important sounds for the re-sonification of geographic activity.

To work towards revealing this importance, we have developed an ontological framework to link sounds together through acoustic, semantic, and social information. Using acoustic content in conjunction with user-provided tags, our framework relies on the prior knowledge of acoustic and semantic ontologies combined with community-defined social links between sounds

This material is based upon work supported by the National Science Foundation under Grant No. 0504647.

and concepts. By linking concepts and sounds together with the community-provided information, the ontological framework provides a measure of the relevance of sounds (and concepts) to one another. To re-sonify specified locations through the playback of sounds in a database, the ontological framework is used to create a graph-based generative soundscape model. Similar to the use of textual queries to filter a ranked list of important websites, we use location to determine the soundscape model parameters such that geographically relevant sounds play frequently. Consideration of the size (surface area) of locations allows our re-sonification to scale to communities or regions of varying size. Using sounds recorded from these locations and other locations that are deemed important by an area's community, our methodology aims to create meaningful soundscapes reflective of the geographic sound activity in those areas. User-tests to assess our methodology are needed, though informal reviews by users have thus far been generally favorable.

The remainder of the paper is organized as follows. Section 2 describes our ontological framework to link sounds and concepts together, using acoustic, semantic, and social information. The application of this framework to the automated design of a soundscape model for re-sonifying geographic activity is discussed in Section 3. Section 4 then presents a social networking website currently under development that provides for the collection of and classification of sounds; the site features an interactive map using our re-sonification scheme to allow virtual "soundwalks." Finally, preliminary results are given in Section 5, followed by conclusions and discussion of future work in Section 6.

2. ONTOLOGICAL FRAMEWORK

To automatically compose soundscapes from collections of sounds with user-provided descriptions, some notion of similarity between sounds is necessary to determine what sounds may be relevant to a space. For example, if few sounds are recorded in a location, retrieving perceptually similar sounds provides greater diversity in the synthesis process. We calculate such similarity with an ontological framework that links together sounds and concepts, using acoustic similarity between sounds, social information in the form of links between sounds and concepts, and semantic information in the form of conceptual similarity [17]. Using these separate modalities, the ontological framework determines the relevancy of objects (sounds and concepts) to one another, using available links (acoustic, social, or semantic).

The ontological framework consists of an undirected graph (Figure 1), where nodes in the graph represent sounds ($\mathcal{S} = \{s_1, ...s_N\}$) or concepts ($\mathcal{C} = \{c_1, ...c_M\}$). Nodes are connected by weighted links, and a nonnegative link weight connecting nodes i and j is signified by $W(i, j)$. Links of weight zero represent equivalence between the nodes connected by that link, while a link of infinite weight between two nodes is equivalent to no link being present. Given a subset of nodes, $\mathcal{A}$, and query node, q, a posterior distribution from the network can be calculated as follows:

$$P(a \in \mathcal{A}|q) = \frac{e^{-d^*(q,a)}}{\sum_{b\in\mathcal{A}} e^{-d^*(q,b)}}, \tag{1}$$

where $d^*(q, a)$ is the shortest-path distance in the network between nodes q and a, which can be efficiently computed using Dijkstra's algorithm [18]. Note that, in the case of a query node that does not yet exist in the database, such as a new sound or concept, the distances between the query node and all other nodes of its type can be computed on demand. For example, when a new sound is uploaded to the database ($q \in \mathcal{S}$) and the ontological framework returns a distribution over concepts ($\mathcal{A} \subset \mathcal{C}$) as in Figure 1 (a), we can automatically annotate a new sound file with tags suggested by the community based on the audio content. In a similar fashion, a concept query ($q \in \mathcal{C}$) can return a distribution over sounds ($\mathcal{A} \subset \mathcal{S}$) as in Figure 1 (b). In order to use the ontological framework in this fashion, we must set the values for all link weights. A description of the three types of links we use, sound-to-sound, concept-to-concept, and sound-to-concept, follows below.

2.1. Acoustic information: sound-to-sound links

Sound-to-sound weights can be computed by comparing the acoustic content of each sound. This process begins with acoustic feature extraction, where six low-level features are calculated using a frame-based analysis, where we use 40 ms frames with 50% overlap and a Hamming window. Features are calculated either from the time-domain data or the short-time Fourier Transform (STFT) spectrum. The feature trajectory for a sound file is given by $Y_{1:T}^{(1:P)}$ where $Y_t^{(i)}$ is the ith feature value at frame t.

The six features we use include *loudness*, the dB-scaled RMS level over time; temporal sparsity, the ratio of ℓ^∞ and ℓ^1 norms calculated over all short-term RMS levels computed in a one-second interval; spectral sparsity, the ratio of ℓ^∞ and ℓ^1 norms calculated over the STFT magnitude spectrum; bark-weighted *spectral centroid*, a measure of the mean frequency content for a sound frame; transient index, the ℓ^2 norm of the difference of Mel frequency cepstral coefficients (MFCCs) between consecutive frames; and harmonicity, a probabilistic measure of whether or not the STFT spectrum for a given frame exhibits a harmonic frequency structure. For more details on how these features are calculated, see [19]. This feature set was developed to accurately represent a broad range of environmental sounds rather than any specific class of sounds (e.g. speech or music) while also providing an intuitive and minimal set for efficient retrieval of sounds stored in a database. To compare sounds, [20] describes a method of estimating $L(s_i, s_j) = \log P[Y_{1:T}^{(1:P)}(s_i)|\lambda^{(1:P)}(s_j)]$, the log-likelihood that the feature trajectory of sound s_i was generated by the hidden Markov Model (HMM) $\lambda^{(1:P)}(s_j)$ built to approximate the simple feature trends of sound s_j.

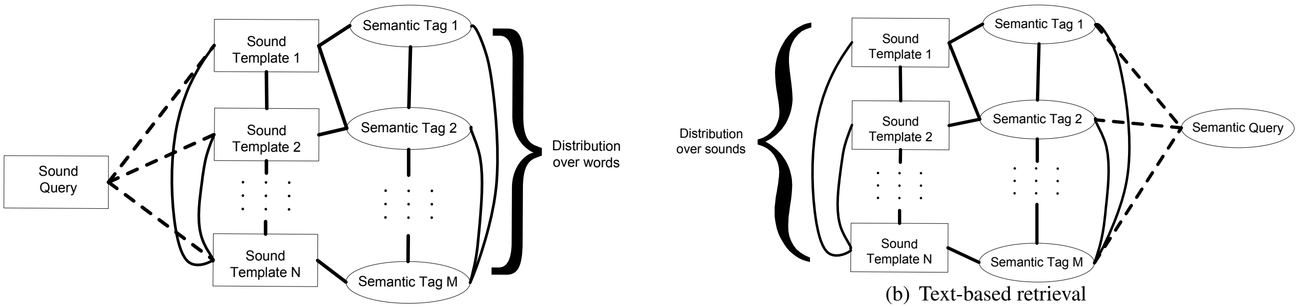
The ontological framework we have defined is an undirected, acyclic graph, which requires weights be *symmetric* ($W(s_i, s_j) = W(s_j, s_i)$) and *nonnegative* ($W(s_i, s_j) \geq 0$). Therefore, we cannot use the log-likelihood $L(s_i, s_j)$ as the link weight between nodes s_i and s_j, because it is not guaranteed to be symmetric and nonnegative. Fortunately, a well known semi-metric that satisfies these properties and approximates the distance between HMMs exists [17, 21]. Using this semi-metric we define the link weight between nodes s_i and s_j as

$$\begin{aligned} W(s_i, s_j) = &\frac{1}{T_i}[L(s_i, s_i) - L(s_i, s_j)] \\ &+ \frac{1}{T_j}[L(s_j, s_j) - L(s_j, s_i)], \end{aligned} \tag{2}$$

where T_i and T_j represent the length of the feature trajectories for sounds s_i and s_j, respectively.

2.2. Semantic information: concept-to-concept links

To calculate concept-to-concept link weights, we use a similarity metric from the WordNet::Similarity library [22]. Specifically,



(a) Annotation

(b) Text-based retrieval

Figure 1: Organization of ontological framework for two common indexing and retrieval tasks. Dashed lines indicate links added at query time. Text-based retrieval use words to search for (unlabeled) sounds. Annotation automatically describes a sound file based on its audio content and provides suggested tags to contributors.

we use the `vector` metric, because it supports the comparison of adjectives and adverbs, which are commonly used to describe sounds. The `vector` metric computes the co-occurrence of two concepts within the collections of words used to describe other concepts (their *glosses*) [22]. For a full review of WordNet similarity, see [23, 22].

By defining $Sim(c_i, c_j)$ as the WordNet similarity between the concepts represented by nodes c_i and c_j, an appropriately scaled link weight between these nodes is

$$W(c_i, c_j) = -\log\left[\frac{Sim(c_i, c_j)}{\max_{k,l} Sim(c_k, c_l)}\right]. \tag{3}$$

2.3. Social information: sound-to-concept links

We quantify the social information connecting sounds and concepts using a $M \times N$ dimensional votes matrix V, with elements V_{ji} equal to the number of users who have tagged sound s_i with concept c_j divided by the total number of users who have tagged sound s_i. By appropriately normalizing the votes matrix, it can be interpreted probabilistically as

$$P(s_i, c_j) = V_{ji}/\sum_k \sum_l V_{kl} \tag{4}$$

$$P(s_i|c_j) = V_{ji}/\sum_k V_{jk} \tag{5}$$

$$P(c_j|s_i) = V_{ji}/\sum_k V_{ki}, \tag{6}$$

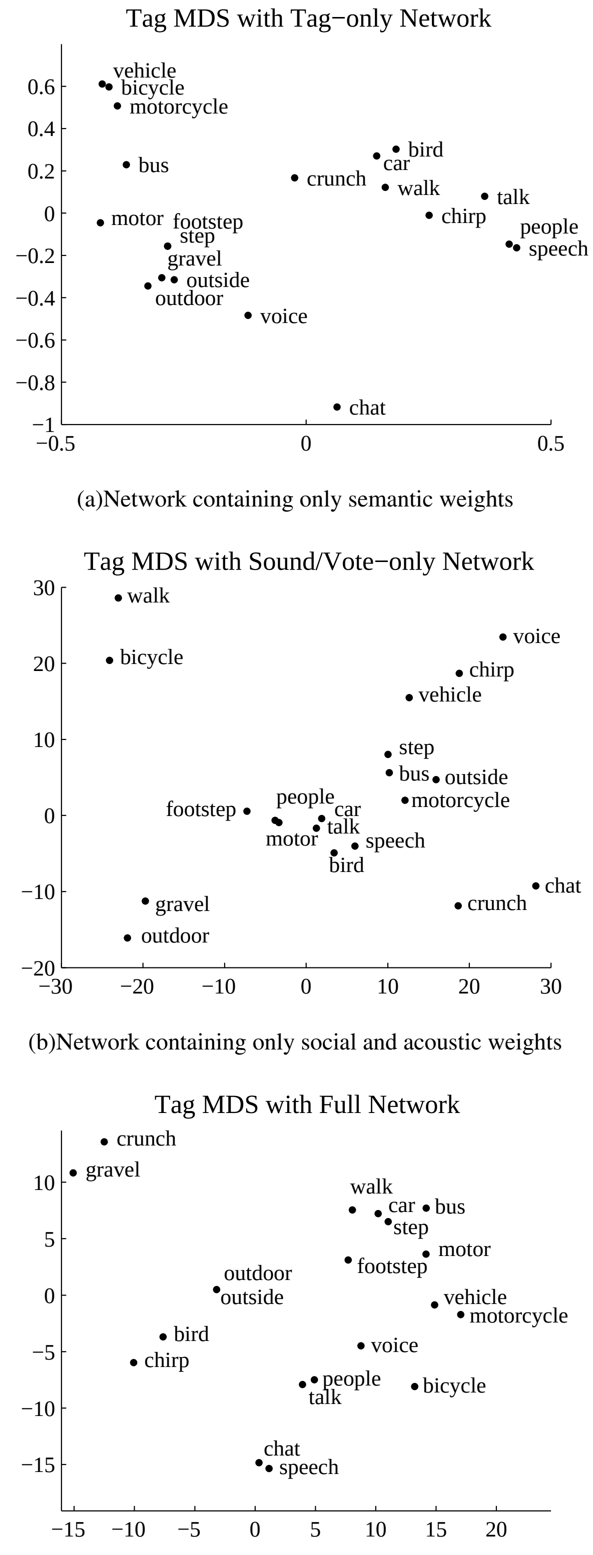
where $P(s_i, c_j)$ is the joint probability between s_i and c_j, $P(s_i|c_j)$ is the conditional probability of sound s_i given concept c_j, and $P(c_j|s_i)$ is defined similarly. Our goal in determining the social link weights connecting sounds and concepts is that the probability distributions output by the ontological framework using (1) are as close as possible to the conditional distributions from the votes matrix in (5) and (6). One way of measuring the distance between probability distributions is the Kullback-Leibler divergence [24]. The link weights between sounds and concepts are then optimized to jointly minimize the Kullback-Leibler divergence between the distributions obtained from the ontological framework and those from the votes matrix, using each sound in the database to obtain a distribution over concepts and each concept in the database to obtain a distribution over sounds. Complete details on this weight optimization process are provided in [17]. Empirically, we have found that a simple approximation of the optimized weight values is to set them to a value inversely related to the joint distribution (4), i.e., $W(s_i, c_j) = -\log P(s_i, c_j)$.

Presently, the votes matrix is obtained using only a simple tagging process. In the future we hope to augment the votes matrix with other types of community activity, such as discussions, rankings, or page navigation paths on a website. Furthermore, sound-to-concept link weights can be set as compositional parameters rather than learned from a "training set" of tags provided by users. For example, sounds can be made equivalent to certain emotional concepts (happy, angry, etc.) through the addition of zero-weight connections between specified sounds and concepts. These emotional connections will then affect the display and soundscape re-synthesis processes discussed in subsequent sections. Similarly, relative scalings of weights between different types of information (e.g., semantic versus acoustic) can be used to explore different relationships amongst the collected sounds.

2.4. Multidimensional scaling

In order to conveniently summarize the social, semantic, and acoustic information contained in the ontological framework for soundscape re-synthesis and visual representation of sound activity on a map, we use multidimensional scaling (MDS) [25]; this embed each sound or concept node in the graph into a low-dimensional space in such a way that retains the distance relationships between nodes. MDS operates on a distance matrix, which is obtained by finding the shortest-path distance between all node pairs using Dijkstra's algorithm.

To provide an example of how the MDS embeddings of our ontological framework represent social, semantic, and acoustic information, Figures 2(a), 2(b), and 2(c) display the two-dimensional MDS for a subset of selected tags. In Figure 2(a) the distance matrix is calculated from an ontological framework containing only tag nodes, i.e., only semantic information, while Figure 2(b) contains both sound and tag nodes but only uses acoustic and social links, excluding concept-to-concept semantic connections. Figure 2(c) shows the MDS that uses all available nodes and links, i.e., acoustic, social, and semantic information. The differences between the absolute scales of the axes in the figures result from the different distance matrices, but by comparing relative tag positions we can see how information is organized in the different frameworks. From Figure 2(a), we can see that natural clusters



(a)Network containing only semantic weights

(b)Network containing only social and acoustic weights

(c)Network containing semantic, social, and acoustic weights

Figure 2: MDS of tags for networks with and without non-semantic information using the `vector` semantic similarity metric.

form from the semantic information, such as {*vehicle*, *bicycle*, *motorcycle*, *bus*}. Similarly, synonyms such as *outside*/*outdoor* are near each other. However, some concepts we might expect to group together do not, e.g., *chat*/*speech*. Similarly, in Figure 2(b), concepts such as *car* and *motor* are close, but *bird* and *chirp* are not.

By including social and acoustic information in the framework, as in Figure 2(c), the concepts organize into clusters that are informed by which concepts sound alike. For example, *chat*/*speech* are now quite near each other. Similar new clusterings can be seen between word pairs such as *gravel*/*crunch* and *bird*/*chirp*. Some clusterings are more vague in the reorganization, such as that which formerly grouped all vehicles, as we are now also capturing information of what concepts typically are heard together or in similar circumstances. This clustering behavior is then considered in our method of soundscape synthesis, described in the following section.

3. SOUNDSCAPE DESIGN

In the ontological framework, the relevance and importance of different sounds to one another is quantified. Using this information, we have developed a method to automatically design graph-based generative soundscape models to re-sonify geographic sound activity. This methodology aims to provide automated soundscape design that 1) is scalable to geographic regions of any size, and 2) meaningfully incorporates community knowledge to address re-sonification when locally obtained sounds are scarce or overly plentiful. Our design and synthesis of soundscapes assumes the availability of a database of sounds with GPS locations and community-provided tags; our process of collecting this data is presented in Section 4.1. Figure 3 displays how the components of our soundscape synthesis system interact, described as follows in detail.

3.1. Markov Transition Networks

Many of the recent approaches to soundscape generation use probabilistic generative models to sonify activity, focusing generally on designing the overall distribution of sounds in the soundscape. By using a stochastic generative model, all generated soundscapes provide different experiences that are consistent in their meaning, yet unique in each instance. In [10], for example, a method of ambiance generation is developed, where short isolated sounds are mixed with longer atmospheric recordings to generate a soundscape described by user-provided textual queries. The addition of the short, isolated sounds to the mix is determined probabilistically such that they most often appear in periods of relative silence in the overall mix. Other methods, such as that of [15], generate a soundscape from the randomized playback of pre-classified sounds where the expected temporal density of different types of sounds is determined through user interaction and design. Recent work [12, 13] uses a manually designed graph-based model where sound sequences are determined stochastically but subject to the set of sequences allowed by the graph's design. This model is particularly useful in modeling representations of complex sources of sound events. Drawing upon this and other work, we use graph-based modeling for soundscape synthesis, but with additional focus on the overall expected temporal density of sounds.

To generate soundscapes for our application, we have chosen to use an emerging compositional structure that we call a Markov Transition Network (MTN), a variation of the models introduced

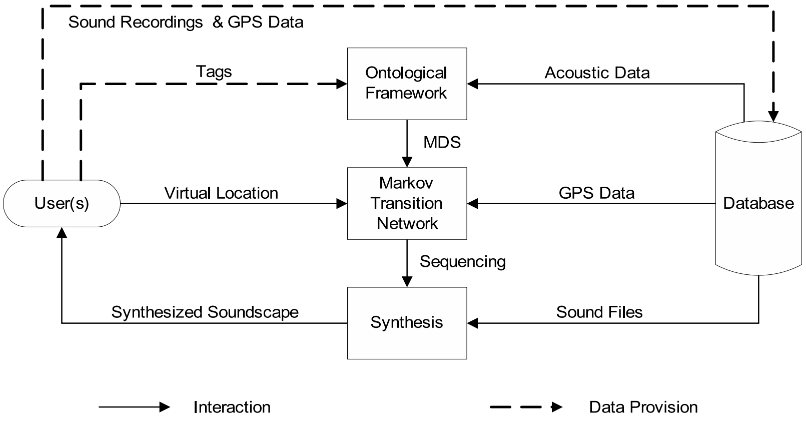


Figure 3: Diagram of the soundscape synthesis system.

in [12, 13, 26]. An MTN is a directed graph with N nodes, with possible directed edges from each node, i, to another node, j, including $j = i$. Figure 4 displays and example MTN. The MTN is used by an *actant process*, $A(t)$, that "travels" to the various nodes of the network, with its behavior dictated by the edges present in the graph. The actant process takes on values from 1 to N, representing the node at which the process is located at a given point in time. Each edge has an associated transition time, $\Delta(i, j)$. When $A(t)$ "enters" node i, the choice of the "next node," j, is determined by an associated probability, $P(i, j)$, and the actant process waits a time of $\Delta(i, j)$ before making the transition. If no edge exists between any two nodes, the associated probability is zero. Given these properties, we note that $A(t)$ is not a Markov process, as transition times depend on the origin and destination nodes, though it is a Semi-Markov process with deterministic transition times. Figure 4 displays an example MTN, with nodes and transition times of edges labeled. (Edge probabilities are omitted for clarity.)

Sound synthesis is performed by the sequenced playback of sounds as determined by the actant process. Sounds in the database are uniquely associated with a node, i, and a duration, $D(i)$. Upon $A(t)$ reaching a new node, the associated sound is played back in full, regardless of the chosen transition time to the next node or length of the following sounds. We presently mix together all sounds being played back into a single soundscape, though we note that a more complex multi-channel scheme could be adopted, and various effects (e.g., reverb) may be applied to individual sounds or the entire mix. Note that multiple actant processes may be active at any time, independently triggering sounds.

Using an MTN, the sequencing of sounds is made random, but it may be limited by the connections made between nodes. If only a single edge is directed from a node, then the sequencing upon the actant process's selection of that node will be temporarily deterministic. However, if all nodes in an MTN are fully connected, the behavior of the actant process becomes less predictable (dependent on the transition probability distributions). By limiting the number of edges connecting nodes, the sequencing determined by actant processes may be made variable, yet confined by the parameters of the network. This is considered in [12, 13], where limited connections are made between clusters of nodes to specify the behavior of complex sources of sound as predictable sequences. We recognize this effect of limiting connections, but we also wish to examine the overall expected properties of the synthesized output. Therefore, we consider the expected temporal density of all available sounds.

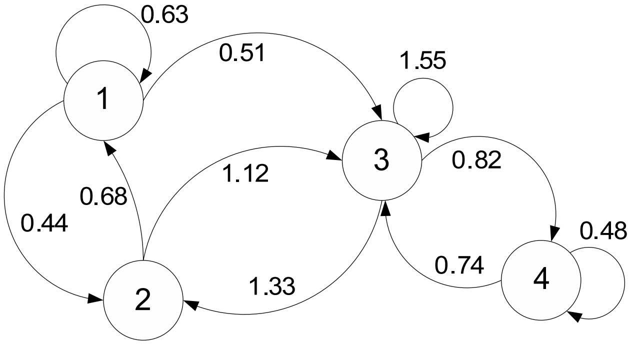


Figure 4: Example MTN for soundscape synthesis. Edges are labeled with transition times. Transition probabilities are not shown.

For a sound i, with an intensity value (this may be any meaningful chosen measure, such as loudness), $V(i)$, we define the expected sum of intensities of any instances of sound i at a given time to be the density, $Density(i)$, given by

$$Density(i) = \frac{D(i)V(i)}{T(i,i)}, \tag{7}$$

where $T(i, j)$ is the expected time for the actant process to travel from node i to node j, including indirect paths. ($T(i, j)$ is then the expected time for the actant process to leave and return to node i). If the actant process travels directly to j, the transition time will simply be the delay, $\Delta(i, j)$, else it will be the delay, $\Delta(i, k)$, to an intermediary node, k, and the time taken to then reach j. Therefore,

$$T(i,j) = \sum_{k=1}^{N} P(i,k)\Delta(i,k) + \sum_{k=1,k\neq j}^{N} P(i,k)T(k,j). \tag{8}$$

Letting Δ, P, and T be $N \times N$ matrices with with elements $\Delta(i, j)$, $P(i, j)$, and $T(i, j)$, respectively, we may express (8) in matrix-vector form as

$$T(:,j) = C + Q_j T(:,j), \tag{9}$$

where $T(:,j)$ is the j^{th} column of T, the i^{th} element of the $N \times 1$ vector, C, is

$$C(i) = \sum_{k=1}^{N} P(i,k)\Delta(i,k), \qquad (10)$$

and Q_j is P with the j^{th} column zeroed out. This gives

$$T(:,j) = (I - Q_j)^{-1}C, \qquad (11)$$

which may be iterated over j.

While this allows us to analyze the density of sounds in a soundscape, for the purpose of design, we seek the ability to specify network parameters to create a desired density of sounds (this density may be determined by the interaction to which the resulting soundscape is applied). As the available sounds and their properties are fixed, specifying the density value of sounds fixes the desired diagonal elements of T. This leaves flexibility in determining the MTN parameters, Δ and P, as there are N equations (one for each of the diagonal elements of T) and up to $2N^2$ unknowns. Therefore, we allow P to be chosen by the designer (human or computer). By choosing P, the connecting edges of the network may be defined, and connections between relevant or logically successive sounds may be reinforced with high probability. The desired densities may then be achieved through the necessary values of Δ.

To determine Δ, we first define $F = E\Phi$, where $E \in \mathbb{R}^{N\times N}$, $\Phi \in \mathbb{R}^{N\times N^2}$, and the i^{th} row of E is given by

$$E(i,:) = e_i + q_i(I - Q_i)^{-1}E_i, \qquad (12)$$

where Q_i is P with the i^{th} column and row removed, q_i is the i^{th} row of P with $P(i,i)$ removed, e^i is the i^{th} row of the size-N identity matrix, E_i is the identity matrix with the i^{th} row removed, and Φ consists of all zeros except for

$$\Phi(i, i + N*(j-1)) = P(i,j), \qquad (13)$$

where i and j are iterated from 1 to N. Finding Δ may then be achieved by solving the quadratic program:

$$\begin{aligned} \text{Minimize} \quad & \|F \cdot \mathbf{vec}(\Delta) - \tau\|_2^2 \\ \text{subject to} \quad & \mathbf{vec}(\Delta) \succeq b \end{aligned}$$

where b is a vector of elements greater than or equal to zero, and $\tau \in \mathbb{R}^N$ is the column vector where the i^{th} element is the value of $T(i,i)$ necessary to achieve the desired density of sound i. The inequality constraint is introduced to allow future extensions where a minimum delay time between certain sounds may be desired. We note that the amount of nontrivial elements of Δ is limited by the edges of the network, and that in some cases the actual set of achieved densities may be the best approximation of densities in a squared error sense.

3.2. Automated Model Design

Using information from the ontological framework and the sounds themselves, we have developed a method of automatically designing an MTN to re-sonify the sound activity of a specified "virtual location" that corresponds to a physical location. Seeking to play the sounds from and relevant to the location, we use our ontological framework to make connections between relevant sounds in the MTN and specify the other parameters such that the expected densities of local sounds are relatively high. By making local sounds dense in the soundscape, they will clearly be heard often, making the available local sounds a key component of the soundscape. As this also implies that the actant process will often travel to local sounds, the creation of edges based on relevancy and importance may aid the actant process in traversing nodes corresponding to sounds relevant to the recorded local sounds. This method is executed as follows.

The edges between vertices are determined by performing a Delaunay triangulation (the dual graph of a Voronoi tessellation) on the sound locations in the previously described two-dimensional MDS. Where a line is drawn between two vertices in the triangulation, edges will be created in both directions; self-connections are not made. The results of Delaunay triangulation on the MDS vary with the placement and clustering of sounds, but it generally connects sounds to those nearby (i.e., sounds deemed relevant by the ontological framework) in the MDS. These connections allow the playback of local sounds to often be preceded and/or succeeded by relevant sounds. The triangulation, however, also makes some connections between sounds considered irrelevant to one another, but the inclusion of such connections can help to ensure that actant processes do not always concentrate near certain nodes when the local sounds are spread in the MDS. Use of Delaunay triangulation also guarantees that every vertex will be connected to at least two other vertices, which can help to prevent repetition.

The desired density of sounds is specified to be inversely related to the distance between the sound's location of recording and the user's virtual location. Currently, we implement this relation as a Gaussian function, referring to the standard deviation as the "listening radius," which sets the size (in surface area) of the region to be explored. The total density of all sounds may be adjusted (so that soundscapes are not overly sparse or dense), perhaps most usefully to a constant value. As described in Section 3.1, specification of the densities determines the values of the transition times, but requires transition probabilities to be provided. The probabilities may be set arbitrarily, but the choice of probability distribution will affect the achievable densities of sounds. Currently, we set the probabilities so that they may further "encourage" the actant process to travel to local sounds. We achieve this by setting the transition probabilities between nodes such that the ratios between the probabilities of edges emanating from a node are equal to the ratios of the desired densities of the nodes toward which they are directed. In practice, we have observed that our current distribution scheme typically provides better actual densities than a uniform distribution.

As this method of soundscape synthesis only requires a virtual location as input when sounds and their corresponding ontological framework are available, it may be applied to various interactions, static or dynamic. Presently, we have created an offline interaction that exactly implements our method of automated design. We have also used this scheme (with sub-optimal calculation of the transition times) in an interactive map, to allow virtual soundwalks, where the network's parameters are periodically updated as the virtual location is changed. This map is also a component of a larger social networking website we are developing that can aid in the collection and tagging of sounds.

4. SOUNDWALKS: AN APPLICATION

We are currently developing a social network website, called "Soundwalks" (http://www.soundwalks.org), to facilitate the collection and re-sonification of sounds and community information.

Through this website, we aim to gather sounds to extensively represent geographic sound activity, and especially where there are too few or too many sounds, utilize user-provided information to determine the importance of individual sounds in the re-sonification of geographic sound activity. On the website, users can upload recorded sounds along with GPS data and provide tags to any sounds. Information, including plots of acoustic features over time, for individual sounds is availabe to view. The site also features an interactive map, similar to systems such as [3, 4, 5, 27], in which the location of sound recordings are marked with icons. With these icons, users can listen to the recordings or retrieve relevant data about them. In addition to allowing users to inspect and playback individual sounds, our map has a "virtual soundwalk" mode that allows users to listen to synthesized soundscapes by moving a token across the map. The following subsections describe the major components of the website.

4.1. Sound Capture and Organization

Presently, sounds in our database have been gathered via mobile sound recording equipment used in conjunction with GPS recording devices. Recordings, which range from short transient events to minute-long soundscapes, consist of both selective recordings and those extracted by automated event segmentation from continuous recordings [19]. An interface allows the uploading of the sounds and GPS data from real soundwalks. After uploading, acoustic feature analysis is performed on the sounds, and they are added to the database and ontological framework. To use existing technology to make the capture and uploading of soundwalks easier and more accessible, we are currently developing a mobile phone application for concurrently recording sounds, GPS data, and any other relevant information (e.g., time).

Once a soundwalk is uploaded and analyzed, users may see a list of sounds from the soundwalk and a small map showing where the sounds were recorded. Each sound may be individually inspected by clicking to navigate to a page that presents all current information about the sound. This includes typical computer-file relevant data (e.g., size, creation time) but additionally displays plots of acoustic feature data and a tag cloud (to which the user may contribute).

4.2. Interaction

An interactive audiovisual display, in the form of a map, is the integral component of our application. This map provides a geographic view of the recordings, displayed as icons at their location of recording on the map. To provide a visual cue of sound content, each icon is colored by mapping its location in the two-dimensional ontological framework MDS to a hue-saturation space. Similarly, all tags on the website are colored by their location in the MDS.

As a user navigates through the map, sound dots within a certain distance threshold, as determined by the map zoom level, will merge together as clusters represented as colored dots. The color and the radius of the cluster are dynamically adjusted by calculating the mean of colors and the number of the child sounds, respectively. When the user clicks on a sound icon, an information window will appear, displaying tags and other information about the recording along with an option to play the sound.

Users may also navigate the map using a virtual soundwalk mode, "scrubbing" a virtual token across the map, creating a virtual soundscape. The soundwalk mode has a variable "listening radius" that may be thought of as the radius of a circle that contains the sounds most expected to be heard. It is effectively the size of the area considered in creating the soundscape. The listening radius may be varied from small to large so as to create soundscapes that range from simulating observable soundscapes at specified locations to providing sonic summaries of large geographic regions. The soundscape is created from an automatically generated MTN (as specified in Section 3), using a single actant process. Periodic updates of the network parameters are made to adapt to the user's movement. The actant process (which is initialized to the sound recorded nearest the virtual location) functions continuously, using the MTN as it is updated. A screenshot of the interactive map (with an open information window) in the virtual soundwalk mode appears in Figure 5.

5. RESULTS

The described system has thus far been informally reviewed by select users and the authors with generally favorable assessments. Our database is presently sparsely populated with sounds recorded in a university environment particularly rich with human activity, traffic (both ground and air), and birds. Using our interaction, we have noted that in exploring areas with few recordings, the inclusion of sounds from elsewhere has provided a soundscape we believe to be indicative of the areas with which we are familiar. The most frequent problem we have observed is the inclusion of keynote sounds (e.g., the beep of a light rail car, or the cheer of a stadium's crowd) in inappropriate areas. Additionally, some sounds are played too frequently. More sounds and users are needed to thoroughly assess how our method of soundscape design can perform across differing communities and with a dense collection of sounds, but present results are promising. Formal listening tests are planned to assess the quality of our synthesized soundscapes, both in comparison to actual recorded soundscapes and in terms of community knowledge.

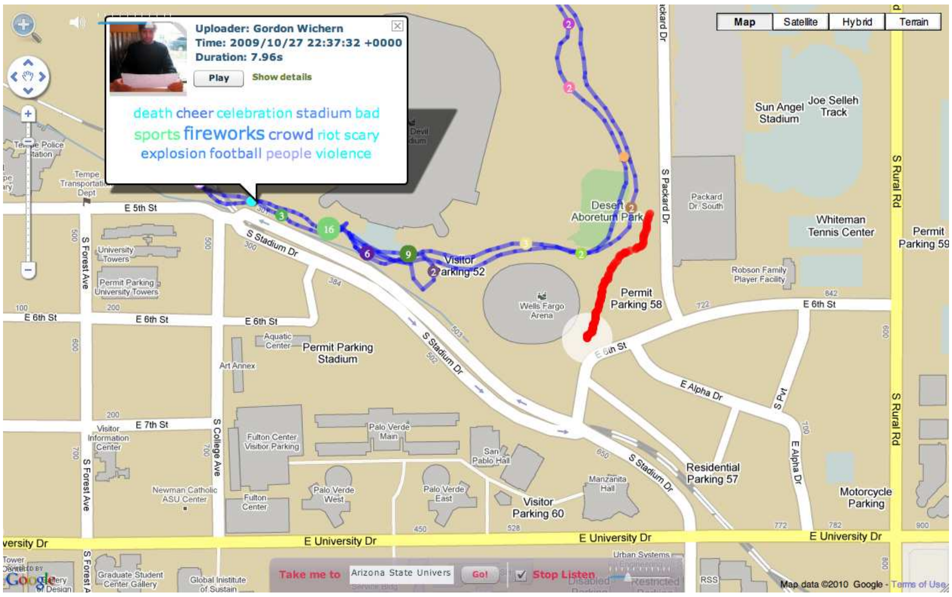
6. CONCLUSIONS AND FUTURE WORK

We have presented a system for the automated re-sonification of geographic sound activity. This re-sonification takes advantage of an ontological framework we have developed that uses acoustic, semantic, and social information that is largely defined by the community. Additionally, we have developed a website application that implements our system for soundscape synthesis. Listening tests and user studies are needed to fully assess our methodology and application.

Areas of future research include use of time data (to provide sound summaries of different points in time) and possible personalization of soundscapes by tracking user activity or preferences. Note that given our current soundscape design method, if certain sounds are precluded (filtered out by time or other data), the MTN may be easily re-designed. Other possible extensions include application of spatialization/reverb effects to enhance the "sense of place" and separate synthesis techniques for different sound types (keynotes, signals, soundmarks), possibly classified by users. Also, we consider the possibility of using geographic-specific information from other sources (such as directory listings or business review sites) to supplement our evaluation of the activity in a space.

7. REFERENCES

[1] Microsoft Live Labs, "Photosynth," http://livelabs.com/photosynth/.



Uploader: Gordon Wichern
Time: 2009/10/27 22:37:32 +0000
Duration: 7.96s

Figure 5: A screenshot of the interactive Soundwalks map in the virtual soundwalk mode.

[2] Google Inc., "Street View," http://maps.google.com/help/maps/streetview/.

[3] Universitat Pompeu Fabra Music Technology Group, "The Freesound Project," http://www.freesound.org/.

[4] Open Sound New Orleans, "Open sound new orleans," http://www.opensoundneworleans.com/core/.

[5] Sound Around You, "Sound around you," http://soundaroundyou.com/.

[6] R. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World*. Rochester, VT: Destiny Books, 1977.

[7] B. Truax, *Acoustic Communication*. Norwood, NJ: Ablex Publishing, 1984.

[8] S. Feld, "Waterfalls of song: an acoustemology of place resounding in bosavi, papua new guinea," in *Senses of Place*, S. Feld and K. Basso, Eds. Santa Fe, NM, USA: School of American Research Press, 1996, pp. 91–136.

[9] A. Misra, P. R. Cook, and G. Wang, "Musical tapestry: Recomposing natural sounds," in *Proceedings of the International Computer Music Conference (ICMC)*, New Orleans, LA, USA, 2006.

[10] P. Cano, L. Fabig, F. Gouyon, M. Koppenberger, A. Loscos, and A. Barbosa, "Semi-automatic ambiance generation," in *Proceedings of the International Conference of Digital Audio Effeccts (DAFx04)*, Naples, Italy, 2004.

[11] S. Serafin, "Sound design to enhance presence in photorealistic virtual reality," in *Proceedings of the 2004 International Conference on Auditory Display*, Sydney, 2004.

[12] A. Valle, M. Schirosa, and V. Lombardo, "A framework for soundscape analysis and re-synthesis," in *Proceedings of the Sound and Music Computing Conference*, Porto, Portugal, 2009.

[13] A. Valle, V. Lombardo, and M. Schirosa, "A graph-based system for the dynamic generation of soundscapes," in *Proceedings of the 15th International Conference on Auditory Display*, Copenhagen, 2009, pp. 217–224.

[14] I. McGregor, A. Crerar, D. Benyon, and C. Macaulay, "Soundfields and soundscapes: Reifying auditory communities," in *Proceedings of the 2002 International Conference on Auditory Display*, Kyoto, 2002.

[15] D. Birchfield, N. Mattar, and H. Sundaram, "Design of a generative model for soundscape creation," in *Proceedings of the International Computer Music Conference (ICMC)*, Barcelona, Spain, 2005.

[16] Google Inc., "Corporate Information - Technology Overview," http://www.google.com/corporate/tech.html.

[17] G. Wichern, H. Thornburg, and A. Spanias, "Unifying semantic and content-based approaches for retrieval of environmental sounds," in *Proceedings of the IEEE Wokshop on the Applications of Signal Processing to Audio and Acoustic (WASPAA)*, New Paltz, NY, 2009.

[18] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, 2nd ed. Cambridge, MA: MIT Press and McGraw-Hill, 2001.

[19] G. Wichern, H. Thornburg, B. Mechtley, A. Fink, A. Spanias, and K. Tu, "Robust multi-feature segmentation and indexing for natural sound environments," in *IEEE CBMI*, Bordeaux, France, July 2007.

[20] G. Wichern, J. Xue, H. Thornburg, and A. Spanias, "Distortion-aware query by example for environmental sounds," in *Proceedings of the IEEE Wokshop on the Applications of Signal Processing to Audio and Acoustic (WASPAA)*, New Paltz, NY, 2007.

[21] B. H. Huang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Tech. Journal*, vol. 64, no. 2, pp. 1251–1270, 1985.

[22] T. Pederson, S. Patwardhan, and J. Michelizzi, "WordNet::Similarity - Measuring the Relatedness of Concepts," in *AAAI-04*. Cambridge, MA: AAAI Press, 2004, pp. 1024–1025.

[23] A. Budanitsky and G. Hirst, "Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures." in *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittburgh, PA, 2001.

[24] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[25] J. Kruskal and M. Wish, *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications, 1978.

[26] A. Valle and V. Lombardo, "A two-level method to control granular synthesis," in *Proc. of the 14-th Colloquium on Musical Informatics*, 2003, pp. 136–140.

[27] Wild Sanctuary, Inc., "Wild sanctuary," http://www.wildsanctuary.com/.

AWESOME - A TOOL FOR SIMULATING SOUND ENVIRONMENTS

Mats Liljedahl

Interactive Institute,
Acusticum 4
SE-941 28 Piteå, Sweden
mats.liljedahl@tii.se

ABSTRACT

Sounds are (almost) always heard and perceived as parts of greater contexts. How we hear a sound depends on things like other sounds present, acoustic properties of the place where the sound is heard, the distance and direction to the sound source etc. Moreover, if the sound bear any meaning to us or not and what the meaning is, if any, depends largely on the listener's interpretation of the sound, based on memories, previous experiences etc.

When working with the design of sounds for all sorts of applications, it is crucial to not only evaluate the sound isolated in the design environment, but to also test the sound in possible greater contexts where it will be used and heard. One way to do this is to sonically simulate one or more environments and use these simulations as contexts to test designed sounds against.

In this paper we report on a project in which we have developed a system for simulating the sounding dimension of physical environments. The system consists of a software application, a 5.1 surround sound system and a set of guidelines and methods for use. We also report on a first test of the system and the results from this test.

1. INTRODUCTION

Human auditory perception has a remarkable and sophisticated ability to identify sounds [1, 2]. Articles in the field of sound design includes a wide range of topics, such as aesthetics, cognition, intuition, design patterns, sound classification etc. [3, 4]. Based on this it can be argued that sound design is an inherently complex task that demands skills in a host of different areas and fields of competence. As a designer of sound for auditory displays of various types you need to have at least basic understanding of things as seemingly disparate as auditory perception, aesthetics of sound and music, acoustics and psychoacoustics together with experience in using digital media editing software and hardware. The field is complex, a large number of competencies are involved in even moderate design efforts and there are vast numbers of possible applications ranging from the design of sounds for kitchen appliances via car blinkers to mobile phones and software for personal computers. Together this suggests there be a large number of easy-to-use, yet powerful professional tools and methods available to support the designers in their roles as creators of sounds that users are, in many cases, likely to hear many times every day. Unfortunately these tools are still, to a large degree, yet to be created.

Auditory perception is based on what is sometimes described as two streams, one in each ear [4, 5]. Our auditory perception is also to a large degree associative, individual and a matter of taste [6]. When designing auditory displays it is therefore important to test different solutions with different groups of potential users to secure the design. Also, when designing sounds that are to be played and heard together with other sounds in a larger auditory context, it becomes important to be able to test the designed sounds in that context or at least in a trustworthy simulation of it [7]. If a sound can not be heard over the background noise, it does not matter how well designed it is.

When designing the individual sounds for auditory displays of various types, there are tools available in the form of recording equipment, sound synthesizing, sound editing systems etc. Often these systems have their origins among artists and in the music industry or in radio and TV broadcasting. There is also an emerging and growing set of design methods available influenced and inspired by theater, film, industrial design and game development just to name a few [8, 9]. Still, there are plenty of research and development to be done to be able to compare what is available in the field of sound design to what is available in, for example, the graphical field. Then again, when it comes to testing the designed individual sounds in their larger context for effects such as masking, repetition and position in space, the tools are largely missing.

Visual artists and designers have always been sketching and graphic designers have a large palette of tools available for trying out and testing ideas. Lately, with the ability to easily record sound and with the advent of computers and music synthesizers, also musicians have been able to sketch and test musical ideas, from small fragments to large orchestral pieces. But for the industrial designer working with the design of sounds as part of a solution, the situation is radically different. There are very few, if any, systems available from the shelf that can be used, without great investments in training, equipment etc. to sketch out a whole soundscape and to test ideas and solutions. The systems available today are most often created for other purposes, such as music creation and editing or for general sound editing. From a more general designers perspective, these systems demand of the designer to put great efforts in learning the systems in order to be productive. One can suspect that often, these learning efforts are not perceived as corresponding to the benefits of the outcomes.

The tools available today for general sound design can broadly be put into the following categories:

- Tools for music production and DJ'ing, etc.
  - Cubase [10]
  - Logic [11]

  - Ableton Live [12]
  - Traktor Pro [13]
- General tools for audio editing, etc.
  - Audacity [14]
  - Pro Tools [15]
- Software for sound syntesis, etc.
  - Native Instrument's Reactor [16]
  - SuperCollider [17]
  - Max MSP [18]

The products listed above are all very competent with immense possibilities to create, edit, apply DSP effects, mix, etc. sound for numerous situations and contexts. Here, the problem is that these relatively large software systems, with few exceptions, are designed for audio professionals. The functionality of most of them can be greatly expanded by the use of various types of plug-ins and there is an ever-growing number of plug-ins for room simulation, sample libraries, software instruments and sound effects. For the experienced and professional users these softwares mean tremendous opportunities to perform all sorts of sound design tasks. The problem we want to highlight here is that the systems are very demanding and requires special training and long-time experience in order to be productive. As sound and audition are becoming more and more important in the design of today's products and services, it is an increasing problem that the tools for sound design are to a large degree inaccessible to the general, non-expert audience. One special problem sound designers are faced with is to sketch out and evaluate sound design ideas as parts of greater wholes, to be able to judge if a designed sound will work as intended together with other sounds in the target environment.

This paper describes a system that aims at being accessible and useable also for non-expert users, enabling designers without special audio training or experience to sketch, try and evaluate sounding ideas.

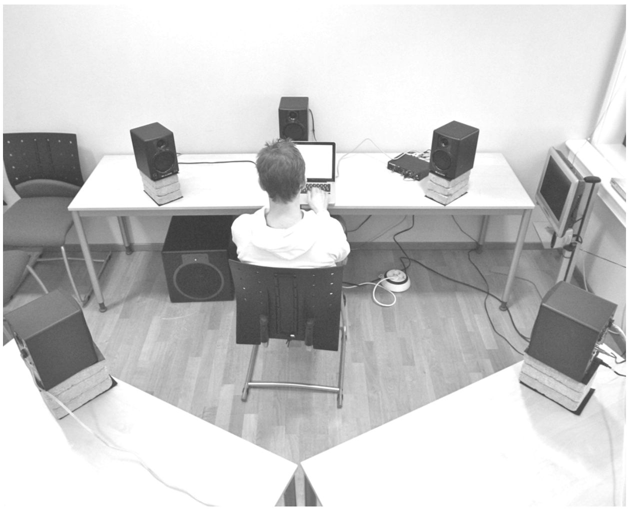


Figure 1. A typical Awesome workstation including computer, audio interface and 5.1 surround sound system.

2. AWESOME – AUDIO WORK ENVIRONMENT SIMULATION MACHINE

AWESOME is a system used for sketching and testing sound design solutions in simulated sound environments (figure 1). The environments are simulated as 2D-spaces and sonically rendered through a standard 5.1 surround sound system. Using the Awesome system you can build sounding replicas of target environments, import designed sounds as audio files to these replicas, position the sounds in 2D-space and move the system's listener object's position and orientation in the environment. In this way the sound design can be evaluated and tested with respect to the overall target environment in which the designed sound or sounds are to be used. This in turn can be used for decision making, to demonstrate ideas and to test design solutions. In focus when designing the system has been a user with only basic knowledge in and experience of traditional sound production equipment. Emphasis has been put on embedding knowledge about general sound design issues into the system in order to make it accessible to users without specialized training and knowledge in more traditional sound design tools. The aim is to create a system that empowers users in a wide range of professional and training situations with a need for sketching and testing sounding ideas.

Traditionally, audio and music editing softwares have emphasized sounds timely dimension. Often these systems are based around timelines that allow the user to organize layers of sound in time. The Awesome system also has a timeline and gives the users similar basic functionality to organize layers of sound in time. But in addition, the system also gives the users functionality to organize sounds in simulated, virtual spaces and to move a listening avatar around in these spaces. In this way, not only the timely aspect of sound but also the space aspect are opened up and can be modeled and worked with.

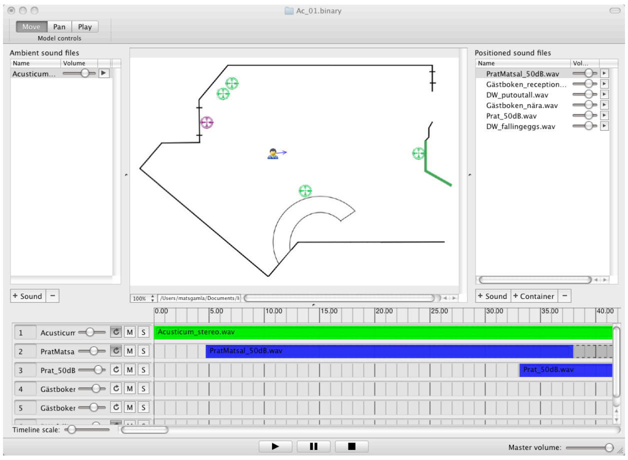


Figure 2. The Awesome main project window.

2.1. The AWESOME system

The total Awesome system consists of five major parts:

- A Mac OSX-based software application.
- A multi-channel audio interface.
- A 5.1 surround sound loudspeaker system.
- A sound library with a set of audio files for a number of typical use-scenarios.

- A set of recommendations, methods and guidelines. The recommendations regard the choice of equipment, for example the surround sound system, audio interface and recording equipment. The set of methods and guidelines includes basic instructions for recording techniques using the recommended equipment, instructions for calibrating the sound system and a general methodology for using the system for a number of typical use-scenarios.

The Awesome software application is built around projects. An Awesome project consists of the following three parts (figure 2):

- The sound library in which the audio files used for the sonic simulation are organized. The library is divided in two parts (figure 3). One part for point sound sources with distinct location in the rendered 2D-soundspace and one part for non-positioned, ambient sounds.
- The graphical model of the simulated environment. The model is represented by a two-dimensional top-view image of the environment (figure 4). One example of a typical model is the blue print of a room or the picture of a car from above. In the model view, sound sources are positioned, animated and moved in 2D space. The model also contains the single listener object that can be moved and rotated within the model and in relation to the individual sound sources.
- The timeline. Sound sources can be positioned in time on the timeline, that is, given a start-time and duration (figure 5). The timeline can be played back creating a pre-defined sequence of sound events.

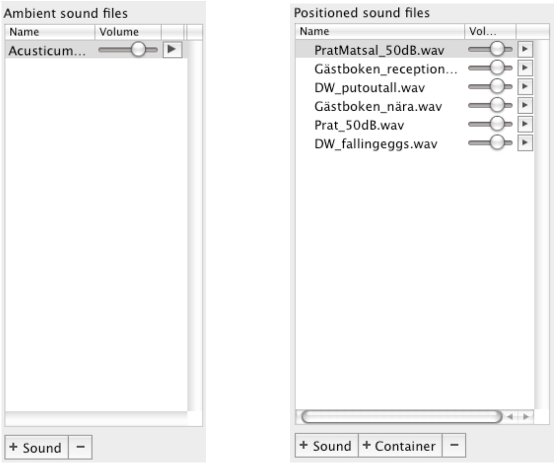
Figure 2 shows the software application's main window with the model view in the centre surrounded by the ambient and the positioned audio file libraries and with the timeline below. The system itself does not synthesize any sounds, instead all sounds used to construct the simulated sound environment are imported to an Awesome project as standard audio files in WAV, AIFF or MP3 format. The resulting soundscapes are rendered through the standard 5.1 surround sound system.

2.2. The audio file library

The audio file library is divided into two parts, one part for ambient, non-positioned sounds and one part for positioned sound objects. Examples of ambient, non-positioned sounds are background sounds such as wind, sounds from ventilation or the sounds from airspeed and tires on asphalt in a car. Examples of positioned sound objects are a bird singing, a human voice, the hum from a refrigerator or the ticking from the blinker relay in a car. In the library, the basic volume of each sound file can be adjusted in order to balance the sounds relative volumes. Each audio file in the library can also be individually played directly from the library window as a complement to playing it from the model and the timeline. (See figure 3.)

The positioned sound sources must be mono audio files. A sound source is positioned in the model by dragging it from the library and dropping it on the model. Once positioned in the model, the playback of the sound object will have a rendered position in the surround sound system relative to the listener object's position and orientation (rotation).

The positioned sound files in the library can be collected in ”containers” for easy comparison between files. Only one audio file in a container is active at a time. Containers can be dragged from the library to the model just as single audio files, but only the active audio file in the container will be played as the container is triggered.



Ambient sound files

Name	Volume
Acusticum...	▶

Positioned sound files

Name	Vol...
PratMatsal_50dB.wav	▶
Gästboken_reception...	▶
DW_putoutall.wav	▶
Gästboken_nära.wav	▶
Prat_50dB.wav	▶
DW_fallingeggs.wav	▶

Figure 3. The ambient and positioned audio file libraries. Note that each audio file has its own volume control and Play-button.

The ambient sound sources can be either mono or multi-channel audio files. The user can assign the channels of a sound file to channels in the rendering 5.1-surround sound system. A stereo-file can for example be assigned to the channels of the surround sound system so that the left channel of the stereo file is routed to both the front left and the rear left speakers of the surround sound system and the right channel of the stereo file is routed both to the front right and rear right speakers.

2.3. The model view

The centre of the Awesome system is the ”model view” (figure 4) where the positioned sound objects are organized in space. The model view also contains the single listener object that can be moved and rotated in relation to the model and the positioned sound objects. The base for the model view is a two-dimensional top-view image of the environment to simulate. The scale of the model view can be set. When a positioned sound object is moved in the model view, the change in position is reflected in the sounds position in the soundscape, rendered through the surround sound system. The same is true when the listener object is either moved or rotated in the model view. The ambient sounds do not have positions and are therefore not affected by the listener object being moved or rotated.

Positioned sounds are placed in the model view by dragging them from the audio file library and dropping them on the model view. Once in the model view, the sound objects can be moved and re-positioned as appropriate. Sound objects can also be animated by assigning an end position relative to the start position. When triggered, the sound's position in the surround sound system will pan from the start position to the end position during the duration of the sound.

Individual sound objects can be triggered from the model view by clicking the sound object's symbol.

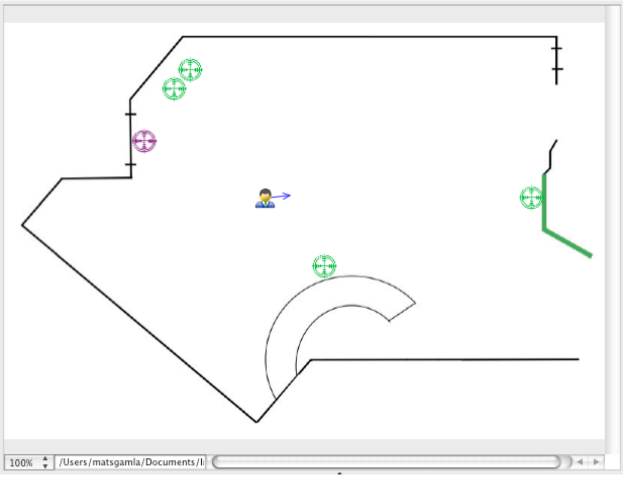


Figure 4. The model view. The arrow on the listener object denotes the direction of the listener in the environment.

2.4. The timeline

Sound objects can be placed on the systems timeline (figure 5) and triggered in sequence as the timeline is played. The timeline consists of a time ruler and a number of ”tracks”. Each track corresponds either to an entry in the ambient sound library or to a positioned sound object placed in the model. Ambient sounds are placed on the timeline by dragging an object from the library and dropping it on the timeline. Positioned sound objects are placed on the timeline by dragging them from the model view and dropping them on the timeline. Individual tracks can be muted, solo’ed or looped on the timeline. Each track also has a volume control for further adjustment of the volume.

Markers can be placed on the timeline and the current position can be moved to a marker by activating a corresponding key on the keyboard. The timeline region between two markers can also be looped.

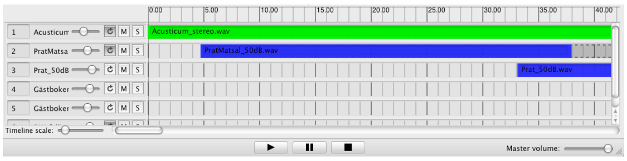


Figure 5. The timeline.

2.5. Built-in acoustics simulation functionality

In the model view, the listener object can be moved and rotated in order to simulate listening from different positions in the real environment. To create an as realistic simulation as possible, given the technical constraints and limitations of the system, two global room characteristic properties can be set on a per-project basis.

- The rolloff-factor. This factor controls how much a positioned sound is attenuated as a function of the distance to the listener object. This factor depends on the scale of the model view and the acoustic characteristics of the simulated environment.
- The reverberation of the room. The system contains a simple yet effective reverberation unit. For each sound object, the balance between the signal affected by the reverb unit (”wet signal”) and the unaffected (”dry”) signal is modulated as a function of the distance between the positioned sound object and the listener object. In a real acoustic environment, the influence of reflections of a sound from walls and other surfaces increases as the distance between the listener and a sound source increases. The Awesome application mimics this effect so that the amount of reverb (wet signal) increases and the amount of direct (dry) signal decreases when the listener object’s distance to a positioned sound source increases.

The relative volumes of the sound sources (audio files) can be adjusted in order to create a realistic balance between the individual sounds comprising the total simulated sound environment.

2.6. Surround Sound System

There are of course several optional methods to arrange the sound playback functionality. The two main options are loudspeakers or headphones. So far we have used a 5.1 mid-range surround sound loudspeaker system exclusively. The system consists of 5 M-Audio Studiophile AV 40 speakers [19] and a M-Audio SBX10 subwoofer [20]. The following motivates the choice of playback solution:

- Loudness. To be able to establish a level of loudness in the simulations with a perceived correspondance to the loudness in the real environment, we want to be able to measure the sound pressure level with a dB-meter. This is not possible when using headphones, which leaves us with the loudspeaker option.
- Collaboration. The aim is to create a system that can be used by groups to collaborate on sound design issues and decision-making. We believe this is easier done with a loudspeaker system rather than with headphones.
- Realism. The psycho-acoustic models available to us does not make it possible to render sound positions using headphones in such a way that “from the front” and “from the rear” are clearly distinguishable. Using a loudspeaker surround sound system this rendering works better, adding to the total realism.

In addition to this we also wanted to keep the cost for the loudspeaker system at a moderate level in order to make the system as accessible as possible for the target groups. The loudspeaker system used and recommended is therefore a mid-range price system.

2.7. Methodology

The intended use of the Awesome system is to sketch, test and evaluate sound design solutions when designing technical systems of various types. A typical Awesome use case scenario can be divided into the following three main parts:

1. Designing, recording and/or synthesizing sounds for both the system under development and other, typical sounds in that systems target environment. These sounds are stored in standard audio files and are used in the following steps

to build the simulated sound environment. Note that Awesome is not intended for this part. Instead traditional recording, synthesizing and editing tools are used.

2. A new Awesome project is created and the audio files from the previous step are imported into this new project. The sounds for the system under development and the sounds for the environment surrounding this system is positioned in the project model view, relative volumes are adjusted, the simulated environments basic acoustic parameters are set in the project and the total soundscape is simulated as closely as possible, given the skill level of the user and the technical constraints of the Awesome system.
3. The sounds for the system under development are tested and evaluated in the context of the total soundscape simulation. Different design solutions can be compared, decisions made and the most appropriate solution selected.

Note that the Awesome system is not intended for the basic design of sounds and the creation of the corresponding audio files in the first part above. Instead the system is used for parts two and three to create the experience and simulation of the whole soundscape including both the designed sounds and sounds of the environment in which they are going to be used.

The development of the Awesome software application is paralleled with the development of a methodology to help and support users to recreate and simulate the sounding aspects of physical environments as realistic as possible.

3. VERIFYING TEST

A first test to verify the system has been conducted. Six test subjects were asked to use the system to recreate the sounding dimension of two physical environments, the interior of a car and the reception of an office building. In focus for this test were the test subjects subjectively perceived experiences of the two environments and how well the Awesome system let them recreate this experience. It should be noted that this time, we were not interested in an objectively measured similarity between the real environment and the replica simulated by the system, but instead the subjectively perceived similarities and differences.

Prior to the test, a number of characteristic sounds in each of the two environments were identified and recorded by the test leaders. In the tests, these recordings were then used by the test subjects as building blocks when recreating the experienced soundscapes of the two environments.

Figure 6. The interior of the car used for tests.

The first environment was the interior of a car driving with the speed of 70 km/h. Apart from the ambient noise from tires, wind, motor and power transmission, the sounds from the horn, the blinker relay, the wipers and the radio were simulated. The second simulated environment was the reception of an office building with people talking, telephones ringing, the sounds from the entrance to the in-house restaurant and the sounds from two interactive applications. The test subjects where sound engineer, music composition and studio musician students at the Department of Music and Media, Luleå Technical University in Piteå, Sweden. Unfortunately only male students volunteered for the test sessions.

Figure 7. The reception used in the tests.

Each test sessions was divided in two parts, one for each environment to simulate. Each simulation sessions was also divided into two parts. First, one of the target environments was experienced firsthand. The test leader and the test subject visited the environment (car and reception respectively) and listened to the predefined characteristic sound sources and the overall ambience of the environment. The test subjects were asked to pay special attention to and try to memorize the following aspects of the environments.

- The overall loudness of the environment.
- The loudness of the individual sound sources relative to the whole.
- How the experience of the sounds depend on the test subjects position and orientation in the environment.
- The positions of the sounds in the environment.

Immediately following this firsthand experience, the test subjects were put in front of the Awesome software application with the task to simulate the soundscape of the environment as close as possible based on the experience and memories from the firsthand experience. In half of the test sessions, the car-case was done first followed by the reception case and the reception was done first and the car last in the rest of the sessions.

Working with the simulations was documented in writing by the test leader and followed the same basic structure in both cases.

1. The test persons where presented an Awesome project with a library of pre-imported audio files recorded in the actual environment as a start. The library contained one ambient

audio file and a number of audio files for positioned sound sources. The project also had a model of the environment set up with basic settings. No sounds had been added to the model view or to the timeline prior to the test sessions and nothing else in the project file was pre-set or manipulated in advance.

2. In order to establish a basic sound pressure level the test subjects were instructed to start with one of the audio files and adjust the volume of that sound so the perceived loudness of the simulation matched the memory of the loudness in the real environment. At this point no equipment for measuring the sound pressure was used, the sound volume adjustment was based solely on the test subject's memory of the perceived loudness.
3. The sound pressure of the simulated ambient sound was measured using a dB-meter. The test leader recorded the result in writing.
4. Next, the positioned sound sources were placed in the model view. The volumes of the audio files were adjusted to match the loudness of the sound in the real environment as perceived and remembered by the test subject.
5. The acoustic characteristics of the simulation were adjusted to match the characteristics of the real environment as close as possible, given the technical limitations of the system.
6. A small questionnaire with seven statements with accompanying seven grade Likert-scales was filled out. The test subjects were asked to relate to what extent the experiences of the simulations made with the Awesome system corresponded to the experiences of the real environments. The aspects asked for were: the total loudness of the environments; the loudness of the individual sound sources in relation to the whole soundscape; the acoustics of the environments; self-movement and the experience of listening from different positions in the environment; aspects of simulating sequences of sounds and finally the total experience of the simulation compared to the real environment.

The same structure was then repeated for the second simulation session. The test subjects were not given any restrictions in time to complete the simulations, but were instructed to take the time needed. The subjects worked with each simulation between 10 and 15 minutes to be satisfied and the simulation considered to be as close to the original as the system allowed. Finally, a small, semi-structured interview with the test subject was carried out led by the test leader. In this interview, the test subjects were given the opportunity to express complementary thoughts and ideas not covered by the more formal questionnaire.

Due to a blizzard on one of the days for the test sessions, the sessions for three of the subjects had to follow a slightly different route than initially planned. The blizzard made driving the car impossible, therefore these three test subjects did the reception case on one day and then came back two days later to do the car case. This was not deemed to affect the test and the results negatively.

4. RESULTS

On an overall level, the results show that the Awesome system is already in this stage working relatively well, allowing relatively trustworthy simulations of the two rather different environments used in the test. In the questionnaires, the test subjects were asked to rate how well the experience of the simulation as a whole corresponded to the experience of the real environment. For the car-case one subject rated the similarity 7 on the seven-grade Likert scale, three subjects rated it 5 and two rated it 4, resulting in a mean value of 5.0. In the reception case, four subjects rated the overall similarity between the simulated and the real environment 5 on the seven-grade Likert-scale and two rated it 4, resulting in a mean value of 4.7. In the interviews, a majority of the subjects expressed the opinion that the quality of the simulations was better than expected and that the similarity between the simulated and the real environments was also better than expected. This is especially true in the smaller and more restricted car-case.

Two weak aspects of the system were identified. The first is the simulation of room acoustics and reverberation. The second is related to sound pressure, loudness, frequency response and the general sound quality of the system used to finally render the soundscape. To create a sense of acoustics and reverberation, the built-in reverb unit of the OpenAL implementation in Mac OS X was used. The car-case did not present a problem in this aspect, since no reverb or other acoustic simulation was considered necessary to simulate the acoustically limited and dry car environment. Instead the acoustics built in the recordings made in the car was in this case sufficient. In the reception-case on the other hand, the situation was different. The reception is a relatively large room with stone floor, several glass walls and an irregular shape, which together creates special acoustic conditions. Using the technology available it was only possible to partly mimic the acoustic experience of the reception. In the interview, the test subjects pointed out the following aspects that presented problems when working with the simulation:

- The simulated reverberation was deemed harder and more metallic than the softer, more diffused reverberation of the real room. Having access to a more advanced reverb unit would increase the realism of the simulation.
- Several of the test subjects reported problems finding settings that made it possible to simulate listening to a sound source both close up and at a distance in a way that corresponded to the experience in the real environment. Once again, a technically more advanced reverb unit together with more developed algorithms for filtering and attenuation of sound sources as functions of distance to the virtual listener are needed.

In the first phase of each simulation session, the test subjects were asked to establish a basic loudness level based only on their ears and memory of experienced sound pressure in the environment to simulate. When these basic loudness levels were established, it was measured using a dB-meter. The results show differences between the two cases. In the car-case, the ambient sound of the car interior was used. In the reception case the ambient was too quite to effectively use for this purpose and instead the recording of a conversation between two persons were used to establish the basic loudness level. It turned our that in the car-case, the experienced and measured levels differed significantly. Measurements of the sound pressure in the simulations showed that, when using only ear and memory, all test subjects adjusted the loudness level to between 5 and 10 dB lower than the level measured in the real environment. In the reception-case, when adjusting the loudness level of the conversation, the measured level did not differ more than a few dB up or down.

5. DISCUSSION

In this paper we have reported on a system that can be used to create simulations of the sounding dimension of physical environments. The system consists of a software application, a 5.1 surround sound system and a set of guidelines and methods for use. We have also reported on a first test of the system and the results from this test.

Our aim has been to create an audio work environment with the space-time dimension in focus. Awesome differs from most other digital audio workstations (DAW's) and sound editing systems in one important aspect. In the Awesome-system the listener object can freely be moved and oriented with respect to the simulated environment and the sound sources in it. Traditional systems for editing surround sound, such as Cubase, Logic and Soundtrack Pro [21] are built on a metaphor where the listener is always statically in the center of the environment and cannot be moved in relation to it. Instead, to simulate the listener moving, each sound source must be moved individually. In these traditional systems it is therefore very difficult to make scenarios where the listener is moving relative to the sound sources in real-time. Most often you end up with a couple of audio snapshots or frozen "still lifes" of possible sound scenarios. The Awesome system makes it possible to create more dynamic scenarios and simulations of environments and the listeners relation to them.

Larsson et al. argues that "[o]ne can think about three different levels of technological sophistication when building mixed or virtual auditory environments – physical, perceptual, and cognitive" [22]. On the physical level, this means that high sophistication in the rendered sound scenes means a close approach to the physical sound properties of the real environment, the sound from the loudspeakers sounds more or less exactly as the original sound source. On the perceptual level, Larsson et al. notes that today's knowledge in auditory perception allows for technical simplifications, and gives as example audio coding schemes (i.e. MP3 compression) and reduced temporal and spatial quality of recorded impulse responses. This means that some technical sophistication can be sacrificed if it can be assumed that the listener will not perceive a sound or part of a sound. Larsson et al. also points out the recent notion of ecological psychoacoustics and the importance of cognition when creating auditory scenes. We know that the sound of airplanes usually comes from above, this knowledge tends to steer our perception, so the sound of an airplane is perceived to come from above even though the sound is presented to the listener at their head level. Also in this case some level of technical sophistication can be sacrificed and replaced with the listener's experiences and knowledge about sounds in the world.

When these ideas are applied on the case presented in this paper, it can be argued that the technique used to render the simulations, not necessarily have to be capable of a one-to-one physical copy of the sounds in real environments in order to give a satisfactory and useful experience. Instead of seeking technical perfection, it becomes important to find a relevant balance between a number of parameters, that together build the whole experience of the soundscape simulation. When designing the Awesome system, we have focused on the following aspects:

- **Accessible**. The Awesome system is intended to be a cost-effective system based on a Mac OS-computer and a mid-range 5.1 surround sound system. The aim is to create a system affordable to any SME that can benefit from using the system. It must also be possible to place the system in a wide range of locations and rooms, without the need for anechoic chambers etc. and still create useful results.
- **Ease-of-use**. The target user group is non-sound experts with the need to simulate sound environments and to test and evaluate sounding solutions. As such, the aim is to create a system with a relatively low learning threshold and a clear learning path. The aim is also to complement the software system with guidelines and methods for use that support non-expert users and make it possible also for them to create useful results.
- **Embedded knowledge**. Along the same line is the strive for embedding general knowledge about acoustics, psychoacoustics and sound perception in the system. The (non-expert) target users should be able to focus primarily on their design tasks and be relieved from distracting demands from the system.

The project is currently in its first stage of implementation. The first test carried out shows that the system has come some way in all three aspects above. The system does not have to be extremely expensive and it can be placed in an ordinary office or conference room for example and it is still possible to create useful results. The test subjects considered the system easy to use and they were all able to create results they deemed satisfying in short times. The system can and need to embed more general knowledge about acoustics, psychoacoustics and sound perception. A better and more technically advanced reverb unit is needed to be able to more accurately simulate the acoustics of a wider range of environments. Today, the Awesome system does not have any functionality for calibration. Several parts of the total system could benefit from ways and means to more or less automatically calibrate parameters. The test performed revealed the following parameters as highly prioritized: frequency response; overall loudness; sound pressure roll-off; frequency response and the amount of reverb applied as functions of distance to sound sources.

6. FUTURE WORK

The Awesome system has been verified to be a tool capable of recreating, to a certain extent, the sounding experience of two environments. Next steps in the project are to spread the system to a wider audience of users and to have the system used in a wider range of contexts and projects. Anyone interested is hereby invited to use the Awesome system, to comment on it and to suggest improvements and additions.

It is still to be verified that the system can also be used and productive for persons not specialized in audio engineering or audio design. To further strengthen the system in this aspect, the intention is to include also an automatic or semi-automatic calibration module. This module will assist in the basic calibration and balancing of the loudspeaker system used in the local setup. The calibration module will also assist when establishing the basic loudness and reverberation levels of a project.

The system does not only consist of the software application, but does also include methods and guidelines for the successful simulation of audio environments and the test and evaluation of new sound experiences in these environments. To further develop these methods and guidelines is therefore as important for the future as the development of the software application itself.

7. ACKNOWLEDGEMENTS

Ulf Blomqvist, Stefan Lindberg and the author created the specifications for the basic functionality of the Awesome system. Björn Kempén does all programming of the software. The verification test was designed and conducted by Stefan Lindberg and the author.

Thanks to Johan Fagerlönn for invaluable input and ideas and to Nigel Papworth for inspiration and proof-reading. Thanks also to valuable feedback from the reviewers of the paper.

8. REFERENCES

[1] A. Nykänen, *Methods for Product Sound Design,* Luleå, Sweden: Luleå University of Technology, 2008.

[2] C. Frauenberger, T. Stockman, “Auditory display design – An investigation of a design pattern approach,” *Int. J. Human-Computer Studies*, vol. 67 (2009), pp. 907-922, June 2009.

[3] W.W. Gaver, “What in the World Do We Hear? An Ecological Approach to Auditory Event Perception,” *Ecological Psychology,* 5 (1), pp. 1-29, 1993.

[4] P. Polotti, D. Rocchesso, *Sound to Sense, Sense to Sound – A State of the Art in Sound and Music Computing*. Logos Verlag Berlin GmbH, Berlin, Germany, 2008.

[5] C. O'Callaghan, “Auditory Perception,” *The Stanford Encyclopedia of Philosophy*, Summer 2009 Edition. Retrieved January 20, 2010 from Stanford Encyclopedia of Philosophy http://plato.stanford.edu/archives/sum2009/entries/perception-auditory/.

[6] J. Fagerlönn, M. Liljedahl, “Tapping into effective emotional responses via a user driven audio design tool,” in *Proceedings of Audio Mostly 2009 – a conference on interaction with sound*, Glasgow, UK, 2009, pp. 89-92.

[7] G. W. Coleman, C. Macaulay, A. F. Newell, “Sonic Mapping – Towards Engaging the User in the Design of Sound for Computerized Artifacts,” in *Proceedings of NordiCHI 2008: Using Bridges*, Lund Sweden.

[8] V. Alves, L. Roque, “A Proposal of Sound Design Guidelines for User Experience Enrichment,” in *Proceedings of Audio Mostly 2009 – a conference on interaction with sound*, Glasgow, UK, 2009, pp. 27-32.

[9] S. Pauletto, D. Hug, S. Barrass, M. Luckhurst, “Integrating Theatrical Strategies into Sonic Interaction Design,” in *Proceedings of Audio Mostly 2009 – a conference on interaction with sound*, Glasgow, UK, 2009, pp. 77-82.

[10] Cubase. http://www.steinberg.net/en/products/musicproduction/cubase5_product.html

[11] Logic Studio. http://www.apple.com/logicstudio/

[12] Ableton Live. http://www.ableton.com/

[13] Traktor Pro. http://www.native-instruments.com/#/en/products/dj/traktor-pro/

[14] Audacity. http://audacity.sourceforge.net/

[15] Pro Tools. http://www.digidesign.com/index.cfm?navid=349&langid=100&itemid=33116

[16] Reactor. http://www.native-instruments.com/#/en/products/producer/reaktor-5/

[17] SuperCollider. http://www.audiosynth.com/

[18] Max MSP. http://cycling74.com/products/maxmspjitter/

[19] M-Audio Studiophile AV 40, Destop speaker system, http://www.m-audio.com/ products/en_us/StudiophileAV40.html

[20] M-Audio SBX10, http://www.m-audio.com/products/en_us/SBX10.html

[21] Soundtrack Pro. http://www.apple.com/finalcutstudio/soundtrackpro/

[22] P. Larsson, A. Väljamäe, D. Västfjäll, A. Tajadura-Jiménez, M. Kleiner, “Auditory-Induced Presence in Mixed Reality Environments and Related Technology”, *in E. Dubois et al. (eds.), The Engineering of Mixed Reality systems, pp 143-163*. Springer-Verlag, London, 2010.

CONTENT-BASED RETRIEVAL FROM UNSTRUCTURED AUDIO DATABASES USING AN ECOLOGICAL ACOUSTICS TAXONOMY

Gerard Roma, Jordi Janer, Stefan Kersten, Mattia Schirosa, Perfecto Herrera

Universitat Pompeu Fabra
Music Technology Group
Roc Boronat 138, Barcelona
firstname.lastname@upf.edu

ABSTRACT

In this paper we describe a method to search for environmental sounds in unstructured databases with user-submitted material. The goal of the project is to facilitate the design of soundscapes in virtual environments. We analyze the use of a Support Vector Machine (SVM) as a learning algorithm to classify sounds according to a general sound events taxonomy based on ecological acoustics. In our experiments, we obtain accuracies above 80% using cross-validation. Finally, we present a web prototype that integrates the classifier to rank sounds according to their relation to the taxonomy concepts.

1. INTRODUCTION

Virtual environments based on realistic simulations of physical space are becoming a common use of the Internet. Most of them can be divided among multiplayer games and social environments used to meet and chat. In some cases, people have even become interested in purchasing virtual goods and hence virtual economies have emerged. However, the cost of designing such amount of 3D spaces is very high. Virtual environments have followed the trend towards user-centered technologies that dominates the web. Many programs allow users to create and upload their own models and to design their spaces. Sites such as *Google 3D Warehouse* are available as centralized repositories of 3D models that can be placed in different environments.

So far, these environments offer very sophisticated visual simulations but quite basic audio functionality. Still, applications like *Second Life* allow users to upload custom sounds for objects. In this context, open, user-contributed sound repositories such as *freesound.org* [1] can be used to improve the acoustic experience of virtual environments. However, searching for sounds in user-contributed databases is still problematic. Sounds are often insufficiently annotated and with very diverse vocabularies [2]. Some sounds are isolated and segmented, but others consist of very long recordings containing mixtures of environmental sounds. In this situation, content-based tools can help improving the search and retrieval of sounds. For specific domains, such tools can be based on *a priori* knowledge and constraints. For the case of designing the acoustic experience of virtual environments, we do not consider voice and music audio, for which separate channels are typically allocated, i.e. real user voices for avatars and background music streams. With respect to indexing and retrieval of environmental sounds for virtual spaces, we are interested in categorizations that take into account the way we perceive everyday sounds. In this context, the ideas of Gaver have become commonplace. In [3], he emphasized the distinction between musical listening (as defined by Schaeffer [4]) and everyday listening. He also devised a comprehensive taxonomy of everyday sounds based on the workings of ecological acoustics, while pointing at the problems with traditional organization of sound effects libraries. The taxonomy categorizes sounds according to the type of interacting materials (solids, liquids, gases) and the kind of interaction (e.g for solid bodies, sounds are classified as impact, deformation, scraping or rolling). One example of the use of this taxonomy can be found in the *Closed* project [5], where it was used to develop its physically-based sound models [6].

In this paper, we analyze the use of this taxonomy for retrieving audio from unstructured, user-contributed audio repositories. We test different approaches to description and classification of sounds according to this taxonomy using SVM. We then use the learnt models to rank sounds. In addition to the traditional text search interface, we add the option to filter and sort the results according to a category in the taxonomy. This interface makes it easier to retrieve iconic sounds that represent basic auditory events. However, the focus of this paper is the classification part, and we don't formally evaluate the search interface.

2. RELATED WORK

Automatic analysis and categorization of environmental sounds has been traditionally related to management of sound effects libraries. The taxonomies used in these libraries usually do not attempt to provide a comprehensive organization of sounds. It is common to find semantic concepts that are well identified as categories, such as animal sounds or vehicles. This ability for sounds to represent or evoke certain concepts determines their usefulness in representation contexts such as video or multimedia content creation.

Content-based techniques have been applied to limited vocabularies and taxonomies from sound effects libraries. For example, good results have been reported [7], [8] when using Hidden Markov Models (HMM) on rather specific classes of sound effects. There are two problems with this kind of approach. On the one hand, working with non comprehensive taxonomies omits the fact that real world applications will typically need to deal with much larger vocabularies. Many of these approaches may be difficult to scale to vocabularies and databases orders of magnitude larger. On the other hand, they commonly employ small databases of sounds recorded and edited under controlled conditions. This means that it is unclear how these methods would generalize to noisier environments and databases. In particular, we are concerned with user-contributed media, which involves a wide variety of situations, recording equipment, motivations and skills.

Previous research works have explored the scalability issue by using more efficient classifiers. For example in [9], the problem of extending content-based classification to thousands of labels was approached by using a nearest-neighbor classifier. The system presented in [10] bridges the semantic space to the acoustic space by deriving independent hierarchical representations of both. In [11], scalability of several classification methods is analyzed for large-scale audio retrieval.

With respect to noise and real world recordings, another trend of work has been directed towards the classification of environmental sound using only statistical features, that is, without attempting to identify or isolate sound events [12]. Applications of these techniques range from analysis and reduction of urban noise, to detection of sonic context for mobile phones.

In a way, our problem shares some characteristics of both sound effects and environmental sound classification. This situation comes from the different perceptions and motivations of users at a site like *freesound.org*. Some users will upload sound effects, and many users are interested especially in downloading clean sound effects for using them in music or multimedia productions. But also it is common to upload raw field recordings of different locations and situations as a way to share personal experiences.

The specification of a general taxonomy for environmental sounds remains an elusive problem. Gaver's taxonomy [3] organizes sounds according to how the mechanics of the production of sounds are perceived, from the point of view of ecological acoustics. Further research has given some support to his hypothesis with respect to the perception and categorization of environmental sounds [13]. The taxonomy offers a coherent and general categorization of environmental sounds that is well defined for simple auditory events. However, despite being frequently cited, we don't know of other attempts at automatic classification using this taxonomy.

Gaver proposed a hierarchical classification space, from broad classes to simple sonic events (see figure 1). The root class can be called *Interacting Materials*, since most generally sounds are produced as a result of an interaction of materials. At the next level, the taxonomy divides sounds in three general categories: those involving vibrating solids, gases and liquids. Finally, basic level sonic events are shown at the third level, they are defined by the simple interactions that can cause solids, gases and liquids to produce sound.

3. CLASSIFICATION OF ENVIRONMENTAL SOUNDS

3.1. Overview

We analyze the use of Gaver's taxonomy for general audio segments databases by training and testing a Support Vector Machine (SVM) classifier over a dataset collected from various sources. Our first experiment consists in comparing the performance of different sets of features for the task, in order to assess the importance of describing temporal evolution. A second experiment analyzes two different definitions of the classification problem: as a hierarchical classification or as a direct multiclass problem.

3.2. Datasets creation

For our experiments, we manually selected and labeled sounds according to the taxonomy's categories. We created three datasets: *SoundEvents*, *SoundFx* and *Freesound*.

The *SoundEvents* dataset [14] provides examples of many classes of the taxonomy, although it does not match it completely. Sounds from this database are planned and recorded in a controlled setting, and recordings are repeated multiple times. For example, the sound of metal balls running on plywood is recorded several times in the same session. We discarded the sounds that would correspond to complex events due to the interactions of different materials. A second dataset was collected from a number of sound effect libraries, with different levels of quality. A small number of sounds in this dataset was downloaded from online repositories. Sounds in this dataset generally try to provide a good representation of specific sounds. For instance, for the *explosion* category we select sounds from gunshots, for the *ripple* category we typically selected sounds from streams and rivers. Some of these sounds contain background noise or unrelated sounds. Finally, our third dataset consists of sounds downloaded from *freesound.org*. This set is the noisiest of the three, as sounds are recorded in very different conditions and situations. Most contain background noise and many are not segmented with the purpose of isolating a particular sound event.

The collection of sounds in the dataset presented a number of issues. We now describe the main criteria we used in order to provide a coherent interpretation of the taxonomy.

First of all, in order to allow our classifier to generalize to user-submitted audio, we needed to search sounds with a variable recording setup, recording quality and relative microphone position. Second, we needed to search samples with a less stringent segmentation than the one used in the *SoundEvents* database, where the researchers tried to include just one instance of a basic event in each recorded sample. Thus we considered samples presenting: i) complex temporal pattern repetition of basic events, ii) sounds generated from compound interaction and iii) sounds generated by hybrid interaction. Compound interaction happens when a sound results from the interaction between more than one type of basic event. This is the case of specific door locks, where the sound is generated by a mix of impacts, deformations and scrapings; or the case of missiles, where the sound is generated both by whoosh and explosion. Contrastingly, the sound generated by hybrid interaction happens when a given material interacts with one of a different kind, as in the case of the hybrid event *impact-drip*, when water drips onto a solid surface, or in the case of bubbles, a hybrid between liquid and gas. In order to extend the dataset, we included sound instances that still can be classified at the basic level, but under somewhat less restrictive constraints: repetition patterns of atomic events of the same type, samples containing only a tiny amount of compound or hybrid interactions and samples representing different microphone positions, recording setups and noise conditions.

The taxonomy provides also some parameters related to source attributes that are percieved through sounds. These parameters were useful to qualitatively determine whether samples belonged to a class or not. Some examples: rain was a problematic case, following the original definition, the *Drip* basic event is just water falling into water, with the parameters *viscosity*, *object size*, *object shape* and *force*. In comparison, the parameters of the *Pour* basic event are *viscosity*, *amount* and *height*. Depending on the par-

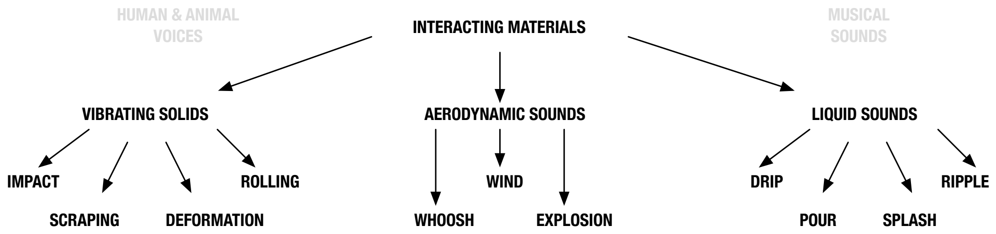


Figure 1: Representation of the Gaver taxonomy.

ticular sample analysed, rain could belong to *Drip*, if individual raindrops were clearly identifyable, or to *Pour*, if the temporal fine structure was undiscernible and the signal closer to noise. Also, if the sound clearly indicated water falling down on a surface, as in the case of rain tapping on a window, the sample was considered to be a hybrid event and discarded, but if the rain contact with the surface was faint, we included it in the *Liquids* category, and it still needed to be categorized into either *Drip* or *Pour*.

3.3. Audio Features

One important question in the discrimination of general auditory events is how much of our ability comes from extracting properties of the instantaneous spectrum, and how much results from following the temporal evolution of the sound. A traditional hypothesis in the field of ecological acoustics was formulated by Vanderveer, stating that interactions are perceived in the temporal domain, while objects determine the frequency domain (quoted in [3]). In several fields involved with discrimination of audio data it has been common to use the *bag of frames* approach, meaning that the order of frames in a sound is ignored, and only the statistics of the frame descriptors are taken into account. This approach has been shown to be sufficient for discriminating different sound environments [12]. However, for the case of sound events we think that time varying aspects of the sound are determinant to recognize different classes. This is especially true for impulsive classes such as impacts and explosions or splashes, and to a lower extent by classes that imply some regularity, like rolling.

In this paper we analyze the performance of some descriptors extracted from the time series of frame level descriptors for our classification task. We test two sets of frame-level features:

- *MFCC*: An implementation of Mel Frequency Cepstrum Coefficients using 40 bands and 13 coefficients.
- *Spectral*: A collection of spectral shape descriptors such as spectral centroid, kurtosis, skewness, crest, decrease and rolloff, along with an estimation of pitch and pitch salience.

We use MFCCs as a reference as they are one of the most commonly used representations of the spectrum. Our second set includes descriptors of the spectral shape that were popularized by the MPEG-7 standard [15]. We also include an estimation of pitch and pitch salience, which have been shown to be relevant for the discrimination of environmental sounds [13]. We compute the mean and variance of every frame level descriptor, as well as mean and variance of its first and second derivative. We also test several descriptors computed from the temporal evolution of frame level features, such as the log attack time, a measure of decay [16] and temporal descriptors derived from statistical moments: temporal centroid, kurtosis and skewness (table 1).

Name	Description	# desc.
mv	mean and variance	2
mvd	*mv*, derivatives	6
mvdad	*mvd*, log attack time and decay	8
mvdadt	*mvdad*, temp. centroid, kurtosis, skewness	9

Table 1: Sets of descriptors extracted from the temporal evolution of frame-level features, and the number of descriptors per frame level feature.

Features	*mv*	*mvd*	*mvdad*	*mvdadt*
MFCC	69.35	75.76	74.98	77.80
Spectral	73.17	78.04	80.02	81.29

Table 2: Average classification accuracy (%) for different sets of features.

3.4. Experiments

We use a Support Vector Machine (SVM) classifier [17] in order to assign a given feature vector representing one sound to one of the classes in the taxonomy. Our first experiment consists in an evaluation the performance of the temporal descriptors applied only to MFCC features. We repeatedly evaluate a *one vs one* multi-class SVM classifier using a set of MFCC descriptors where we progressively add temporal evolution descriptors. We then repeat the procedure with the second set of descriptors and compare the results.

The second experiment consists in comparing the *one vs one* classifier to a hierarchical classification scheme, where we train separate models for the top level classes (solids, liquids and gases) and for each of the top level categories (i.e. for solids we train a model to discriminate impacts, scraping, rolling and deformation sounds). For this experiment we combine both MFCC and spectral shape features with their corresponding temporal descriptors.

Our general procedure starts by resampling the whole database in order to have a balanced number of examples for each class. We then evaluate using ten-fold cross-validation. We run this procedure five times and average the results in order to account for the random resampling of the classes with more examples. We estimate the parameters of the model using grid search only in the first iteration in order to avoid overfitting each particular sample.

Method	accuracy
Direct	84.10
Hierarchical	80.61

Table 3: Average classification accuracy (%) for direct vs hierarchical approaches

4. RESULTS

Table 2 shows the accuracy of the multiclass SVM model for each set of descriptors. While the most important improvement is typically obtained by adding derivatives, the experiment shows that adding the temporal descriptors does help in the discrimination of the different kinds of event. This is true for both MFCC and spectral shape descriptors. On the other hand, it shows that it is possible to obtain reasonably good results with a simple and scalable approach to the description of the temporal evolution of the spectrum. Our best result is obtained when combining both descriptor sets (table 3). As a further step, we plan to compare this results with more complex approaches such as HMM.

Table 3 shows the comparison of the hierarchical approach to the direct classification. While in the hierarchical approach more classification steps are performed, with the corresponding accumulation of error, results are still above 80% on average. This seems to support the underlying hierarchy in Gaver's proposal, in the sense that basic events involving a main class of materials (solid liquid or gas) share some features and can be discriminated from other main classes. This approach has the advantage of providing a model for the top level and consistent models for the lower levels, which may be used for browsing sound databases.

The results of the classification experiments show that a widely available and scalable classifier like SVM may suffice to obtain a reasonable result for such a general set of classes over noisy datasets. Next, we describe the use of the direct approach to rank sounds in the *freesound.org* database. The rank is obtained by training the multiclass model to support probability estimates [17]. The probability estimate is then used as a rank for a query containing one concept of the taxonomy.

5. APPLICATION

A principal objective of the present research is to facilitate the search of environmental sounds in user-contributed audio databases. With that purpose, we integrated the SVM models as a front-end for querying the *freesound.org* database with a combination of textual input and terms from the ecological acoustics taxonomy. First, we review how the taxonomy under study is currently represented as metadata in the *freesound.org* database by social tags.

5.1. Taxonomy concepts in the Freesound folksonomy

Since its inception in 2005, *freesound.org* has become a renowned repository of sounds with non-commercial license, building an active online community at the same time. Currently, it stores 84, 222 sounds, labeled with approximately 18000 unique tags. Sounds are collaboratively labeled with tags, a practice known as folksonomy [18], leading to an unstructured audio database.

Looking at the database content, one can distinguish three main types of sounds: environmental (e.g. nature recordings), musical (e.g. instrument samples, loops) and speech (e.g. individual, conversational).

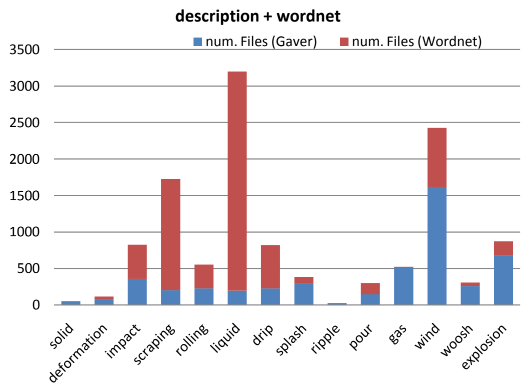


Figure 2: Number of sound files in *freesound.org* containing tags or descriptions with Gaver taxonomy's terms (in red), or their synonyms from Wordnet (in blue).

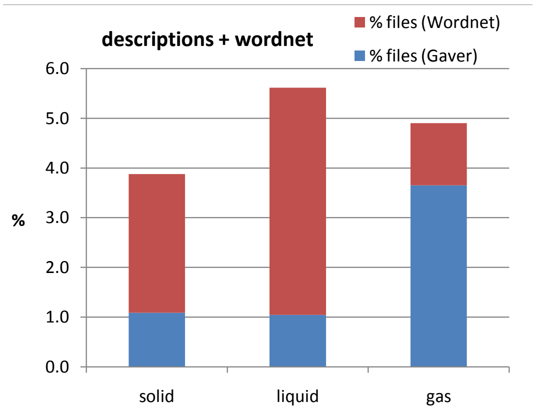
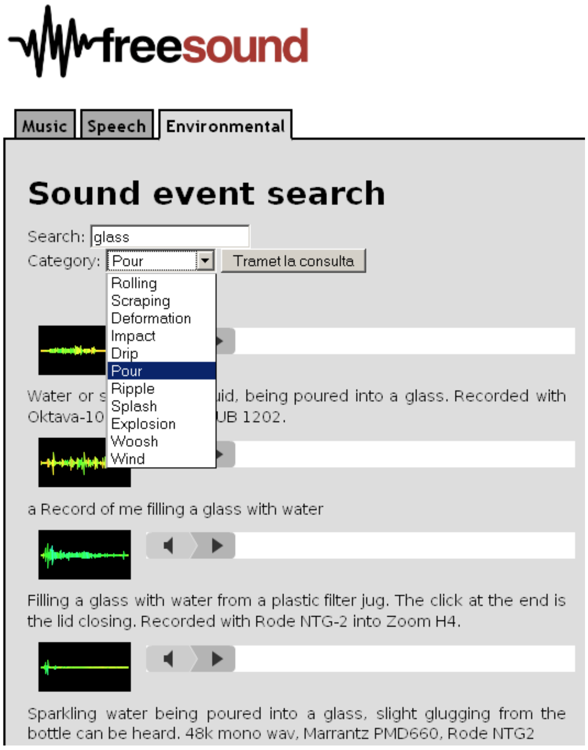


Figure 3: Percentage of sound files in Freesound containing tags or descriptions with Gaver taxonomy's terms (in red), or their synonyms from Wordnet (in blue). Results are grouped by top categories (solid, liquid and gas).

Regarding the environmental sounds tagged in *freesound.org*, the presence of the studied ecological acoustics taxonomy is scarce. Figure 2 shows the histogram of the taxonomy's terms (in blue), grouped by the top-level categories (solid, liquid and gas). In order to widen the search, we aggregated to each term various semantically-related tags that appear as a *synset* (synonym set) in the Wordnet database [19]. The number of files retrieved with the *s*ynset are shown in red. Figure 3 shows the histogram of files grouped by the top categories in the taxonomy, i.e. solid, liquid and gas. In this case, values indicate the percentage of files compared to the total files in the database (84, 222). Patently, the concepts used in ecological acoustics are infra-represented in the the folksonomy. Next section describes a practical application of how content-based retrieval can assist the search with these concepts.



Sound event search

Search: glass
Category: Pour

Water or s[illegible]uid, being poured into a glass. Recorded with Oktava-10[illegible] UB 1202.

a Record of me filling a glass with water

Filling a glass with water from a plastic filter jug. The click at the end is the lid closing. Recorded with Rode NTG-2 into Zoom H4.

Sparkling water being poured into a glass, slight glugging from the bottle can be heard. 48k mono wav, Marrantz PMD660, Rode NTG2

Figure 4: Screen-shot of the web-based prototype.

5.2. Extending text-queries with a content-based classifier

In this application, we are only concerned with environmental sounds. Hence, the retrieval of musical and speech sounds would have a negative effect on our search results. A pre-process to automatically classify these three sound categories is currently beyond the scope of this paper. Instead, here we opted for filtering the dataset using tags related to environmental sounds in order to reduce the retrieval of musical and speech sounds. The final subset is built by departing from the “field-recording” tag, which has become one of the most general of the database. We consider this tag as one of the main themes of the site along with “voice” and “loop”, which respectively represent speech and music samples. We compute the cosine distance of all tags to the “field-recording” tag and keep all tags within a distance below an empirically determined threshold. We limit the search to files labeled with these tags. Also, we limit the file duration to 10 seconds in order to avoid the retrieval of long soundscape recordings, reducing the whole database to 1934 sounds.

On this restricted dataset, our prototype allows searching by ecological acoustics terms. A free word is input in a text box, and a term from the taxonomy is selected from a list. The idea behind this scheme is that the query is composed of a *subject* represented by the free text word, and a *predicate* represented by a class of event. The query is matched to tags and descriptions, and the results are ranked according to the output of the automatic classifier for the given class, as described in section 3. Figure 4 shows the GUI of the search prototype.

5.3. Discussion

While we didn’t formally evaluate the search interface, we informally analyzed its viability by trying several common queries composed of a word plus a term of the taxonomy. We compared the results to the ones returned from multi-word queries by the regular search engine at *freesound.org*, which matches text queries to tags and descriptions, and ranks the results by popularity (number of downloads). We observed that for some queries (e.g. glass+pour) the content-based approach represents a significant improvement over the traditional text-based search. In many cases, content-based indexing helps reducing the effects of ambiguity and incomplete or noisy text descriptions. As a side effect, we observed that the content-based search is much more restrictive. Depending on the query, it may return an empty list, if none of the matched sounds were classified into the specified category. A development version of the web prototype is publicly available [20].

6. CONCLUSIONS

This research aims to improve the search of environmental sounds in large-scale unstructured audio databases. Specifically, we contribute with a content-based analysis and classification framework built upon the ecological acoustics taxonomy proposed by Gaver [3]. To our knowledge, previous approaches on content-based analysis of environmental sounds, have only addressed very concrete sound categories (e.g bird calls, sirens, car engine), without tackling the usage of a general taxonomy.

We proposed a supervised learning approach, and created a annotated database providing several examples of all categories present in the taxonomy. By means of an automatic classifier, the system ranks the sounds according to the acoustic similarity to each class in the taxonomy. We implemented a search interface to *freesound.org* using this system, with promising results. We plan to experiment with other ways to interact with the database using this framework, such as a more exploratory browsing interface.

7. ACKNOWLEDGMENTS

This work was partially supported by the ITEA2 Metaverse1 Project (http://www.metaverse1.org).

8. REFERENCES

[1] Universitat Pompeu Fabra. (2005) Freesound.org. Repository of sounds under the Creative Commons license. [Online]. Available: http://www.freesound.org

[2] E. Martínez, O. Celma, M. Sordo, B. De Jong, and X. Serra, “Extending the folksonomies of freesound.org using content-based audio analysis,” in *Sound and Music Computing Conference*, Porto, Portugal, July 2009.

[3] W. W. Gaver, “What in the world do we hear? an ecological approach to auditory event perception,” *Ecological Psychology*, vol. 5, pp. 1–29, 1993.

[4] P. Schaeffer, *Traité des objets musicaux*, 1st ed. Paris, France: Editions du Seuil, 1966.

[5] http://closed.ircam.fr.

[6] D. Rocchesso and F. Fontana, Eds., *The Sounding Object*. Edizioni di Mondo Estremo, 2003.

[7] M. Casey, “General sound classification and similarity in mpeg-7,” *Organised Sound*, vol. 6, no. 2, pp. 153–164, 2001.

[8] T. Zhang and C.-C. Kuo, “Classification and retrieval of sound effects in audiovisual data management,” in *Conference Record of the Thirty-Third Asilomar Conference on Signals, Systems, and Computers*, vol. 1, 1999, pp. 730–734.

[9] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, N. Wack, and P. Herrera, “Nearest-neighbor automatic sound annotation with a wordnet taxonomy,” *Journal of Intelligent Information Systems*, vol. 24, no. 2, pp. 99–111, 2005.

[10] M. Slaney, “Semantic-audio retrieval,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’02)*, vol. 4, 2002, pp. IV–4108–IV–4111.

[11] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, “Large-scale content-based audio retrieval from text queries,” in *Proceeding of the 1st ACM international conference on Multimedia information retrieval (MIR ’08)*. New York, NY, USA: ACM, 2008, pp. 105–112.

[12] J.-J. Aucouturier, B. Defreville, and F. Pachet, “The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music,” *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[13] B. Gygi, G. R. Kidd, and C. S. Watson, “Similarity and categorization of environmental sounds,” *Perception & Psychophysics*, vol. 69, no. 6, pp. 839–855, 2007.

[14] http://www.psy.cmu.edu/~auditorylab/AuditoryLab.html.

[15] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005.

[16] F. Gouyon and P. Herrera, “Exploration of techniques for automatic labeling of audio drum tracks instruments,” in *Proceedings of the Workshop on Current Directions in Computer Music (MOSART)*, Barcelona, Spain, 2001.

[17] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[18] A. Mathes, “Folksonomies - cooperative classification and communication through shared metadata,” December 2004. [Online]. Available: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

[19] C. Fellbaum *et al.*, *WordNet: An electronic lexical database*. MIT Press Cambridge, MA, 1998.

[20] http://dev.mtg.upf.edu/soundscape/freesound-search/.

SOCIO-EC(H)O: FOCUS, LISTENING AND COLLABORATION IN THE EXPERIENCE OF AMBIENT INTELLIGENT ENVIRONMENTS

Milena Droumeva

Faculty of Education,
Simon Fraser University
mvdroume@sfu.ca

Ron Wakkary

Associate Professor,
School for Interactive Arts and Technologies
Simon Fraser University
rwakkary@sfu.ca

ABSTRACT

In this paper, we aim to conceptualize the connection between embodied interactions and the experience of understanding a dynamic auditory display response. We have termed this concept *aural fluency* and hereby continue previous work documenting in more detail the listening patterns that emerge in users' experiences with ambient intelligent environments. Aural fluency describes the acquired listening competency and focus on sonic feedback that users form over time in systems utilizing responsive ambient audio display and collaborative embodied interaction. We describe *listening positions* that characterize the concept and show the stages of aural fluency. The concept arose from the design, analysis and evaluation of an embedded interaction system named socio-ec(h)o – a project upon which we also build on from previous work in the hopes of elucidating further the complex experiences of listening attentions and thus offer insights to the field of auditory displays.

1. INTRODUCTION

The need for new concepts of how users understand and make use of their knowledge of system displays arise as tangible computing, embodied interaction and ambient intelligence systems become increasingly possible. Our understanding of the design of interaction has advanced considerably within a traditional human- computer desktop view that emphasizes visual perception and mental cognition, however there is the additional need to explore concepts related to embodied interaction, sensory perception and ambient audio- visual displays with an emphasis on social interaction or at the least multi-user settings. We see real value in adding to the emerging literature of case descriptions of collaborative and embodied interaction systems but even more critically the need to explore new concepts that emerge from in-depth empirical studies of systems.

In this paper, we aim to conceptualize and describe the connection between embodied interactions and the experience of understanding a dynamic auditory display response. We have termed this concept aural fluency, and we build on past work to not only describe better the stages and levels of aural fluency, but to also offer ways of codifying and examining them in context and over time. We believe there is a need for ideas of interaction and meaning-creation with respect to auditory perception to evolve and reflect the different reality that embodied and tangible interactive systems offer in shared, collaborative, temporally and physically persistent contexts. Aural fluency describes the acquired listening competency that users form over time in systems utilizing responsive ambient audio display and collaborative embodied interaction. The concept arose from the analysis and evaluation of an embedded interaction system named socio-ec(h)o. While we have already introduced it to the auditory display community [1], we now hope to build upon that with a discussion on listening patterns as they relate to embodied interaction and problem-solving in ambient intelligent environments.

First we'll explore the concept of aural fluency by theoretically explicating notions of listening, acoustic interpretation, and auditory training, and their relationship to embodied action and interactivity, situated cognition and perceptual dimensions of interpretation and context. We reference auditory perception and auditory display design literature, as well as the acoustic ecology and acoustic communication frameworks put forth by Schafer and Truax [2,3]. Finally we touch on a discussion of multiliteracies in educational discourse and Gilford's dimensions of fluency [4]. Using extensive qualitative analysis from speech transcripts, post-evaluation survey and video coding, we then begin to paint a picture of aural fluency as a concept that is central to the relationship between ambient system and human user through documenting listening patterns in ambient intelligent environments. We further situate these emerging concepts into both paradigms – that of auditory training and that of acoustic communication, in order to enrich our understanding of both in light of embodied interaction. We conclude with a conceptual synthesis of its role in the technological ecology of ambient, embodied interactive environments, as a vehicle for interpretation, meaning-creation and communication with the system.

2. FOCUS, LISTENING AND AURAL FLUENCY

Rather than delving into the philosophical and epistemological foundations of 'fluency' we focus on the connection between aural competency and embodied interaction in responsive environments particularly in its relationship to a distributed cognition model of problem solving. There are three paradigms that we propose will help flesh out and develop the concept of aural fluency: one comes from the auditory display design field and refers to facilitating the user's ability to derive meaningful information from sound; the other refers to patterns of [everyday] listening from the acoustic communication tradition;

and finally a perspective from education and cognitive science – four dimensions of 'fluency' as intellectual abilities.

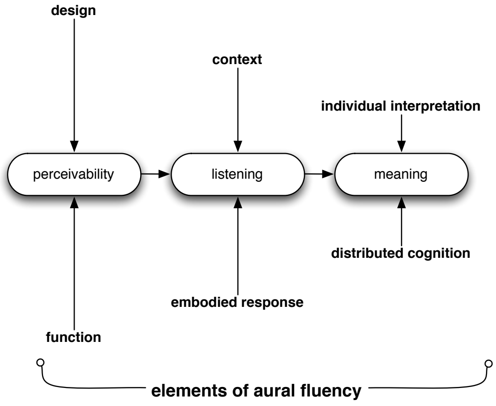


Figure 1. Elements playing a part in users establishing and cultivating an aural fluency for a system's sonic feedback.

In auditory perception research, the problem of aural training and its relationship to "perceivability" is well known. Auditory training is an essential element in areas dealing with the display of information using non-speech sound. Both in earcon/audio icon design practices [5, 6, 7] and especially in the field of sonification [8, 9], efficiency is largely defined and achieved through a careful balance between perceptual intuitiveness of the sonified information, meaningful mapping to data, and facilitation of learning how to make sense of the auditory display. Implied in this task is not only auditory perception but also auditory memory as well as associative and connotative epistemic aspects of listening. Yet there is relatively little research focusing on precisely this area of auditory display design – the nature, qualities and characteristics of how users learn to interpret complex sonic feedback. Further, there is little research into collaborative tasks that require group work and group auditory perception of feedback. This is where our project focuses and so in this paper we attempt to offer usable heuristics of the way in which collaboration, focus and listening patterns work together to constitute an "aural fluency" of a given system's feedback (see Figure 1 for our conceptual proposition).

Our design for socio-ec(h)o, albeit taking a novel direction in sonifying movement or intensity (second-level interpretations from data), is still inspired by some of the foundational literature addressing the balance of 'perceivability' and auditory training, including Flowers and Hauer's research in auditory scatter plots and graphs [10]; the works of Bonebright et al. [11] and Brewster and Brown [12] in perception of multiple data series sonified simultaneously; and Hunt and Hermann's research into interactive sonification [13]. Hunt and Hermann specifically found embodied gesture pairs well with sound feedback, and offloads some attention focus so as to facilitate comprehension from sound. In the more general field of sonification, Smith and Walker's article [14] is an example of applying contextual and peripheral auditory cues to facilitate learning of audio response in sonifications of financial information. The authors concentrate their study on the use of "additional context" elements – auditory tick-marks, axes and labels – and find it moderately helps with comprehension, however what is significant is their finding regarding the role of a reference tone in accurate value estimation by the users. In another study, referenced by Walker, Lindsay and Godfrey, the researches create a principle curve from dense scatter plots, and sonify individual data points "around the curve" as amplitude fluctuations of the main frequency, experienced as Doppler shifts [14] – thus taking advantage of the ear's natural sensitivity to the directivity of sound.

Both of these aural facilitation techniques – context-based and a continuous auditory graphing - have been shown to dramatically impact accurate value estimation, general perceivability of data, response time and accuracy of comprehension [14, 16, 17, 18, 19]. Most of these studies however do not necessarily address the *temporal* and/or *collaborative* effect of aural competency, perception and fluency when interacting with a technological system, nor do they identify discreet learning states and design for them or with them in mind. The type of auditory fluency that is needed and seems to develop in more physical, situated technological environments such as embedded interaction spaces is much more akin to everyday listening. Thus, in order to enrich our understanding of the activity of listening we look to the way it has been articulated by Truax [2] and Schafer [3] (in natural settings, from a holistic perspective), as well as Gaver [20], Bregman [21] and Ballas [22] (from an auditory perception perspective). By analysing everyday and natural acoustic environments, Truax and Schafer build an understanding of listening as an epistemic activity – a complex and dynamically shifting process, nested and interdependent on context and setting, and not simply a mechanical or a psychological process of perception (see Figure 1). Based on Schafer's typology of the natural sonic environment [3] and the World Soundscape Project's ethnographic work on local soundscapes, Truax articulates several ways of listening that he calls 'listening positions' [2]. These positions – modes or states of listening allow us to 'tune in' or 'tune out' of our sonic environment, picking out cues when needed and ignoring others. Understanding these positions then is crucial to design of auditory displays, especially ones that are nested in physical systems of embedded interaction, given that such conditions most resemble everyday contexts.

Listening-in-search is a position where the listener is actively searching either for particular cues in the environment (for example, the cocktail party effect – where we fine-tune our ears to pick out a faint familiar voice amongst a crowded noisy space) or is actively listening for any auditory cue that might be of use/meaning. Listening-in-readiness refers to a combination of non-active listening (background listening) and a pre-cognitive attention to a specific (typically familiar) sound that is expected to occur – such as a baby cry at night or the car of a familiar person. Background listening is one where we intentionally tune out the surrounding soundscape in order to focus on other modalities of feedback or tasks at hand, and finally, analytical listening is one of the listening positions that Truax attributes to the technological development of electroacoustics, recording and reproduction of sound. It is an active, deconstructive listening that occurs in situations where the user knows that sound is designed for an artistic or informational purpose, and there is a cognitive, critical engagement with the nature, characteristics and properties of sound. Lastly, an important point Truax makes about listening in everyday, physical environments is one relating psychoacoustic properties of listening to the context – namely, the fact that our ears attune, depending on the setting and our associations with it, to particular sound properties and changes [2].

Based on this brief bridging of literature on auditory training as well as characteristics of listening, we define aural fluency as a developed competency in using a system that responds through sound. In the sections to follow we hope to explicate the details of this competency and show how it may be developed by learning and acquiring key listening positions. The positions allow users to become increasingly competent and efficient in interpreting the system's response, as well as to link, in an embodied manner, their ability to affect the system through actions, and finally – to do so in a collaborative manner, as a group.

When touching on 'fluency' however, it is important to note some of the cognitive and educational connotation of this concept. Many theorists studying creativity and originality touch on 'fluency' and one of the most prominent voices is that of Gilford's 1962 [4] 'intellectual abilities' expose, where he defines four kinds of fluency (as related to literacy): word fluency, ideational fluency, associative fluency and expressive fluency. If we re-interpreted these notions to relate to aural fluency, they too, offer a usable framework for thinking about the discreet skills that players must develop in order to effectively use sonic feedback as guidance in collaborative problem-solving.

3. PROJECT BACKGROUND

Since we have already published the design details of our project elsewhere, including in the ICAD community, and the fact that we are specifically focussing here on second-level interpretive findings, we'll only provide a brief description of the project in this section. socio-ec(h)o is a prototype environment for a playful collaborative puzzle-solving activity, whose ultimate goal it is to explore the design, use and evaluation issues of embedded interaction systems and social interaction. The overall research goal is to understand how to support groups of participants as they learn to manipulate an embedded interaction space, to understand and interpret feedback reliably and effectively and achieve competency by problem-solving as a team. The research questions are numerous in a project of this nature and yet immersive and embodied interaction does not lend itself to reducible and isolatable variables that can be measured independently. Given this, our evaluation protocol focused on ecological investigation and theory-building, aiming to provide broad, yet particular set of heuristics that help describe and make sense of the different system and display components of such environments.

The socio-ec(h)o prototype is a six-level puzzle game played by a team of four people in a "black box" space with controlled immersive lighting and a surround sound auditory feedback (see Figure 2). The puzzles' solutions are collaborative whole-body physical configurations that players have to achieve and hold for a duration of time, as a team. Their movements in space are tracked by a 3-d setup of infrared cameras. Each level is represented by a unique set of light and sound feedback and the team's progression is interpreted in real time by a reasoning engine resulting in a change in the feedback's intensity intended to inform and guide players towards the right solution. The levels are progressively more challenging in terms of body states as well as in terms of transparency or 'perceivability' of feedback as represented through changes in the environment in light and audio.

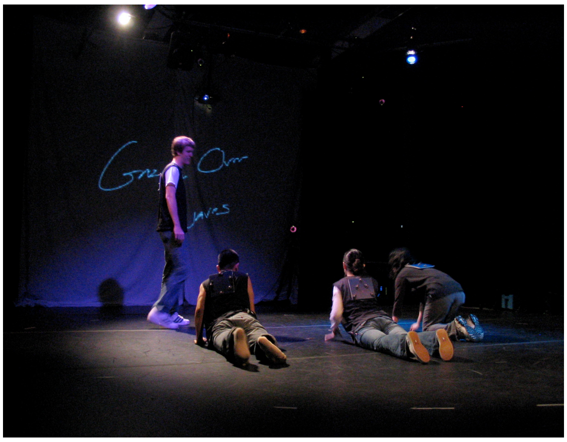


Figure 2. socio-ec(h)o gameplay, Level 5. Participants work together to solve the puzzle "Gazing Over Waves" and slow down fast pulsating water streams into one coherent wave.

3.1. Auditory Display Schema

Again, since we have detailed the audio display approaches in socio-ec(h)o elsewhere [1, 23] we will just briefly explain the main innovation, which was our *intensity gradient* mapping of parametric sound change to a real-time interpreted value of the team's progress in the game. The main approach to the sonic feedback is intensity-based parametric change of a continuous ambient sound adapted from the area of sonification, however, once players reach an intensity of around 3.5 we also implemented a confirmatory signal, which they were introduced to before commencing the game (see Figure 3 below).

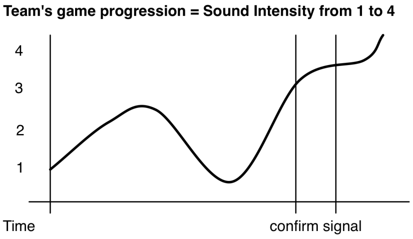
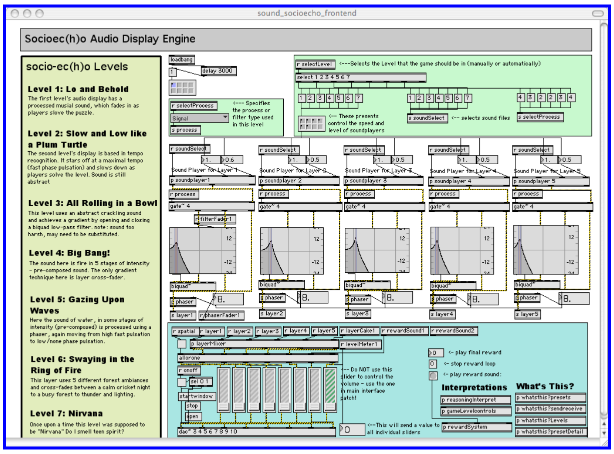


Figure 3. A schematic of a given game progress and its corresponding sound feedback mechanism.

While the base ambient sound was different for each level, attempting to *narratively* fit the puzzle and required movement at hand, the confirm signal (reward – short granulated glass clink) was consistent throughout the game levels. Figure 4 presents a screenshot of the audio engine used, where we see a 5-layer cross-fader with a selection of parametric effects attached to them, which get called up with the beginning of each level. The base ambient sounds for the levels included environmental recordings such as water, fire and abstract soundscapes such as percussion impact sounds (e.g. from marble or dice) and a polyphonic tonal drone. The way sound display fit into the game as a core mechanic was in following the natural skill mastery progression of the game – the soundscape in Level

1 was perceptually easy to interpret (an abstract musical drone where progress is mapped only to amplitude) while the soundscape for Level 4 was significantly more ambiguous (a fire of increasing intensity, as represented by five different fire crackle/bonfire sound files mixing dynamically). Thus ambiguity, perceptual and aesthetic, rather than being avoided, became a core part of the game experience itself.



sound_socioecho_frontend

Socioec(h)o Audio Display Engine

socio-ec(h)o Levels

Level 1: Lo and Behold
The first level's audio display has a processed musial sound, which fades in as players slove the puzzle.

Level 2: Slow and Low like a Plum Turtle
The second level's display is based in tempo recognition. It stars off at a maximal tempo (fast phase pulsation) and slows down as players solve the level. Sound is still abstract

Level 3: All Rolling in a Bowl
This level uses an abstract crackling sound and achieves a gradient by opening and closing a biquad low-pass filter. note: sound too harsh, may need to be substituted.

Level 4: Big Bang!
The sound here is fire in 5 stages of intensity - pre-composed sound. The only gradient technique here is layer cross-fader.

Level 5: Gazing Upon Waves
Here the sound of water, in some stages of intensity (pre-composed) is processed using a phaser, again moving from high fast pulsation to low/none phase pulsation.

Level 6: Swaying in the Ring of Fire
This layer uses 5 different forest ambiances and cross-fades between a calm cricket night to a busy forest to thunder and lighting.

Level 7: Nirvana
Once upon a time this level was supposed to be "Nirvana" Do I smell teen spirit?

<--- Specifies the process or filter type used in this level

<---Selects the Level that the game should be in (manually or automatically)

<-- These presents control the speed and level of soundplayers

<-- selects sound files

<-- Do NOT use this slider to control the volume - use the one in main interface patch!

<--This will send a value to all individual sliders

<- play final reward

<- stop reward loop

<- play reward sound:

Figure 4. The audio display engine, in Cycling 74's Max/MSP software.

3.2. Study Design And Results

As mentioned earlier, this project involved many research questions and points of interest. Thus we'll hereby briefly describe the study protocol and results and then we'll focus specifically on findings related to the reception of sonic feedback. In that, we'll elaborate on our second-level analysis of stages of aural fluency, as well as exemplify the role of listening, focus and collaboration in that.

For socio-ec(h)o's evaluation we had 14 teams of four playing the game, for a total of 52 participants. Teams had an hour and fifteen minutes to complete all six levels (if they could), with a break in the middle. All study sessions were videotaped for the purpose of later video analysis. We also collected times of completion measures, verbal transcripts, and a post-activity survey. The survey instrument contained a combination of Likert-scale questions and open-ended questions relating to the effectiveness of the system's response in helping, guiding the problem-solving, and creating an enjoyable experience for users. In terms of the auditory display, we hoped to see some consistency in performance in certain levels across teams – this could point to a design success or flaw of a particular approach to sound feedback taken in that level. The survey and transcripts in turn were meant to serve as indicators of what participants thought about the design. In fact what we saw in the results is no consistency in the times of completion (ToCs) of different teams, even within the same level – for level 4 alone some teams took less than 3 minutes to solve and other teams took over 45 minutes! What is curious and cannot be explained by our initial results is participants' ratings of the effectiveness of soundscapes and audio response in different levels. According to the survey results, most participants both preferred the soundscape of level 4 and thought it worked best (this level's puzzle was titled "Big Bang" and its ambient sound base was a gradation of fire). On average, participants thought highly and positively of the system's accuracy and effectiveness with regard to sonic and other response, regardless of their individual performance. There is no correlation between performance and preference for sonic feedback. So we were left wondering how it is that teams uniformly found the sound response to be supportive and the experience positive regardless of their ability to complete the puzzles in a timely fashion or play the game "well"? The answer would have to be that something else acted as intrinsic motivation within the game, and/or there were other skills besides the ones we hypothesized about that were developing – social and team cohesion, focus on gameplay and efficacy in problem-solving together, as well as fluency in interpreting cues from the system. This question led us to explore the idea of aural fluency as a skill that developed over time and was intrinsically motivating. This concept is constituted through embodied interaction and offers its own rewards besides the game completion goal – a sense of mastery, familiarity and understanding of how to interpret and manipulate the system even if the explicit goal is not achieved.

4. LISTENING, FOCUS AND COLLABORATION

We devote the remainder of the paper to describing heuristics for aural fluency derived from our study. By way of theoretical descriptions and rich accounts we aim to demonstrate a level of transferability of our concept that goes beyond the scope of our study to apply to other embedded interaction settings, particularly collaborative AmI environments.

4.1. Listening Positions of Aural Fluency

As detailed in the theoretical section on aural fluency in the beginning, one of the paradigms that could be used to frame and explicate how and why aural fluency forms in situations of embodied interactive systems is that of listening positions [2]. Users in fact form these listening positions over time, in order to make sense of and cope with the auditory feedback presented to them. We hereby present several positions that we articulated based on our observations and video analysis; we describe them; offer vignettes from the study and then present three stages of aural fluency as design guidelines contribution. As shown in Figure 5 below, one thing we want to emphasize is the importance of what we call a "narrative transformation" as a cognitive link between listening as a perceptual activity and interpreting sonic feedback as a co-constructed, collaborative meaning-making activity.

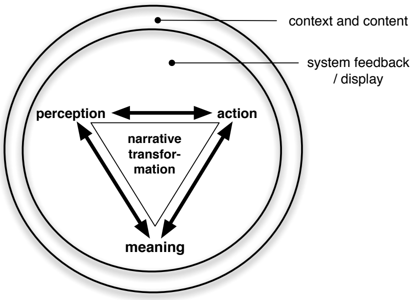


Figure 5. A schema of the interplay between perception action and meaning formation through narrative transformation.

In short, we propose that unlike individual interaction with sonic feedback, where users may employ more analytical and discriminate stances towards listening and interpreting feedback, the nature of collaborative listening has the added factor of the group listening *together*, and negotiating the shared meaning through translating sound changes *narratologically* into common references – i.e. "sounds like fire…a camp fire?..."… "we got something here.. let's keep it going" … "oh, now it's cold…" … "that's a good sign isn't it? Warmer?" "It's fast, that's good right? … no slow is better, fast is cold.", etc. It may be simply that single users interacting with an auditory display don't have to *externalize* their interpretation, however, that is precisely why it is imperative that we look into the unique way in which groups make sense of auditory displays. Following are several listening positions that we saw emerge as patterns during our observation and video coding.

4.1.1. Resetting

This is a listening position where players are often static – not moving, or not playing, listening in full attention to the sound in preparation of a strategy for an embodied configuration. It may or may not involve verbal communication about the sound feedback, but it is clearly focused on understanding and familiarizing themselves with it. In its first iteration, novice players who are taking time to familiarize themselves with the soundscape of a new level often utilize this position. As an extension of this pattern comes a process we named *auditory memory flush*, where players exit the game/space or otherwise bring feedback to its minimum in order to study it or start fresh. It is another well-known issue in sonification that several researchers have worked to address - essentially, auditory memory is fairly short and if people are asked to compare two or more sounds, or continuously compare a dynamically shifting sound in order to derive meaning, auditory memory is fragile. Eventually it gets 'full' and people start losing confidence in the just noticeable differences between different sounds or sound portions. Having a bottom-line reference tone is one way to solve this problem. In our case, the way people solved this problem for themselves was that they would knowingly go into a state or area of low intensity, in order to re-adjust their 'listening position' and start fresh with interpreting the gradient intensity change from there on.

4.1.2. Experimental Listening

In this listening position players are typically in motion and have formed embodied configurations while actively listening for a change in the soundscape that would indicate to them whether they are on the right track or not. While *resetting* is usually characterized by verbal references to the sonic response, participants in an *experimental listening* position typically indirectly reference sound. Verbal references are related to whether players are "doing something right" or not; whether the movement or position seems "hot, warm or cold" and whether (the system) "likes it" or not. The significance of this listening position to aural fluency is that it is the strongest form of communication between system and users, and thus plays the strongest role in learning and skill acquisition. Once users are more comfortable interpreting a particular soundscape – its approach to intensity and rate of change, there is a marked improvement in speed with which they are able to experiment, listen and make judgements about their progress.

Related to this 'listening to change' is the notion of *tolerance of ambiguity* - a concept that surfaced as an important feature of ambient sound feedback in ubiquitous computing spaces in general. Changing sound, especially complex changing sound, always has degrees of ambiguity as a form of feedback, however, in this situation of embodied interaction – a physical and spatial relationship to sound, participants were accommodating of less than clear auditory feedback. In respect to acoustic communication and auditory display paradigms, this concept speaks to the type of everyday listening that people are proficient in utilizing already – making sense of confusing, unclear, complex sonic situations by selectively focusing or shifting attention on different aural elements, and fine-tuning their ear to particular sound structures and qualities.

4.1.3. Narrative Listening

When examined through the lens of everyday listening practices, as developed by Schafer and Truax [2, 3], we see numerous examples of both players forming a narratological association with the soundscape – "no, no we got no fire, somebody has to keep the flame!"… as well as seamless connections to embodiment – "so the fire builds up and we're all still…what if we move towards the sound? What if…when the fire gets all crazy we move more with it?" This narratological relationship could also serve to explain why negative polarity was not a problem in Level 2, when the solutions required moving very slow (so the slower they move the slower the sound gets) but it was a problem in Level 5 where the relationship between puzzle, movement required and sound feedback was less seamlessly related, more abstract. Prior associations and familiarity with sound definitely coloured user experiences with the sonic feedback and their interpretation of its meaning, range, connotations, etc. In fact, interestingly, recognizable sounds seemed, in our preliminary tests, to present a more intuitive range and scaling (to use sonification terms) to users than abstract sounds. This principle also relates to acoustic communication theory, which discusses people's ability to naturally derive information from the structure and quality of sound at the micro level of perception, depending upon the context. In other words, it is the context that shapes listening in such a way that it can fine-tune information retrieval from subtle sound changes by recognizing patterns in the sound. As Truax points out, it is these patterns that mediate the relationship between people and environment [2].

4.1.4. Anticipatory Listening

This listening position usually emerged once players achieved the preceding listening positions and were able to anticipate sonic responses. For example, often players came close to solving a puzzle and would then hear an intermediate reward sonic cue. Through actively interacting, moving, and at the same time listening they could anticipate the upcoming reward sound. This listening position is usually accompanied by verbal references like "we almost had it!" or "we heard the sound, let's keep it up" and then verbal exclamations when the intermediate reward is activated again.

4.2. Aural Fluency Accounts

In order to illustrate the typology of listening positions in systems of embodied interaction, we offer several vignettes from our 14 sessions, of selected levels. We describe the progression and interplay of the listening positions and show screenshots of our video coded data demonstrating the relationships between listening, focus and collaboration. In other work {REF}, where we focus more on the interactional qualities of socio-ec(h)o, we have already identified two other team states as relevant heuristics that characterize the experience and success of teams' problem-solving. These are game focus (the extent to which players are focussed on playing the game) and team cohesion (the extent of collaboration within a team in any given moment). In the figures to follow, we not only present coded versions of the listening positions we have identified, but we also juxtapose them along our coding of the teams' focus and cohesion in order to paint a fuller picture of the complex process of listening in a collaborative, problem-solving, ambient intelligent environment.

<u>*Team G, Level 4*</u> In this vignette, Team G has managed to reach level 4 however they are still developing their collaborative aural fluency, and they are thus not successful at solving the puzzle for level 4. Level 4 is in many respects a "true test" for teams since it is possible to reach this level with some luck, not just skill, and is the first level with two sequential stages of solution, which requires a heightened awareness of the system's response. Yet players spend a considerable amount of time only passively listening (in black on Figure 6 below), along with many breaks in their game focus. Not understanding why they have been unsuccessful, the players attempt to retrace their steps in this level for a while, crouching in the centre waiting "until we have something" [*resetting*] and not moving and intently listening as a means of *experimental listening*. The lack of communication about the sound [resulting form *narrative listening*] and focused experimenting may in fact be an indicator for their failure to solve the puzzle. Further, we see in the zoom-in section below that their shifting of listening attention often coincides with shifts in game focus – perhaps a greater indication of attentional issues.

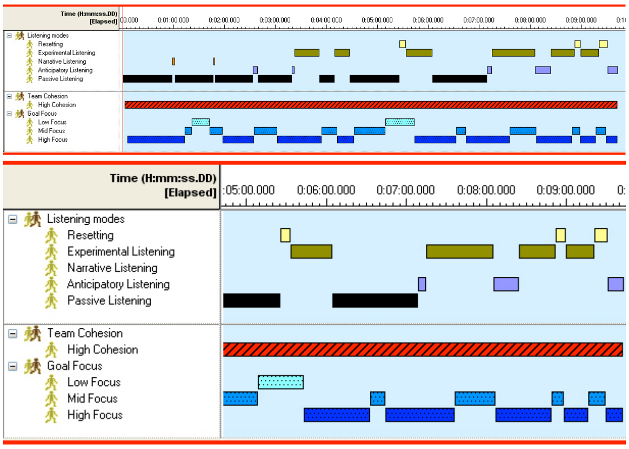


Figure 6. Video coded sections from Team G, level 4.

<u>*Team C, Level 4*</u> This is one of the lengthy examples of level 4, taking almost 50 minutes to solve, and in actuality this team ran out of time and wasn't able to solve it. Yet the timeline more than ever demonstrates clear patterns of fluency over various aspects of the auditory feedback and as such – manipulating of system response. Since this level has two stages of completion, when teams achieve stage 1 they hear the confirmatory signal, and like many others, Team C employed a mixture of *experimental* and *narrative listening* in order to achieve stage 1; and after that they alternated between anticipatory and experimental listening stances in order to try and complete the puzzle. What is interesting to note is that once they had figured out stage 1, they employed anticipatory listening when they came together in physical formation for it as a group, without any explicit communication about listening. Only if someone moved too early they would comment: "no, we haven't heard the sound yet…" or "see, it's getting hotter here, the fire's bigger". After mastering stage 1 the team is clearly displaying aural fluency over the system's response, eliciting it with ease and precision. Every subsequent time they achieved stage 1, they continued either to anticipate the aural completion of the level poised in a passive physical position, or attempted different formations while listening experimentally. In all of these cases, listening was performed 1) while multitasking with movement and even team communication, and 2) together as a group, in an implicit (or occasionally – explicit) listening agreement. Finally, we see even more clearly here many shifts and transitions between different listening and game focus states over time, with more breaks in cohesion and durations of passive listening towards the end.

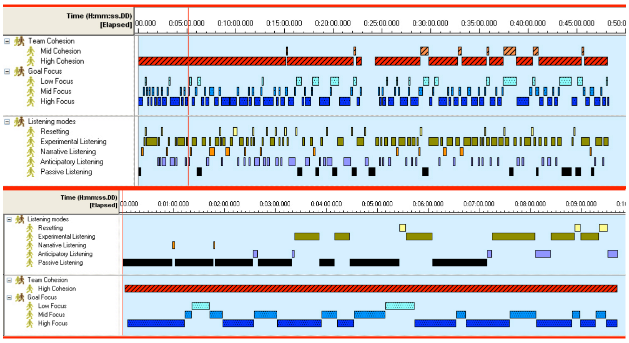


Figure 7. Video coded sections from Team C, level 4.

<u>*Team D, Level 3*</u> In this example, Team D illustrates a mastery of aural fluency. The vignette shows a shared understanding of the listening positions: Even though there is no mention of sound a team member utters: "It got cold…" [*narrative listening*] halfway through it, suggesting that everyone has been playing dormant yet monitoring (listening to) the system's response. When they finally move ahead, a player comments while moving: "ok, this is getting higher…I think that's good" [*experimental listening*] – addressing the rising pitch of the soundscape of bouncing marbles. When the intermediate reward plays, another player immediately takes notice: "hey there's the sound – what are we supposed to do, hold it?" [*anticipatory listening*] – and he pauses for a brief moment before he continues to move like the others (see Figure 8).

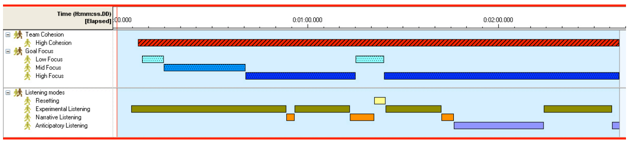


Figure 8. Video coded section from Team D, level 3.

Team A, Level 3: Key to aural fluency is that the listening positions can be seen as progressive levels of competency. Once achieved, they are utilized in a complex interplay as listeners alternate between positions. It is also clear that aural fluency undoubtedly supports collaboration and shared understandings. In this vignette, players in Team A strategize and collaborate together seamlessly supported by their shared aural fluency: Users first notice sound when they stop to strategize and pause to decide on a strategy: “let's do it to the sound of the dice…” [*narrative listening*]. When the intermediate reward comes on they are moving/playing and continue on, making a verbal notice: “oh!..there we go!” [*anticipatory listening*]. When re-evaluating their strategy they do so in motion – “it doesn’t like it that way” (referring to system = sound/light) [*experimental listening*]. Again in motion, they acknowledge the sound’s narrative relationship to the activity: “so how do you do dice? Oh…like rolling the dice I see…let’s just move in circles”… [*experimental listening* and *anticipatory listening*]. After regrouping again to talk about strategy, they start moving and right away comment, “oh there it is...so…warmer, warmer” [*anticipatory listening*]. Another regrouping [*resetting*], and in motion again, they reflect on their embodied configuration and strategize, referring to the sonic feedback as guidance: “I don't know how small a circle we need. When it was making the sound we were much closer” [*experimental listening* and *anticipatory listening*].

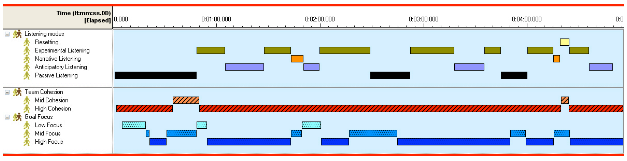


Figure 9. Video coded section from Team A, level 3.

4.3. Stages of Aural Fluency

Based on the listening positions related to listening in embodied interaction environments, and our observations, we further formulate three stages of aural fluency. There is an evident progression through the stages marked by listening positions. It is also important to note that our observations were in a collaborative setting and it is in large part integral to our description of aural fluency and its stages. We believe it was evident in the vignettes the underlying social as well as embodied characteristics of the concept. We feel that the stages provide design parameters that could in turn be used in the generative stages of a prototype to account for different and shifting competency in listening and interpreting sonic feedback of an embedded interaction system.

Stage One: at first users need to understand the logic and actuality of the sonic display response. For example, users need to define for themselves and their collaborators what a constant intensity is, what it means, and how is it affected by their actions, configurations and movements. *Narrative listening* and *experimental listening* are in evidence at this stage, followed by *resetting*.

Stage Two: In this stage of aural fluency users have created a serviceable and shared understanding of what the sound means to them and how it responds to movement on for example a gradient scale. They may however still not know what exactly affects change, or how to interpret granularities of the feedback. At this stage, there is a tolerance for some ambiguity, but there is an ongoing need to stop and reflect on the response of the system. *Experimental listening* and some *anticipatory listening* at an intermediate level with the need for *resetting* are in evidence at this stage.

Stage Three: In this stage of aural fluency, users have acquired enough familiarity with the particular soundscape, approach to intensity and rate of change, in addition to having accomplished a systematic approach to experimentation. They are able to very quickly shift between attempting solutions, strategizing, and experimentation since they can almost immediately tell by listening to the sound whether their new approach is favourable or not. There is less narrative listening and more focussed experimentation and exercising of players’ fluency of the system’s ‘language’. Often in this stage, users can easily reconstruct past attempts and incrementally change them. Ease with *experimental listening*, *anticipatory listening* and the strategic use of *resetting* are in evidence at this stage.

5. CONCLUSION

Ubiquitous computing environments and spaces of embodied interaction increasingly rely on guiding and informing, ambient feedback embedded within the environment and able to serve and inform groups of participants as opposed to individual users. Understanding how people collaboratively familiarize themselves with the meaning, properties and structure of ambient feedback, and designing for various levels of expertise is what we have tried to address in this paper with the notion of aural fluency. The contribution of this paper is in articulating interaction characteristics related to listening positions and acquisitions of the overall skill of listening. Moreover, we focus on evidence of acquiring the skill of interpretation over time – a listening competency we termed *aural fluency*. As articulated here, this concept addresses not only static listening positions as characteristic but also their development overtime, with potential patterns of sequence and alternation during the embodied interaction with the ambient intelligent space. As a result of our current exploration, there are several main criteria that sound feedback for embodied systems must adhere to: it has to support embodied learning (competency building in physical and temporal conditions); and it has to respond to and manage users’ attention (their listening position shifts) and still allow for effective collaboration. From the detailed accounts of playing socio-ec(h)o an emerging pattern points to the experience of learning competencies within our system as an intrinsic motivating factor and source of satisfaction with the interactive experience. The specific conditions that collaborative, problem solving ambient contexts constitute a process of acquiring competency in interpreting feedback over time, in motion, while strategizing. Often understanding sonic feedback is instant and shared, without needing explicit conversation but is supportive of implicit agreement. Users dynamically alternate between different listening positions and it's important that they don't lose or miss crucial opportunities for interpretation of the sound display and potential actions. At one time they may be listening attentively, or analytically, while in another they may be listening associatively, and in yet another – completely embodied and experimentally. These core patterns we feel constitute and elucidate the notion of aural fluency and are a critical part of ambient intelligent system design, as well as auditory display design.

6. REFERENCES

[1] Droumeva, M. & Wakkary, R. (2008) Understanding aural fluency in auditory display design for ambient intelligent environments. In Proceedings *International Conference on Auditory Displays*, 2008

[2] Truax, B. (2001). Acoustic communication (2nd ed.). Westport, CT: Ablex.

[3] Schafer, R. M. (1977). The tuning of the world. Toronto: McClelland and Stewart.

[4] Guilford, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill.

[5] Brewster, S.A., Wright, P.C. & Edwards, A.D.N. (1992). A detailed investigation into the effectiveness of earcons. In

G. Kramer (Ed.), Auditory display, sonification, audification and auditory interfaces. Santa Fe, NM: Addison-Wesley, pp. 471-498.

[6] Walker, B., J. Lindsay and J. Godfrey. (2004). The Audio Abacus: Representing Numerical Values with Nonspeech Sound for the Visually Impaired. Presented at ASSETS '04 in Atlanta, Georgia, ACM Press.

[7] Brazil, E., Fernström, M. & L. Ottaviani (2003). A new experimental technique for gathering similarity ratings for sounds. *Proceedings of the 9th International Conference on Auditory Display (ICAD2003).* pp. 238-42.

[8] Walker, B., & Kramer, G. (1996). Mappings and metaphors in auditory displays: An experimental assessment. Presented at the third International Conference on Auditory Display (ICAD 1996), Palo Alto, California.

[9] Kramer, G. et al. (1999) "The Sonification Report: Status of the Field and Research Agenda. Report prepared for the National Science Foundation by members of the International Community for Auditory Display," International Community for Auditory Display (ICAD), Santa Fe, NM 1999.

[10] Flowers J. H., & T. A. Hauer, (1995) "Musical versus visual graphs: Cross-modal equivalence in perception of time series data," Human Factors, vol. 37, pp. 553 – 569.

[11] Bonebright, T. L., M. A. Nees, T. T. Connerley, and G. R. McGain, "Testing the effectiveness of sonified graphs for education: A programmatic research project," presented at International Conference on Auditory Display, Espoo, Finland, 2001.

[12] Brown, L., Brewster, S. et al. (2003). Design guidelines for audio representation of graphs and tables, *Proceedings of the 9th International Conference on Auditory Display* (ICAD2003), pp. 284-287.

[13] Hunt, A., & Hermann, T. (2004). The importance of interaction in sonification. Proceedings of the Meeting of the International Conference on Auditory Display (ICAD), Sydney, Australia.

[14] Smith, D. R., & Walker, B. N. (2002). Tick-marks, axes, and labels: The effects of adding context to auditory graphs. Presented at International Conference on Auditory Display (ICAD 2002), Kyoto, Japan.

[15] Hermann, T.T., P. Meinicke, and H. Ritter (2000) "Principle curve sonification," presented at International Conference on Auditory Display, Atlanta, GA, 2000.

[16] Flowers, J. H. , L. E. Whitwer, D. Grafel, C, and C. A. Kotan (2001), "Sonification of daily weather records: Issues of perception, attention, and memory in design choices," presented at International Conference on Auditory Display, Espoo, Finland, 2001.

[17] Peres C. P., and D. M. Lane, "Sonification of statistical graphs," presented at International Conference on Auditory Display, Boston, MA, 2003.

[18] Kramer, Gregory, & Walker, Bruce. (2004). "Auditory Displays." In J. Neuhoff (Ed.), Ecological Psychoacoustics. Boston: Elsevier Academic Press.

[19] Nesbitt, K. V. and S. Barrass (2002), "Evaluation of a multimodal sonification and visualization of depth of stock market data. *Proceedings of the 8th International Conference on Auditory Display* (ICAD2002), Kyoto, Japan.

[20] Gaver, W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. Ecological Psychology, 5(1), 1-29.

[21] Bregman, A. (1994). Foreword. In G. Kramer (Ed.), Auditory display: Sonification, audification, and auditory interfaces. Reading, MA: Addison-Wesley.

[22] Ballas, J. (1993) Common Factors in the Identification of an Assortment of Brief Everyday Sounds. Journal of Experimental Psychology: Human Perception and Performance, 19(2), 250-267.

[23] Wakkary, R., M. Hatala, Y. Jiang, M. Droumeva, M. Hosseini. (2008) Making Sense of Group Interaction in an Ambient Intelligent Environment for Physical Play, TEI 2008: Second conference on Tangible and Embedded Interaction, ACM Digital Press

SONOPHENOLOGY : A TANGIBLE INTERFACE FOR SONIFICATION OF GEO-SPATIAL PHENOLOGICAL DATA AT MULTIPLE TIME-SCALES

Steven R. Ness, Paul Reimer, Norman Krell,
Gabrielle Odowichuck, W. Andrew Schloss and George Tzanetakis

University of Victoria
Department of Computer Science and School of Music
Victoria, BC, Canada.

ABSTRACT

Phenology is the study of periodic biological processes, such as when plants flower and birds arrive in the spring. In this paper we sonify phenology data and control the sonification process through a tangible interface consisting of a physical paper map and tracking of fiducial markers. The designed interface enables one or more users to concurrently specify point and range queries in both time and space and receive immediate sonic feedback. This system can be used to study and explore the effects of climate change, both as tool to be used by scientists, and as a way to educate members of the general public.

1. INTRODUCTION

The study of the yearly timing of biological processes is called phenology. Examples of this include when a particular species of tree first flowers in the year, when birds return from their migrations, or when frogs first emerge after winter. It comes from the Greek words for “to show or appear” (phaino) and “reasoning, or rational thought” (logos). It is an ancient field of study and people have recorded this type of information since the dawn of time. For example farmers have kept records about the emergence of their crops from year to year in order to help them determine the optimal time to plant crops in a specific geographic location.

In just the last few years, internet enabled collaborative websites have transformed the collection of phenology data from a discipline where the individual scientist or farmer records their data onto paper for primarily their own research or use, into one where large numbers of amateurs from the general public can all collect and enter their own phenological observations. Sites such as the National Phenology Network [1] and Nature’s Calendar [2] allow citizen scientists to record their own observations about when certain natural phenomenon happen. This type of approach is commonly referred to as Crowdsourcing.

Tangible interfaces are a new metaphor in Human-Computer interactions, where instead of having the user interact with only the screen and keyboard, the person interacts with physical objects in the real world. These types of interfaces can involve cameras, sensors, motors, actuators and displays, and merge the real and virtual worlds into a single unified user interface.

In this paper, we propose to sonify phenological data and let people explore these datasets using a tangible interface. Our proposed system could be used with both historical phenological data and also with the large quantities of crowdsourced phenological data that is just now becoming available. There are several aspects of phenology data that make it a particularly interesting candidate for control through a tangible interface and sonification. Ideally a system for exploring phenology data should allow the specification of both spatial and time range queries in addition to simple point queries i.e render the data from Tokyo and Osaka between 1985-1990. We design a tangible interface based on tracking of fiducial markers that can be used for specification of point and range queries in time and space over a printed map. Of particular interest is the relative timing of different events such as flowering happens earlier in the South than the North. Synchronicity and relatively timing are clearly conveyed in our sonification. We are particularly interested in installation and public outreach environments therefore the sonification has also been designed to be aesthetically pleasing and not intrusive.

2. RELATED WORK

In “The Climate Symphony”, [1] the author presents a sonification of 200,000 years of ice core data in an artistic presentation that is a combination of sonification and story-based narrative structure. Eight sets of time series data of the relative concentrations of a number of ions in this ice core were examined, and using Principal Component Analysis, these time series were reduced to three sets of time series data. These time series data were sonified with simple sine waves which were then amplitude modulated by the amount of ice sheets coverage. Interesting contributions of this paper include the idea that because of the variety of different cycles in global temperatures that are driven by climate forcing from the sun (on time scales of 400,000, 100,000, 40,000 and 22,000 years) there are natural periodicities to this data. By using the natural ability of humans to hear periodic structure in audio signals, this paper demonstrates that this type of data is amenable to sonification. Another important contribution of this paper is that it attempts to create a system that will engage members of the general public by providing an interesting and pleasant way to explore climate data.

In “Broadcasting auditory weather reports - a pilot project”, [2] a sonification system is described that generates a sonified summary of a days worth of weather data which is then broadcasted on a local radio station. The data that is sonified includes time markers, wind, rainfall, temperature, cloudiness, humidity as well as discrete events such as thunder, hail and fog. They then sonify a 24 hour period in a 12 second audio clip, and comment on the different mappings of weather data to sound that they tried. One interesting contribution of this paper was that they found it useful to explore the emotional content of music, and that the authors tried

[1] http://usanpn.org

[2] http://www.naturescalendar.org.uk/

to map pleasant weather events (like bright sunshine) to musical phrases that evoked pleasant emotions, and less pleasant weather conditions (such as rain) to more melancholy musical phrases.

Another related paper is "Sonification of Daily Weather Records" [3], in which the authors describe a system that sonifies the weather data from Lincoln, Nebraska. In this paper, the authors choose three different parameters to sonify, temperature, rainfall and snowfall. For the temperature, they take daily high and low temperature measurements and convert these to MIDI notes. Because of the sizable difference between the high and low notes, this produces a sonification with two independent melodic lines, which humans are able to independently track as separate streams, as previous research by Bregman has shown [4]. They also propose mappings for rainfall and snowfall, for these the authors use one, two and three note sequences to encode different amounts of rainfall and snowfall, for example, for rainfall events less than 0.05 inches, only a single note is sounded, and for rainfall events over 0.5 inches, a sequence of three consecutive notes are played. They chose this mapping in order to follow the metaphor that light rain makes only light plinks and that heavier rain "comes down harder".

In "Atmospherics/Weather works: A multi-channel storm sonification project" [5], the authors describe a system for sonifying the meteorological data associated with weather storms using multi-channel audio. In this paper they present sonifications for two storms, one of which was a typical strong hurricane, and one was an extremely violent storm that was not predicted by existing meteorological models. They choose these two storms in order to test if their sonification of storms could help meteorologists develop insights into the differences between these storms. Besides the very interesting idea of using multi-channel audio to help users understand the data better, they also present ideas for a variety of different sonifications of weather patterns. They first identified a number of variables, including temperature, wind speed and humidity at a variety of elevations, and then did a simple mapping of this data to pitches. Another interesting idea that was employed in this paper was to correlate each geographical point on the map to a speaker and then to use loudness as an indication of wind speed. The authors report that this gave a dramatic spatialization effect to the data.

In the majority of existing system for sonifying scientific data the result of the sonification process is a monolithic audio signal and the amount of influence users have in the sonification process is minimal or non-existent. In contrast in our system we have tried to make the sonification process an interactive, exploratory experience. Our design has been informed by several different research topics: phenology, crowdsourcing, tangible interfaces, and sonification. In the following section we describe these different topics and show how they relate to our work. The resulting system which we call Sonophenology integrates these different influences in a coherent whole.

3. BACKGROUND AND MOTIVATION

3.1. Phenology

Phenology is the study of the timing of biological processes as they occur during various times of the year. The timing of biological processes are intimately linked to the environment in which the organisms exist, and one of the most important determiners of the timing of seasonal changes is the average local temperature. For example, during a warm year, cherry blossoms will flower earlier than they would during a year with a colder spring.

Recently, phenological data has been used in a number of research projects in climate change [6, 7, 8, 9]. In these studies, a general conclusion has been reached that changes in local and global temperature affect the timings of phenological processes, and that these processes are exquisitely precise measures of climate change. Currently these results are typically compared using using either statistical measures, such as the ANOVA (Analysis of Variance) tests such as in Doi [10] or using visual representations of this data such as graphs that show histograms of the timing of various events across years.

Data about winter temperatures have been recorded for the last 2000 years in China [11], and this data has been used to study climate variations. Another set of phenological data that has been used to study climate change is that of Burgundy grapes in France [12]. In this study, spring and summer temperatures from 1370 to 2003 were studied, and using the data from the ripening of this species of grape, it was possible to look at variations in temperature over this time span. These types of studies show that phenology data can be used as a source of proxy data for studying the climate. Karl Linnaeus, the founder of modern taxonomy, studied phenology extensively, and by making observations of the flowering of 18 different plant species across Sweden. In his research, he came to the conclusion that flowering plants are exquisitely sensitive weather instruments.

3.2. Crowdsourcing

Crowdsourcing is a relatively new phenomenon that has been enabled by the pervasive spread of the internet in society, and allows members of the general public to help scientists collect or analyze data. It is a new type of collaboration where non-specialists help expert scientists [13] and has been used to great advantage in a number of research programs [14] [15] [16]. Hong [17] presents results that show that a group of problem solvers with a diverse background can outperform smaller groups of experts.

Whereas it used to be the case that obtaining phenological datasets used to be a difficult and time consuming process, the advent of these websites will mean that there will soon be huge archives of phenological data. One of these sites that has already started to distribute data is the Nature Watch [3] website in Canada. This website has subprojects including IceWatch, PlantWatch, FrogWatch and WormWatch that monitor the timing of various physical and biological processes, including when ice is present, when plants emerge and bloom and when worms and frogs emerge from hibernation.

With the advent of these new crowdsourced sites for the collection of phenological data, the concept of phenology is becoming more well known in the general community. These websites have thousands of observers located in many geographical regions, and with this data becoming available, it can be anticipated that these citizen scientists will want to observe the results of their observations. Currently, results are usually presented in the form of a map with an associated timeline which allows the user to go back and forth in time to observe which plants are flowering at which places over time.

[3] http://www.naturewatch.ca

3.3. Tangible Interfaces

Tangible computing interfaces using tokens detected by computer vision techniques, such as the reacTable proposed by Kaltenbrunner, Jorda, and Geiger [18] have been tailored specifically for designing multimedia processing algorithms. The shape, translation, and rotation of tokens placed on a planar desktop surface controls some aspect of a multimedia processing pipeline. Early versions of these interfaces had an audio focus, to complement the visual process of designing an audio processing interface (e.g. an musical instrument). Tokens designed specifically for detection, classification, and spatial location/orientation are known as fiducial markers.

Fiducal marker detectors and trackers operate by identifying fiducials in a video frame, based on information that is known a-priori. Several visual properties can serve to identify a fiducial marker, (e.g. colour, geometry); several popular, state-of-the-art detectors use fiducials designed and identified by the topology of a hierarchy of shapes contained within the fiducial design, as described by Costanza and Robinson in [19].

Costanza, Shelley and Robinson[20] describe the application of this approach to detecting fiducial markers via the use of a region adjacency graph (RAG) to encode a two-level topology (e.g. black and white) of binary shapes into a tree structure representing that shape. Several constraints are imposed on marker designs by this choice of detector; markers must consist of white shapes wholly surrounded by black shapes, which in turn may enclose another level of black shapes. Detectors work on a binary-thresholded version of the input image, which allows some variation in the detected colour (e.g. off-white, near-black). A region adjacency graph can be generated for any number of levels, but in practice the number of levels is limited to three, denoted in [20] as root, branches and leaves.

Bencina et al. improve on the topological fiducial detector in [21], where the centroid of clusters of shapes contained in a lower level of the topology are used to rapidly reject candidate fiducial matches that do not conform to the structure of expected fiducials, and use this centroid information to discriminate between different fiducials.

Using a fiducial detector based on marker topology presents a tradeoff between marker complexity (and hence increasing size of the marker at the same level of resolution), and number of possible markers represented by different topologies of the same size.

Tangible interfaces based on positioning multiple fiducial markers placed on a multitouch table or desktop surface have many advantages over a conventional interface using a keyboard a mouse. These interfaces are a pure direct-manipulation modality: the user can intuitively see the structure they have created. Physical controls for parameters, visual representations of those parameters, and visualizations of the output produced by each processing unit, are located spatially nearby the fiducial token. The display plane and the control/interface plane are often aligned, preventing confusion common from a mouse/screen arrangement. Affordances are offered in multiple dimensions, for each fiducial marker detected (i.e. marker id, position, and rotation). This implies a simple marker printed on paper can yield more information than a dedicated, wired, peripheral such as a computer mouse. Indeed the costs of fiducial tracking hardware are little more than the cost of a conventional webcam and a printer for producing fiducial markers.

Each marker can be positioned by a separate person, and so marker-based interfaces lead easily to collaborative interfaces, since multiple people can use the same desktop surface with a separate collection of markers, or become more productive in assembling a single algorithm using multiple simultaneous operators.

In addition to using fiducial markers as physical controls independent from other markers, commonly a system of rules is designed to relate multiple fiducial markers present on the same desktop. Possible interactions include varying parameters of one or more markers, varying processing steps of one or more markers, or establishing an application specific chain of markers which interact in a pre-determined way.

We extend the concept of a tangible, fiducial marker-based interface used to create an aesthetically pleasing, and usable environment for exploring phenological data.

3.4. Sonification

Sonification can be described as the use of audio to convey information. In other words, scientific data is represented not as a visualization, like a graph, but instead as a collection of sounds that are played at different times.

The manner in which a given set of data is mapped to audio is a challenging problem, there are an infinite number of ways to transform data into a sonification[22]. Many aspects of any sound can be modified: we can perceive changes in amplitude, pitch, timbre, directional, and temporal information. Any of these auditory aspects, or audio parameters, can be modified by a data set. The best choice when selecting audio varies, depending on the content of a given set of data. The direction, or polarity, of the datasets that are being compared can also affect the perception of a sonification. For example, temperature is often described aurally as a tone with increasing pitch [23]. The scale of the relationship between a one-dimensional data set and the audio parameter modified by that data must also be considered. If we consider the temperature to pitch example, we must consider how quickly will the pitch increase, and whether the relationship will be linear or non-linear [24], that is to say, one wants to preserve the ratios, not the differences in frequency. The aesthetics of sonification are also an important consideration. The goal is to create a collection of sounds that represents a dataset accurately, and is also pleasing to listen to.

4. SYSTEM DESCRIPTION

4.1. Overview

Our system consists of a number of separate sub-components that interact together to provide a tangible interface for the exploration of geo-spatial phenological data. The overall organization of this system can be seen in Figure 1.

The phenological data sources that we obtained for this paper are quite diverse, and contain various types of information that could be used for sonification. In this application, we constrained our analysis to include only the species name, the latitude and longitude of the observation and the date when this observation was taken. Other data that we are not using for this paper includes the type of observation, for example, was the observation of the first bloom of the lilacs or when they were in full bloom. Many of the observations also include comments from the observers. These additional sources of data could be used in the future to enhance the audilization and visualization in our interface. We first sanitize

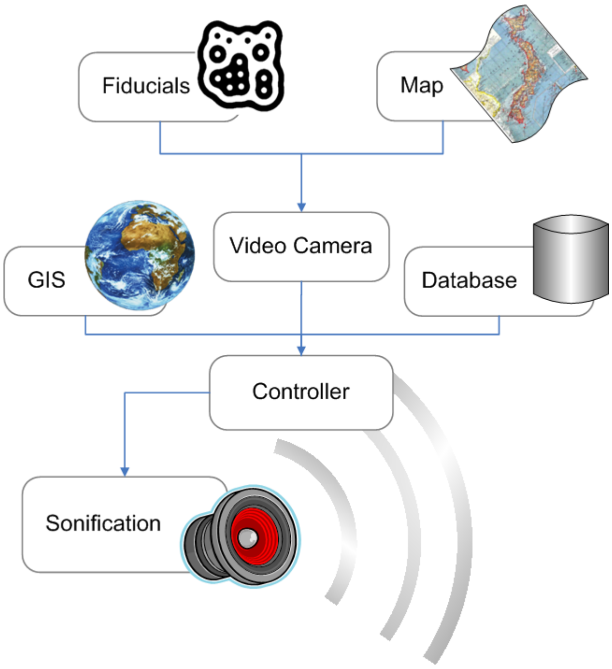


Figure 1: A flowchart of the system organization of our system. This system has at its core a Controller module that communicates with the phenology and GIS databases as well as the video camera and sonification engine. It generates sonifications by tracking the positions of fiducials on a printed paper map.

these data sources and read them into our GIS-enabled database system.

The second section of our system is the fiducial tracking interface. This uses a consumer grade webcam and tracks pre-printed fiducial markers on a surface. We also use fiducial markers to determine the position and orientation of the physical map underneath the fiducial markers. We then create a mapping from the set of coordinates of the fiducial markers to the physical latitude and longitude on the map. When the user places fiducial markers on the map, this system then takes the latitude and longitude of these points and queries the database to obtain corresponding phenological data points.

The final step in this system then involves taking these phenological data points, which include latitude, longitude, species and observation date, and sonifying them.

4.2. Phenology - Japan lilac

For this project, we are concentrating first on a set of observations of the flowering of the common purple lilac *Syringa vulgaris* in Japan [25]. Observations on the flowering of this species were collected from 1996 until 2009. Because of the large difference in latitudes between the south and north of Japan, flowers bloom earlier in regions in the south of Japan before they do in the north of Japan. These types of geographical differences are one source in the variation of flowering times. Another difference that may be possible to observe is the effects of climate change on the flowering times of these lilacs, however to truly see effects of climate change, one must of course examine temperature records over longer time spans, on time spans of centuries to millenia. If average temperatures increase over a period of years, one would expect that the phenological processes that respond to temperature would tend to move to earlier times in the year.

4.3. Tangible interface

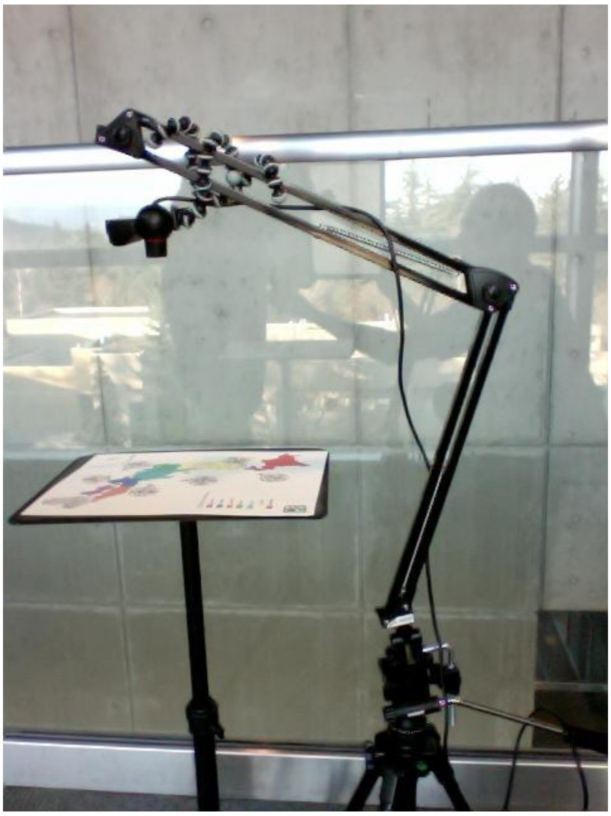


Figure 2: Shown above is a picture of the fiducial tracking interface. Above the computer monitor is a small consumer grade video camera, which is pointed downwards in order to view the fiducial markers which are placed on a printed paper map.

While it would be possible to develop a simple desktop or web-based interface to explore a sonification of this data, a much more intuitive and engaging interface could be a tangible interface, where users interact with a physical interface. We have chosen a fiducial based tag tracking system previously used in the reacTable [18]. A picture of this system is shown in Figure 2.

This interface is inexpensive and easy to deploy, requiring only a consumer-grade webcam, physical printed map and printed fiducial tags, and could be easily deployed within a classroom setting. With such a system in a classroom, a teacher could teach students not just about phenology, climate change and maps, but also about new systems for physical interaction with computers. By moving markers across the map, the students experience a direct correlation with the location of the marker on the map and the associated phenological data.

In order to generate the query of the GIS database that contains the phenology data, we use two different user-interface metaphors. The first, simple method, is to simply use the center of the fiducial

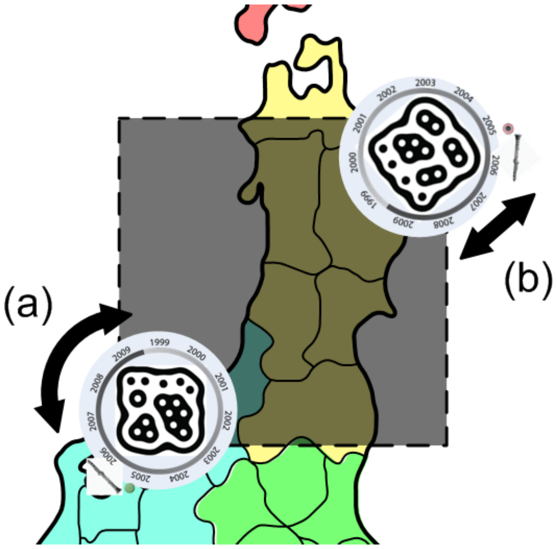


Figure 3: (a) Rotating the fiducials will change the sonification range. (b) Translating a pair of fiducials will sonify all data points found within the enclosed area.

as the latitude/longitude search point and to return all data points that lie underneath that fiducial. A more complex setup that we have also implemented allows users to select a region of the map and a time range for each region. In this scheme, regions are created by placing two fiducials on the map. The first fiducial specifies the top left corner of a bounding box and the second fiducial specifies the bottom right corner. To change the time range that is sonified, the system calculates the relative rotation angle between the two fiducials and maps this to a value of years. This setup is demonstrated in Figure 3.

4.4. Sonifications

There are a number of advantages to sonifying these phenological data over using statistical tools and visual graphs. One advantage is that by using different timbres to represent the different sections of the map that we are sonifying, we take advantage of the fact that humans can distinguish different melodic streams that are rendered in parallel by different timbres. This could potentially allow a user to follow many different lines of data at once. This technique becomes even more powerful because of the distributed geographical and temporal nature of the phenological data, where flowers in the south bloom earlier than flowers in the north. These different melodic lines start and swell at different times, and the combination of different timbres with different start times of these timbres make it even simpler for users to follow the progression of phenological events.

Our primary sonification metaphor is that of a step sequencer, which uses a fixed two-dimensional grid consisting of quantized steps, with the horizontal axis representing time and different steps on the vertical axis being different instruments, or different pitches of one instrument. In our system, the vertical, or pitch axis, corresponds to different years, and the horizontal, or time axis, corresponds to the timing of the phenological event in days since the start of the year. This system allows us to easily hear and compare changes in the timings of different events over years by listening to the organization of pitches. If a phenological event occurs on the same date each year, one would hear a chord of all the notes at the same time. If on the other hand, the date of a phenological event becomes earlier each year, one would hear a descending arpeggio of notes.

The comparison of phenomena over various years is an essential part of this system, as one a primary motivation of this project is to provide a way for people to not just see but also hear and explore the effects of climate change. These different modalities of experience might prove effective in the education of people about phenology and climate change.

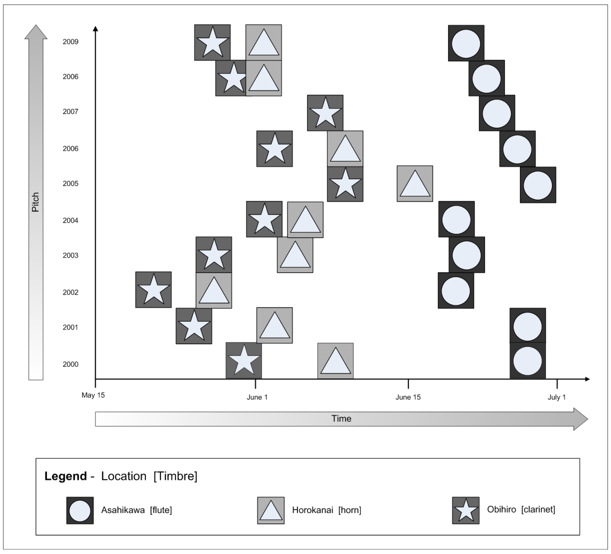


Figure 4: A graphical representation of 10 years of flowering data for the common lilac in three locations in Japan. The three different locations are depicted by different shapes, a circle, a triangle and a star. From this diagram, one can see that there are certain years (2002-2004) in which flowering occurred earlier than in other years.

One mapping that we have found to be useful is a step sequencer. In our system, one axis of the sequencer has pitches that correspond to different years, where earlier years have lower pitches, and later years have higher pitches. On the other axis of the step sequencer we have the day of the year. We also then map each fiducial marker to a different musical instrument or timbre. With this mapping, if the flowers in a specific region all flowered at the same time, then one would hear the notes from all the years sounding at once. On the other hand, if the flowering dates occur at later times each year, one would hear an arpeggio of notes with increasing pitches. A graphical view of three observation locations over a time period of 10 years is shown in Figure 4.

Another mapping that we are exploring is to instead represent each phenological observation as a distinct sonic event. This type of sonification produces a radically different soundscape which is more textural and ambient. One can imagine what this sonification sounds like by thinking of the timing of blooming of plants in the spring. One will often see one or two different plants of a species flower, then as time goes on more plants will flower in almost an exponential fashion until all the plants of the species have flowered. If one were to sonify each of these events as an impulse sound, then the sonification of this data would sound something like the popping of popcorn. What is interesting in this method is that it allows us to perceive the "stochastic" nature of the natural process, where each event is not significant unto itself, but the aggregate events outline a process that can be reflected in an auditory soundscape that reveals subtle differences in the rate of change of a physical system. Our ears are very sensitive to subtle differences in stochastic signals like colored (or filtered) noise.

When converting data into audio, there are a number of different mappings that can be used. The simplest would use a sinusoidal oscillator and would linearly map input data into the frequency of this oscillator. One disadvantage of this mapping is that in the human auditory system, the frequency to pitch ratio is not linear but rather is logarithmic. Because the human ear hears frequencies logarithmically, a logarithmic mapping of data to frequency would more accurately preserve the ratios of data points to each other. There are a potentially infinite number of mappings of data to pitch values, the one that we chose for this application was to map data values onto the equal tempered scale, as seen on the piano keyboard or MIDI note values. However, in our system, we anticipate that several values could occur at one point in time, and if we were to simply map data values to MIDI note values, it would be common to encounter dissonances in simultaneously played notes. To overcome this, one can use different scales or chords instead of the chromatic scale. In our system, we mapped the 10 different year values to the pentatonic scale. We are also developing mappings using chords, for example the notes of a C-sharp major 7th chord, or any other chord, could be used to map each year to a pitch component of the chord. Then the chronological order would determine the position of the year within the chord - in our example chronological order follows pitch height. One could use the same chord for all instruments with or without the same keynote, however, one could also use different chords for different instruments, which might have the advantage that it would further improve distinguishability for people by different melodic lines following the chords.

For most of our work in this paper we have used sampled sounds from the RWC dataset [26]. However, we have also implemented a synthetic instrument model in order to provide more and different sonification parameters. In doing this, we have implemented simple sine sources, plucked strings as well as more advanced synthetic models of physical instruments. The advantage to using these types of synthesized sounds is that it is possible to control different parameters of the sound, for example, the brightness of a clarinet sound, or attack speed of a trumpet, these parameters can then be mapped to the data that is being sonified. Using synthetic instrument models, one could also generate timbres that are intermediate between two instruments, for example, one could make a sound that was half-way between a clarinet and saxophone. This type of fine-grained control is difficult to implement using pre-generated samples. The main disadvantage to using synthetic instruments is that the models are often quite elaborate and are computationally expensive, which limits the amount of simultaneously playing instruments.

5. CONCLUSIONS

In this paper we have presented a system that takes geo-spatial phenology data and allows users to interact with it using a tangible interaction metaphor. The dataset of the flowering dates of Japanese lilacs was a useful dataset to explore with this system as it contained data points of flowering dates that occurred at different times and in different locations from the northernmost to the southernmost areas in Japan.

We have explored this dataset with our system. We have observed a number of interesting properties of the data and of the system. One interesting observation about this data is that in certain years the flowering of trees occurs earlier, and in some years they occur later. This is clearly heard in the sonification of this dataset because in these years, the note that is played for the different instruments is the same, and is repeated earlier in the cycle than those notes from other years. Another observation is that for the data points that occur earlier in later years, a descending arpeggio is indeed heard.

With the inclusion of the tangible interaction interface, this system is quite approachable for members of the general public, and in the few number of interactions that these individuals have had with our system, they find it both interesting and easy to use. We are currently considering doing user studies with this system, with the goal of building a system to help educate students and the public about climate change with an engaging interface.

In future work, we would like to develop a similar system to the one described in this paper but for mobile devices, such as the iPhone. This interface would allow people to interact with a computer generated map of a region, for example, a map of Japan and would allow people to explore the timing of various phenological events on their own personal mobile device. In conjunction with this, we are building a web-enabled version of this app using a combination of Flash and HTML5 technologies. The advantage of these web based and iPhone based applications is that they could have much wider penetration into the general community, at the cost of a more limited interaction metaphor.

This system can also be used for other phenology datasets, and as websites such as the National Phenology Network and Nature's Calendar start releasing their crowdsourced data, we anticipate that there will be a huge amount of phenological data that would be interesting to sonify. In addition, this system could also be used with other geo-spatial datasets, for example, one could develop an interface to allow scientists to sonify the amount and type of ground cover as determined by satellite images.

Although this system was developed as a tool to be used in a single location, in the future we would like to extend it to allow for remote collaboration between scientists. In this system, scientists in different cities could each have their own map, camera and fiducial markers. The fiducial markers would be mapped to unique instruments, so that for example, one scientist could use fiducial markers that correspond to different timbres. By exploiting the ability of humans to do auditory stream recognition, each scientist could choose to either focus on the sounds from the instruments that they are controlling or could focus on sounds that are being generated by a query from the fiducials of the other scientist. This type of multiple user interaction paradigm is often challenging when using visual interfaces because of problems of occlusion, reach and grasp, and could be more intuitive and easy to understand when using sonification instead.

We have made a website[4] that presents visualizations and sonifications of the data used in this paper, along with videos showing the system in action.

[4] http://sonophenology.sness.net

6. ACKNOWLEDGEMENTS

We would like to thank the National Phenology Website and Dr. S. Funakoshi for making the Japanese lilac dataset available to the scientific community.

7. REFERENCES

[1] M. Quinn, "Research set to music: The climate symphony and other sonifications of ice core, radar, dna, seismic and solar wind data." Laboratory of Acoustics and Audio Signal Processing and the Telecommunications Software and Multimedia Laboratory, Helsinki University of Technology, 2001, pp. 56–61.

[2] T. Hermann, J. M. Drees, and H. Ritter, "Broadcasting auditory weather reports - a pilot project," E. Brazil and B. Shinn-Cunningham, Eds., 2003, pp. 208–211.

[3] J. H. Flowers, L. E. Whitwer, D. C. Grafel, and C. A. Kotan, "Sonification of daily weather records: Issues of perception, attention and memory in design choices," J. Hiipakka, N. Zacharov, and T. Takala, Eds., Espoo, Finland, 2001, pp. 222–226.

[4] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Massachusetts, 1990.

[5] A. Polli, "Atmospherics/weather works: A multi-channel storm sonification project," S. Barrass and P. Vickers, Eds., 2004.

[6] E. Post and N. C. Stenseth, "Climatic variability, plant phenology, and northern ungulates," *Ecology*, vol. 80, no. 4, pp. 1322–1339, 1999.

[7] P. J, I. F., and P. Comas, "Changed plant and animal life cycles from 1952 to 2000 in the mediterranean region," *Global Change Biology*, vol. 8, no. 6, pp. 531–544, 2002.

[8] G. Walther, E. Post, P. Convey, A. Menzel, C. Parmesan, T. Beebee, J. Fromentin, O. Hoegh-Guldberg, and F. Bairlein, "Ecological responses to recent climate change," *Nature*, vol. 416, 2002.

[9] A. Menzel, T. H. Sparks, N. Estrella, and D. B. Roy, "Altered geographic and temporal variability in phenology in response to climate change," *Global Ecology and Biogeography*, vol. 15, no. 5, pp. 498–504, 2006.

[10] H. Doi and I. Katano, "Phenological timings of leaf budburst with climate change in japan," *Agricultural and Forest Meteorology*, vol. 148, no. 3, pp. 512 – 516, 2008.

[11] Q. Ge, J. Zheng, X. Fang, X. Zhang, and P. Zhang, "Winter half-year temperature reconstruction for the middle and lower reaches of the yellow river and yangtze river, china, during the past 2000 years." *Holocene*, vol. 13, pp. 933–940, 2003.

[12] I. Chuine, P. Yiou, N. Viovy, B. Seguin, V. Daux, and E. Le Roy Ladurie, "Grape ripening as a past climate indicator," *Nature*, 2004.

[13] J. Howe, *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business, 2008.

[14] J. Surowiecki, *The Wisdom of Crowds*. Anchor, 2005.

[15] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.

[16] J. Travis, "Science and commerce: Science by the masses," *Science*, vol. 319, no. 5871, pp. 1750–1752, 2008.

[17] L. Hong and S. E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers," *Proceedings of the National Academy of Sciences*, vol. 101, no. 46, pp. 16 385–16 389, 2004.

[18] S. Jordà, G. Geiger, M. Alonso, and M. Kaltenbrunner, "The reactable: Exploring the synergy between live music performance and tabletop tangible interfaces," in *Proceedings Intl. Conf. Tangible and Embedded Interaction [TEI]*, 2007.

[19] E. Costanza and J. Robinson, "A region adjacency tree approach to the detection and design of fiducials," in *Video, Vision and Graphics*, 2003, pp. 63–69.

[20] E. Costanza, S. B. Shelley, and J. Robinson, "Introducing audio d-touch: A tangible user interface for music composition," in *6th Intl. Conference on Digital Audio Effects*, no. DAFX-03, 2003.

[21] R. Bencina, M. Kaltenbrunner, and S. Jorda, "Improved topological fiducial tracking in the reactivision system," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops*, 2005.

[22] T. Fitch and G. Kramer, "Sonifying the body electric: Superiority of an auditory over a visual display in a complex, multi-variate system." in *Proceedings of the First International Conference on Auditory Display*, 1992.

[23] B. N. Walker, "Consistency of magnitude estimations with conceptual data dimensions used for sonification," *Applied Cognitive Psychology*, vol. 21, pp. 579–599, 2007.

[24] B. N. Walker, G. Kramer, and D. M. Lane, "Psychophysical scaling of sonification mappings," P. R. Cook, Ed., International Community for Auditory Display. Atlanta, GA, USA: International Community for Auditory Display, 2000. [Online]. Available: Proceedings/2000/WalkerKramer2000.pdf

[25] S. Funakoshi and F. Kanda, "Phenological study of common purple lilac flowering as a study program for environmental education," *Journal of Environmental Education*, vol. 3, pp. 77–81, 2000.

[26] M. Goto and T. Nishimura, "RWC music database: Music genre database and musical instrument sound database," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2003.

AUDIO FRACTAL

Joachim Gossmann

UC San Diego,
Center for Research and Computing in the Arts,
9500 Gilman Drive La Jolla, California 92093-0037
jgossmann@ucsd.edu

ABSTRACT

"Audio Fractal" is an interactive sonification strategy for Escape Time Fractals implying the listener as a recipient of sinusoidal auditory information quanta–an idea correlated with D. Gabor's concept of Acoustical Quanta and the inverse application of Fourier’s Analysis (a.k.a. Additive Synthesis). The volatility and ephemerality of auditory impressions is suspended by creating a relationship of accountability between the sound and positions within a persistent visual representation of the fractal on the explorable 2d-plane. The auditory formations demonstrate transitions between chaos, order and self-similarity as well as the perceptual inexhaustability of timbre space. The iteration path inherent to each and every pixel of the visual escape-time-fractal produces a distinct individual additive spectrum revealing properties and relationships not perceptible from the visual representation, allowing the emergence of a distinctly auditory perspective on the underlying structure.

The current version of Audio Fractal, which was first presented in 2004, now allows an expanded vocabulary of exploratory approaches such as comparison and juxtaposition as well as automation of the spatial exploration.

1. REFERENCES

[1] J. Gossmann, “*Towards an Auditory Representation of Complexity*,” Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, Limerick, Ireland, July 6-9, 2005.

[2] D. Gabor, “*Acoustical Quanta and the Theory of Hearing*,” Nature, vol. 159, 1947, pp. 591-594.

[3] R.N. Bracewell, “The Fourier transform,” Scientific American, vol. 260, Jun. 1989, pp. 86-89, 92-95.

A FRAMEWORK FOR MUSIC-BASED INTERACTIVE SONIFICATION

Nuno Diniz, Alexander Deweppe, Michiel Demey and Marc Leman

Institute of Psychoacoustics and Electronic Music
Ghent University
Ghent, Belgium
nuno.diniz@ugent.be

ABSTRACT

In this paper, a framework for interactive sonification is introduced. It is argued that electroacoustic composition techniques can provide a methodology for structuring and presenting multivariable data through sound. Furthermore, an embodied music cognition driven interface is applied to provide an interactive exploration of the generated music-based output. The motivation and theoretical foundation for this work are presented as well as the framework's implementation and an exploratory use case.

1. INTRODUCTION

The development and application of processes that allow the transmission of information using sound has always been a main concern of music composition practice. Particularly in the 20th century, several theories have been suggested for establishing a meaningful and coherent binding of individual sound streams or events. However diverse these approaches might be, they all address the same problem: how to establish a unified context between hierarchical levels of communication that are exposed simultaneously through time. As in music, it is relevant to take this problem under consideration when presenting multivariable data through sound. For illustration purposes, consider the situation where three variables are sonified at a given moment with the C, E and G musical pitches. The presence of a higher level of meaning (a major chord) as well as the intermediate ones (such as the intervals formed by the combination of the individual elements in the pitch set) should be taken into account with the same degree of importance as the individual pitches.

Therefore, the work presented here is focused on the design of a framework which provides a simultaneous encoding of such interrelated levels.This process is a key element in the definition of structures that allow the constitution of contexts in sound data presentation. Furthermore, it is argued that a system that proposes a scalable approach to content should include an interface that provides both a top-down and bottom-up inspection perspective from the outset in order to facilitate the interactive access of these musically structured levels.

In the following section, an overview of some compositional views of Pierre Schaeffer and Karlheinz Stockhausen is discussed in order to establish a relation between musical composition practice and multilevel sound communication. Afterwards, we present the motivation underlying the use of embodied music cognition theory as an interface paradigm for interactive sonification, followed by the main concepts concerning virtual object-based mediation. Then, the framework's design, the technological aspects and the evaluation of an exploratory use-case are addressed. Finally, a discussion of the present work is provided.

2. MUSICAL COMPOSITION AND MULTILEVEL SOUND COMMUNICATION

The application of music-based approaches in non-speech sound communication has been present in the auditory display since the early stages of this research field as, for example, documented in [1, 2, 3]. Furthermore, there has been an evolving interest in on how the compositional processes can be adopted in the sonification domain. By underlining its systematic validity, aesthetic added value and capacity to generate context, the focus has been on how some of these techniques can help provide design options for improving the perceptual cognition of sonification processes' output [4, 5]. Nevertheless, a brief review will follow of some compositional theory and processes that constitute the point of departure for this work.

In the two main initial trends in electroacoustic music, the french Musique Concrete and the Electronic Music from Cologne, the search for ways of establishing relations between material and form is present in the theoretical and compositional production of their leading advocates, Pierre Schaeffer and Karlheinz Stockhausen. According to Michel Chion's Guide to Sound Objects, the sound object, as defined by Schaeffer, is perceived as an object only in an enclosing context. This dependency between individual and group is further developed in the sense that "every object of perception is at the same time an object in so far as it is perceived as a unit locatable in a context, and a structure in so far as it is itself composed of several objects" [6]. One can extract from such postulates that the dialogue condition that is imposed to the sound object and the structure holds a dynamic perspective shift, that reassures the relationship between these two concepts. From his part, Stockhausen's concept of unity addressed the possibility to trace all musical parameters to a single compositional principle [7]. This concept envisioned the unified control of the musical structures in a given work through the establishment of inherent relationships between the micro and the macro level of the musical discourse. Initially driven by the aims of integral serialism, his search for such mechanisms of scope transposition continued throughout his career. Of such techniques, one can highlight "moment form", a structuring paradigm based on a non linear distribution of "gestalts" known as moments, or the "formula composition", in which all aspects of a given work derive directly or indirectly from an initial short composition. As an example, his over twenty-nine hours long opera cycle "Licht" is based on a three-part, eighteen-bar only score formula.

Although, as argued by Vickers, one can establish a close relationship between sonification and musical composition through a perspective shift [8], it is surely arguable that these concepts can be fully applied outside the art and music realm. Nevertheless, it is our claim that they can encapsulate a set of guidelines that can be of service in functional sound based communication, as defined in [9]. As Delalande pointed out, there is a communality of processes in electroacoustic composition practice that concern the relationship between singularity and regularity of events used in the musical discourse which underlines their structural dependencies [10].

As such, the aim of this work is to transpose the above mentioned compositional concepts to the interactive sonification domain and apply the relationships between material and structure to the micro and macro sound levels of data presentation. As a result, functional contexts are generated by data-dependent hierarchical levels that still preserve their informational identity and significance. As highlighted in the work of Scaletti concerning the specification of the Kyma environment [1], the adoption of Schaeffer's concept of sound object as a base structuring concept is a fundamental design directive for allowing the manipulation of multiples levels of complexity under one unifying abstract structure [11]. On the other hand, and in agreement with Childs, the application of these techniques should synthesize these data structures in such a way that the information transmitted is not cluttered by the presence of non functional musical elements and conveyed to the user in a clear and effective way [12].

Given this conceptual perspective, the next section will present the argumentation for the need and advantages of incorporating an embodied music cognition driven interface in the presented framework in order to more efficiently connect the bi-directional top-down/bottom-up processes of human cognition [13] to this scope variation.

3. INTERACTIVE SONIFICATION AND EMBODIED MUSIC COGNITION

As defined by Hermann and Hunt, interactive sonification is "the use of sound within a tightly closed human computer interface where the auditory signal provides information about data under analysis" [14]. Being so, several questions immediately arise, namely, "how to make these multiple levels of sonification both perceivable and meaningful to the user in order for him to take full advantage of their interrelated nature?", and additionally, "how should this composed information be made available in such a way that the user can interactively manipulate it?".

In order to address these issues, the proposed interface for interacting with the framework's musically structured output follows an embodied music cognition perspective [15]. In electroacoustic music, the concept of musical gesture as materialization of the composer's inner musical intention has always been present at different levels of conception, both within the non-realtime compositional and realtime performance levels. For example, one can point out the expressive use of the mixing board's faders in both the mixing and spatialization processes. It is a trivial but nevertheless good example of an embodiment-based discourse that incorporates the physical factor in the creational process. Even more interesting is the fact that this process can be used across different levels of granularity throughout the work, ranging from individual amplitude envelope of a sound object to post-production panning of entire sections. This process is in tune with the concept of variable resolution [16] in which the hierarchical nature of the human motor system allows a context coherent variation in the resolution of performed actions.

Thus, it is only natural that this architectural similarity between sound objects and gestural/physical behavior is included in the framework's interface design. In other words, an interface that stimulates a user centered approach should be adopted from the start of the design and implementation process, in order to address the impact of context on identification of data structures. With the objective of promoting a fruitful dialogue between the user and the data, an approach based on the expansion of the mediating role of the body through virtual entities is considered within an immersive environment (in relation to [17]). First, virtual objects can act as mediators representing multilevel mapping layers that conform with the premise of a hierarchical object oriented decomposition of sound entities. Second, through the immersion of the natural communication tools of the actors involved, a virtual reality based framework presents itself as an appropriate setting for the investigation and development of interfaces between body and music. More details concerning this methodology will be provided in the next section. In summary, by enabling a configurable location and form representation of the data in space, this methodology invites the user to a physical approach for the inspection process through a shared space of multilevel interaction. As such, an embodied cognition approach is expected to further enable a perceptual link between the data under inspection and the semantic high level representations of the user.

4. VIRTUAL REALITY AS AN INTERFACE FOR SONIFICATION

In this section, the main concepts regarding the virtual object's role in the data exploration are introduced: the inspection window and the inspection tool (Figure 1).

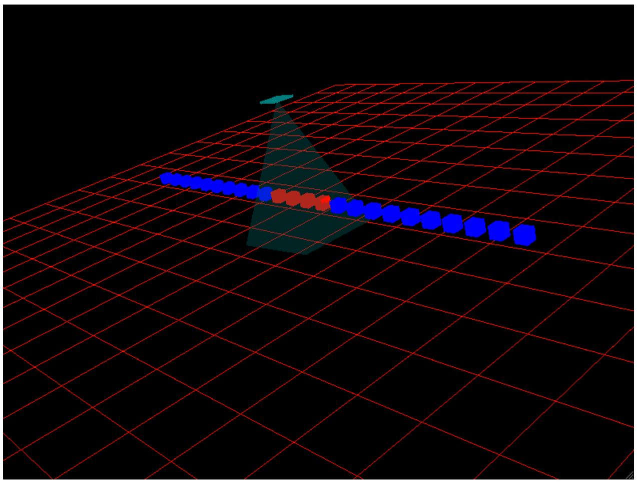


Figure 1: Visual feedback illustrating the inspection window and the inspection tool controlled by the user's hand.

[1] http://www.symbolicsound.com/

4.1. Virtual objects as metaphors for data representation

As mentioned in the previous section, the interface paradigm adopted is based on the interaction with virtual objects, which constitute access points to the variables belonging to a target dataset. To be able to represent datasets of $N > 3$ order in a tridimensional environment, these objects represent an inspection window to the variable's values. For illustration purposes, consider the following example. A dataset containing last year's average day temperatures in 5 cities of a given country is to be sonified. Instead of representing every value for every variable simultaneously, which would easily cause a congestion of the virtual scene and, consequently, difficulties to the inspection process, the representational virtual objects can be configured to allow simultaneous access to a subset of these values. For instance, an array of 30 spheres can be assigned to each variable constituting a temporary access to a period of approximately 1 month. The remaining values of each variable can then be accessed through sliding of the inspection window, which can be controlled by an auxiliary device (Ex. a WiiMote). Furthermore, the 5 arrays can then be placed in various arrangements in order to allow multiple views of the dataset's content. Besides the previously described advantage, this approach constitutes a viable option in the analysis and comparison of real-time data. The values can be made available to the user for a certain amount of time (dependent on the generation rate) and then "hidden" from him, being available for later inspection. Moreover, the morphology of the virtual object(s) that is assign to represent a variable in space can be data dependent, conveying a more informative visual and spatial representation of the values being analyzed.

4.2. The virtual inspection tool

The inspection procedure is conveyed to the user through the inspection tool. This virtual object, composed by a flat surface (that visually represents the user's hand in the virtual space) and an inspection volume, allows the user to interactively investigate the data. This inspection volume behaves as a sonic magnifying glass or virtual microphone, allowing the user to zoom in and out in order to investigate either one element's output, or its relationship with other members of the set. This interaction mode was strongly inspired by Stockhausen's Mikrophonie I [19], composition where the active use of microphones is a base concept in the performance of the piece. Each independent virtual sound source is activated through collision detection when the inspection volume intersects the virtual objects. Then, the activated items are fed into the sonification levels responsible for calculating the respective sonic outputs according to their specific implementation. At this point, the virtual objects, their structure and their relationship with the sonification layers are addressed. Going back to the example described in the last subsection, only the individual virtual elements that compose the inspection window are subject to sonification procedure (i.c. the day's average temperature). As mentioned, the inspection window (the parent object) is composed of 30 spheres (the child objects). Here, the manipulation of the parameters involved in the sonification comes into play. As one gets the inspection tool closer to the activated elements, the distance between them has an effect on the amplitude and depth of the reverberation in the sonification process. As the distance to the user's hand is reduced, loudness increases and reverberation's depth decreases. Although this behavior is sonically implemented by the individual elements, it conveys information about the activated set as a whole. Following the previously referenced theoretical guidance of Schaeffer and Stockhausen, it stimulates a perceptual interpolation between the whole (a month) and the individual nodes (the days). Furthermore, through the use of the inspection tool and the spatial arrangement of the inspection window's elements, the user can group several consecutive "day" and have a "on the fly" composed sound object which conveys the progression of the temperature in one city. On the other hand, several "day" from different cities can be grouped, sonically illustrating the relations between different locations' temperature. Being so, the adopted interface paradigm can convey multiple perspective views between different levels in the form domain, by representing the evolution of the sound object in time, and in structure, by establishing and comparing different groups of N variables (Figure 2).

The framework's implementation and an exploratory use case are the subject of the following sections.

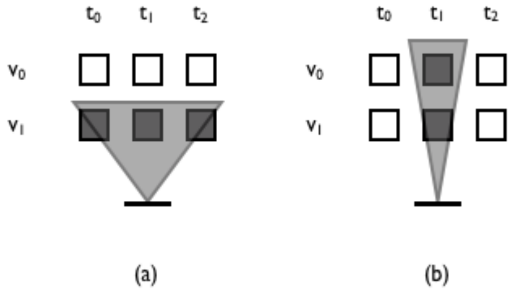


Figure 2: Interactively grouping variables' values in form (a) and structure (b)

5. FRAMEWORK

5.1. Introduction

The present framework's ongoing implementation is a consequence of the need for generic data sonification tools for research and application [18]. As such, it aims to provide a software infrastructure where issues concerning portability, flexibility and integrability are addressed. Furthermore, given the current state of the art in frameworks for sonification, as presented in [5], the framework's design reflects the authors' perspective for the need of incorporating, together with compositional guidelines presented in Section 2, the user's interaction mechanisms as a fully integrated structural paradigm, in accordance with Section 3.

The following subsections address the framework's architecture and the choice of Java as its core technology.

5.2. Architecture overview

The design is based on a functional division of modalities into individual branches around a virtual scene representation. Following a top-down approach, a first level is composed of abstract managing cores and their respective elements per modality - visual, auditory and human interface. A second level is then obtained by concrete implementations of these cores in correspondence to the external libraries chosen by their particular capabilities. A similar decomposition process is also applied to the elements that map the targeted functional implementation. Both the cores and the corresponding elements that they manage implement generic interfaces according to their role in the desired platform. The resulting abstraction layer, combined with a command-based access, enables the simultaneous use and undifferentiated access between elements through their specific cores independently from the specific library that implements them. In addition, such abstraction layer is also extended to other auxiliary managing cores and their respective elements such as Open Sound Control (OSC) drivers for connection with other software platforms.

So, as a result of this encapsulation, the concrete implementations of the virtual worlds, their visual and auditory representations and the human interfaces that enable the manipulation of the virtual objects can be either refined or substituted according to the desired performance, access or functional needs of the intended use cases. The user configured binding between the elements in play follows the observer design pattern. It is provided through the implementation of custom tracker objects that read and update the relevant entities through event triggering or user defined refresh rates. Furthermore, this modular design allows both static and realtime processing of data as well as physical model based interaction.

To further illustrate the framework's design, a concise description of the sonification package structure follows.

- Core/Element - Both core and elements implement generic interfaces concerning the framework's kernel (ISoundCore; ISoundElement) and the external library used in the implementation (Ex. ISoundCoreSC3). It is segmented per library and functional task and contains the implementation of the synthesis controller. Ex. SonificationIntervalSC3 class.
- Sonification - Implementation of the sonification levels.These provides the triggering algorithm for the synthesis controller instances. Ex. SonificationLevel0 class.
- Model - Provides in real time the data for sonification. Defines the specified model for data conversion and source connectors. Ex. WiiPitchValueToFreqConverter class.

Being so, the development has its focus on providing a set of basic elements for non experts to construct a fully functional sonification use case while being open for expansion and more demanding scenarios. In the latter case, the basic interconnection between the elements implemented using different technologies is guaranteed through the generic interfaces. As an example, a sound element class for rendering a string-like physical model implemented using SuperCollider 3 can be swapped for an implementation using JavaSound for web deployment reasons. Since both implement a generic interface IPluckedDataString that specifies the functionality for both cases (Ex. play()), this change has no effect in the remaining elements (visual and HI elements, triggering observers) of the system.

5.3. Java Technology

The framework's kernel was implemented using Java technology [2]. The primary reasons for this choice are Java's object oriented paradigm, cross-platform support, a wide range of modular freely available open source libraries and a robust integration oriented framework with virtually every IT application area. Particularly relevant are databases connectors, mobile and data mining frameworks, web service based access, web start deployment technology and support for various functional and/or interpreted languages (Ex. Python). In the case of specific performance and/or compatibility demands, it is possible to make use of C/C++ code through component wrapping via Java Native Interfaces. Finally, a strong argument in favor of the implementation of real-time software in Java is the continuous evolution in the Real-Time Specification for Java's implementations (RTSJ).

6. USER EVALUATION

6.1. Description

The presented use case consists of the interactive exploration of a one dimension dataset through sound. The main goal was to present the test subjects with a simple use-case in order to extract preliminary issues concerning the framework. The sonification levels' specification and the technologies used are described in the next sections followed by the user evaluation where both the methodology, the tasks performed by the users and the results are presented.

6.2. Sonification Levels

Three independent sonification levels were define in which the data mining processes are driven by musical relations present in the data (Figure 3 and 4).

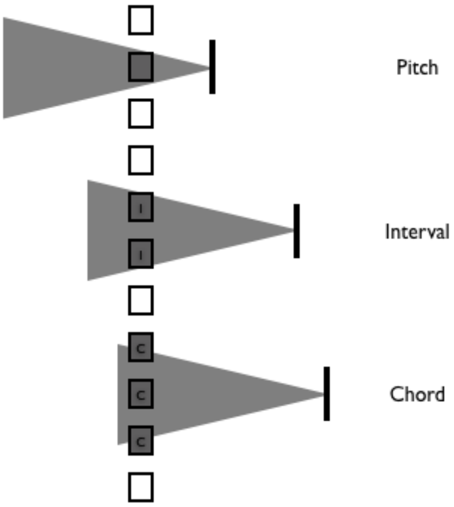


Figure 3: The interactive grouping process and the levels of sonification

- Level 0 - This level manages the sound output concerning the individual entities in the scene. It updates and triggers the assigned pitch of the activated items. This level was implemented through individual sine wave oscillators for each activated element.

[2] http://java.sun.com/

- Level 1 - This level is responsible for detecting and sonically activating musical intervals between two virtual entities under inspection. These relations are defined as a ratio between two given frequencies and used to highlight degrees of variation of the data. For example, a perfect fifth interval can be used for detecting a relation of 3/2 between two elements within the array. This level was implemented through the use of a resonant filter bank per interval. Its application consisted in a percussive type activation each time a given interval was detected.
- Level 2 - This level establishes a relation between several elements and their frequencies in the inspection scope. The presence of a music chord is calculated through the detection of N ratios or intervals from a base frequency. For example, a C major chord is detected through the simultaneous presence of three frequencies: the base F0 and two other that, in relation to F0, respect the 5/4 and 3/2 ratios conditions. By defining and sonically highlighting these relations, further information is provided through a wider view of the data's progression. This level was implemented through the use of a set of delayed sine wave oscillators per chord detection.

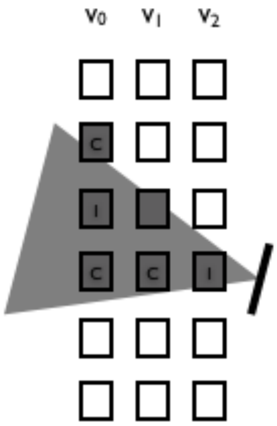


Figure 4: Finding relations through the sonification levels in multivariable inspection.

6.3. Technology

The technologies used were:

- Java 3D Library was selected for the visual engine for its high level scene graph based implementation, well structured overall design and functionalities (e.g. the included support for stereo view).
- The sound engine has been implemented using Supercollider 3 through JCollider, a Java based SCLang implementation [20]. The latter allows not only the instantiation and control but also the definition of the synthesis elements from within the framework's core.
- The NaturaPoint's OptiTrack motion capturing system provided the tridimensional position and orientation tracking through an OSC custom client and a framework's OSC connector element using the NetUtil OSC Java library [20].

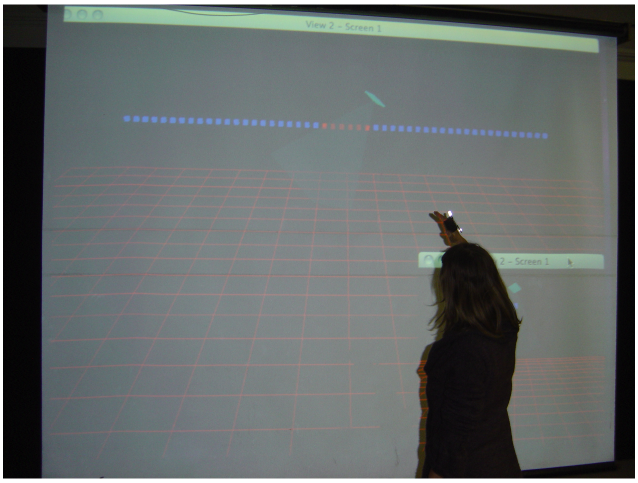


Figure 5: Test subject performing the required tasks.

6.4. User evaluation

To investigate whether the platform functioned as envisioned, two basic user-tests were performed. These consisted of both observations, documenting users' appraisal and feedback regarding the interface. In the latter test, a measurement of their performance was taken while conducting an experiment with a set of predefined tasks. The methods used were adopted from the field of HCI-usability studies [21] [22] and based on techniques such as heuristic evaluation [23] and cognitive walkthrough [24].

The initial and exploratory investigation was performed by a small number of evaluators and consisted on free interaction with the prototype while all the three levels of sonification were activated. They were asked to inspect its basic operation and to comment on it. No further instructions were given at this time. Most users found the interface to be quite responsive and its operation to be intuitive. However, only a small percentage of the test-subject mentioned the different levels in the sonification, so the purpose of these different levels had to be clarified. Other problems that were reported with the prototype included the fact that movements of the inspection tool parallel to the screen were visualized with a certain inclination, which complicated the interaction with the virtual objects. This was due to the use of a perspective based visualization. Finally, some of the users complained about problems concerning their depth-perception within the visual representation. The addition of the second screen which conveyed depth perception improved the spatial awareness of the users and other suggestions that were made by the participants were considered for further implementation. The actual interface (i.e. the MoCap-system), however, was not substituted.

For the second test, a more thorough evaluation of the platform was made, and users were required to perform a series of tasks in which they evaluated the different levels of sonification and the relations between them (Figure 5). The aim of this test was to investigate whether perceptual abilities of the participants were increased by the different sonification levels or not. The proposed tasks were comprised of the exploration of a predefined one dimensional dataset through the use of different combinations of the sonification levels. Within each test, the user would be asked to find a certain relation present in the presented dataset with only the lowest level of sonification after which the same user would be asked to repeat this task using the same level combined with one of higher degrees of sonification to explore a differently ordered representation of that same dataset (Ex. Level1). The test that focussed on the combination of the level 0 and level 1 sonification did not show a rise in effectiveness for any of the participants. This test proved to be problematic because of the fact that a number of different relationships within the dataset (i.e. intervals) were sonified, and it was too difficult for participants that did not have any specific musical training, to discern which. The combination of the level 0 and level 2 of sonification, however, in which the test subjects were required to find a set of relations (i.e. chord) did yield good results. The addition of the level 2 of sonification proved to be very valuable in the discerning task, and it improved the performance of every participant (i.e. the time consumed in performing the task).

After exploration of the interface and performance of the set tasks, participants were required to evaluate the human interface they had used in terms of performance, maneuverability and precision much in the same way as the initial evaluation. Moreover, participants were asked to comment on the completeness of the improved visual output (in order to find the requested relations and to interact with the virtual array) and on the sonification output in terms of distinguishability, information carrying potential and aesthetic qualities. Additionally, a number of remarks and requests were recorded that are being considered for further implementation. A first issue that was raised was that the different sound levels (i.e. the different sound-relations within the data) should be made selectable. Directly related to that, one of the respondents pointed out the need for a test with non-prepared dataset. Admittedly, this was a just critique, which on the other hand indicates that the platform functioned properly as means to evaluate the dataset, because otherwise, the test-subject would not have been able to notice the fact that the values had been retained and only their order had been changed. Finally, some of the test-subjects suggested a number of changes that should be made to the inspection tool, namely the fact that a rectangular as opposed to a square base of the inspection tool would allow for more accuracy in the exploration, and that making the edges of the pyramid-shape visible would enhance the perception concerning the orientation of the tool and thus the precision in the exploration task.

The overall evaluation of this prototype revealed that all participants were able to properly operate this platform. Concerning the sonification's output, the users reported being able to perceive all the sound-levels and were able discern the information that was conveyed by them, although they were not always able to fully perform the set tasks. Furthermore, their performance was considerably improved by the use of the different levels of sonification, and in that respect, the findings of this initial user testing are promising in view of further development.

7. FUTURE WORK

The future development of this project will progress in several aspects. First, we will focus on the expansion of the sonification levels and their intercommunication in order to progressively incorporate higher levels of representation. These will developed not only as a function of the simultaneous data streams at a certain point (in Schaefferian terminology, a Structural analysis) but a time based analysis (Form analysis) in which the result of the sonification process takes into account previously examined samples. Such development will contribute to a more global, musical form inspired perspective of the data's inner relationships by sonically placing its local behaviors within a broader context. Other modes of interaction with the sonification levels will be explored. For example, besides the regulation of the amplitude and reverberation parameters, the relative distance of the virtual microphone and the object(s) under inspection could also be used for the activation and mixing of the sonification levels. Furthermore, in order to the morphology and sonic feedback of the virtual elements to reflect the data's behavior, further investigation in incorporating physical model based interaction will be carried out. As Stockhausen commented about Mikrophonie I, "Someone said, must it be a tam-tam? I said no, I can imagine the score being used to examine an old Volkswagen musically, to go inside the old thing and bang it and scratch it and do all sorts of things to it, and play Mikrophonie I, using the microphone" [25]. Second, spatialization features will be implemented to assist the user's interaction with the virtual objects and to further convey information about the data (Ex. when inspecting N variables, sound spatialization can be useful in informing the user to which variable(s) correspond the heard sonic events). Third, concerning the interface, future testing will include the real time configuration by the user for positioning the inspection window in the dataset, adjusting the inspection tool dimension parameters and the activation of the sonification's levels. Third and finally, all of these features will be subject of a more comprehensive usability study in order to validate the present and future modes of user interaction in new inspection scenarios (Ex. the simultaneous inspection of N variables datasets).

8. CONCLUSIONS

The presented article aims to establish relationships between the interaction sonification field and musical composition practices. Although the present development is still in an initial stage, preliminary testing has shown that the progressive inclusion of the discussed concepts and its related techniques, combined with an embodied music cognition interface approach, can contribute to close the semantic gap between the user and data through sound.

9. ACKNOWLEDGMENTS

The authors would like to give a special thanks to Maarten Grachten for the revision of this paper and Pieter Coussement for the guidance concerning the sound design. This research has the support of the EmcoMetecca project.

10. REFERENCES

[1] J. Alty, "Can We Use Music in Computer- Human Communication?" in *proceedings of HCI '95*, Huddersfield, August 1995.

[2] M. Blattner, D. Sumikawa, R. Greenberg, "Earcons and Icons: Their Structure and Common Design Principles" in *Human-Computer Interaction*, 1989, pp. 11-44.

[3] J. Hankinson, A. Edwards, "Musical Phrase-Structured Audio Communication" in *Proceedings of the 6th International Conference on Auditory Display*, Atlanta, GA, USA, 2000.

[4] J. Gossmann, “Towards an auditory representation of complexity” in *Proceedings of 11th Meeting of the International Conference on Auditory Display*, Limerick, Ireland, July 6-9, 2005.

[5] K. Beilharz, S. Ferguson, “An Interface and Framework Design for Interactive Aesthetic Sonification” in *Proceedings of the 15th International Conference on Auditory Display*, Copenhagen, Denmark, May 18-22, 2009.

[6] M. Chion,“Guide To Sound Objects. Pierre Schaeffer and Musical Research”, (Trans. John Dack and Christine North), http://www.ears.dmu.ac.uk/.

[7] M. Clarke, “Extending Contacts: The Concept of Unity in Computer Music” in *Perspectives of New Music*, Vol. 36, No. 1 (Winter, 1998), pp. 221-246.

[8] P. Vickers, B. Hogg, “Sonification Abstraite/Sonification Concrete: An ’ Aesthetic Perspective Space For Classifying Auditory Displays In The Ars Musica Domain’ ” in *Proceedings of the 12th International Conference on Auditory Display*, London, UK June 20 - 23, 2006.

[9] T. Hermann, “Taxonomy And Definitions For Sonification And Auditory Display ” in *Proceedings of the 14th International Conference on Auditory Display*, Paris, France June 24 - 27, 2008

[10] F. Delalande, “Towards an Analysis of Compositional Strategies” in *Circuit : musiques contemporaines*, vol. 17, n 1, 2007, p. 11-26.

[11] C. Scaletti, “Composing Sound Objects in Kyma” in *Perspectives of New Music*, Vol. 27, No. 1 (Winter, 1989), pp. 42-69.

[12] E. Childs, “Musical Sonification Design”. MA thesis, Dartmouth College, 2003.

[13] O. Kuhl, K. Jensen, “Retrieving and Recreating Musical Form” in *Lecture notes in computer science*, 2008, Springer.

[14] T. Hermann, A. Hunt, (eds.) “Special Issue on Interactive Sonification” in *IEEE Multimedia*, April-June, Vol. 12, No. 2, 2005.

[15] M. Leman, “Embodied Music Cognition and Mediation Technology”. Cambridge, MA: MIT Press, 2008.

[16] R.I. Godoy, “Gestural Imagery in the Service of Musical Imagery” in *Lecture Notes in Computer Science*, 2004, Springer.

[17] A. Mulder, S. Fels, K. Mase, “Mapping virtual object manipulation to sound variation” in *IPSJ Sig Notes*, 1997.

[18] G. Kramer, B. Walker, T. Bonebright, P. Cook, J.H. Flowers, N. Miner, J. Neuhoff, et al., “Sonification report: status of the field and research agenda”, Available at: http://dev.icad.org/node/400, 1997.

[19] K. Stockhausen, “Mikrophonie I (1965), fr Tamtam, 2 Mikrophone, 2 Filter und Regler.” in *Stockhausen, Texte zur Musik 3*, Cologne: Verlag M. DuMont Schauberg, p. 5765.

[20] H. Rutz, JCollider and Netutil Java library. https://www.sciss.de/.

[21] M.M. Wanderley, N. Orio, “Evaluation of Input Devices for Musical Expression: Borrowing Tools from HCI” in *Computer Music Journal*, 26(3), 62-76, 2002.

[22] C. Kiefer, N. Collins, G. Fitzpatrick, “HCI Methodology For Evaluating Musical Controllers: A Case Study” in *Proceedings of the conference on New Interfaces for Musical Expression*, Genova, Italy, 87-90, 2008.

[23] J. Nielsen, R. Molich, “Heuristic evaluation of user interfaces” in *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people*, CHI ’90, 249-256, 1990.

[24] C. Wharton, J. Rieman, C. Lewis, P. Polson, “The Cognitive Walkthrough Method: A Practitioner’s Guide” in *Usability Inspection Methods*, J. Nielsen, Mack, R. L. (eds.). New York, John Wiley and Sons, 1994.

[25] K. Stockhausen, “Stockhausen on Music. Lectures and Interviews compiled by Robin Maconie.” London and New York: Marion Boyars, 1989.

ENVIRONMENTAL SOUNDS AS CONCEPT CARRIERS FOR COMMUNICATION

Xiaojuan Ma, Christiane Fellbaum, Perry R. Cook

Computer Science Department, Princeton University
35 Olden St., Princeton, NJ 08544, USA
{xm, fellbaum, prc}@cs.princeton.edu

ABSTRACT

Sonification, the use of nonspeech audio to represent data and information, has been applied to industrial systems and computer interfaces via mechanisms such as auditory icons and earcons. In this paper, we explore a different application of sonification, which is to facilitate communication across language barriers by conveying commonly used concepts via environmental auditory representations. SoundNet, a linguistic database enhanced with natural nonspeech audio, is constructed for this purpose. The concept-sound associations which are building blocks of SoundNet were validated through a sound labeling study conducted on Amazon Mechanical Turk. We determine the factors that cause a sound to evoke a concept. We examine which aspects of the proposed auditory representations are evocative, and what kinds of confusions may occur. Our results show that sounds can effectively illustrate some concepts, especially those related to concrete entities and actions, and thus can be utilized in assistive communication applications.

1. INTRODUCTION

In everyday life, nonspeech audio such as car horns and fire alarms has been widely used to convey specific information (e.g. alert to danger). People use sound to imply other commonly known messages as well. For instance, people sometimes fake a cough to signify that someone is uncomfortable or ill, and in comedy shows we often hear audience laughing in the background indicating that it is supposed to be a funny scene. These are all examples of sonification, "the use of nonspeech audio to convey information" [21].

Current research on sonification mainly focuses on two areas, industrial human-machine interactions [37][5][29] and computer interfaces (e.g. auditory icons [13] and earcons [6][7]). However, little work has explored the use of environmental sounds to evoke concepts for communication.

Natural language is the primary mode of communication between humans. A concept, whether it is about an entity or an event, concrete or abstract, is encoded in a linguistic form, and can be expressed verbally through words and sentences both within a language and across languages. However, language as a message carrier fails when links between concepts and their linguistic forms are missing, in situations like people trying to communicate through an unfamiliar language, people learning a new language, and people with language impairment. When the associations between words and concepts are either not yet established or corrupted, it is impossible to retrieve information via a language. To bridge language barriers, non-linguistic modalities have been explored to assist comprehension and expressions of concepts. Compared to visual languages [23][24][22][32], less attention has been given to language support through nonspeech auditory stimulus.

One disadvantage that auditory representations have over pictures is that sound requires time to play and has to be played in sequence [38]. Many concepts do not produce a (a distinctive) sound. However, there are still cases where a sound can evoke a concept even better than a picture. For example, "thunder" (unlike lightning) and "chirp" (unlike bird) are harder to visualize; "coughing" and "sneezing" can be distinguished more easily by sounds than by pictures; and "tuning a radio" can be better portrayed via a sound unfolding over time than a static picture.

To explore the use of environmental sounds as concept carriers across language barriers, we built SoundNet, a lexical network which consists of associations between concepts and short environmental sounds, and can be employed in applications like multimodal dictionaries for language learners or people with language disorders to look up concepts for communication (e.g. relaying symptoms to doctors or ordering food). A sound labeling study was conducted to verify the concept-sound associations established in SoundNet. Analysis of our results addresses issues such as what kinds of concepts can be expressed by a nonspeech sound, what aspects of a sound can be perceived, and which sounds are confusable, and guides the improvement of SoundNet.

2. BACKGROUND WORK

2.1. Sonification

Sonification refers to the use of acoustic signals to illustrate data and information. Compared to visualization, audio has been found to have the advantages of evoking temporal characteristics and showing transformation over time [19][25][26]. Furthermore, auditory display does not require users to direct their visual attention, and thus is suitable for eyes-free environments.

Sonification techniques have been applied to catching attention/alerting, and depicting changes in data by the shift of sound frequencies and intensity. Examples of such auditory systems include audio alert/monitoring and guidance systems for airplanes [5][29], nuclear power plants [37], factories [17], and scientific data analysis [30]. There are also attempts to use sound patterns on computer interfaces (e.g. earcons and auditory icons).

2.2. Earcons and Auditory Icons

Earcons are nonspeech synthetic audio patterns designed to provide information about objects, operations, status, and interactions on computer interfaces via auditory features like pitch, rhythm and volume [6][7]. People are not familiar with synthetic sounds and their assigned meanings, and thus the use of earcons requires learning. Compared to earcons, auditory icons are more natural since they encode computer events with everyday sounds

We thank the Kimberley and Frank H. Moss '71 Princeton SEAS Research Fund for supporting our project.

[13]. For example, the sound of throwing into a trashcan is used to indicate the deletion of a computer file. Additional work on auditory icons includes [15][27][13].

Both earcons and auditory icons aim to represent specific information, mainly on computer interfaces. Earcons and auditory icons are metaphor or analogy, instead of a direct translation of the everyday experience embedded in the sounds.

2.3. Everyday Listening

Everyday listening is the perception of auditory events (e.g. the characteristics of the sources of the sounds, their position and interactions), in contrast to musical listening, which captures the pitch, loudness, timbre, and changes of the sounds [16]. [20][18][36] have shown that people can identify significant aspects of environmental sounds from their experience. For instance, people can tell a car engine sound from footsteps on a wooden floor, and detect if the car is approaching or departing. Auditory icons utilize everyday listening to illustrate computer events with sounds from real life with similar effects.

By contrast, we explore the use of environmental sounds to convey everyday concepts to facilitate communication across language barriers. The intended concepts are directly linked to sources, locations, and actions involved in the sound events, and thus can be evoked through everyday listening. People working in an unfamiliar language environment, or people learning a new language, or people with low literacy, or people with language disabilities face difficulties in daily communication due to their failure to comprehend and/or produce languages. [8][9] have shown that many people with language disorders still maintain the ability to identify natural sounds. This suggests that everyday listening is viable for both healthy populations with limited language skills and language-impaired populations. Nonspeech audio has potential to assist language comprehension.

In the following sections, we describe SoundNet, a lexical network enhanced with environmental audio representations, its construction, and its effectiveness in conveying common concepts.

3. SOUNDNET

SoundNet is an environmental sound-enhanced lexical database. Different from auditory icons and earcons, the SoundNet backend vocabulary consists of hundreds of concepts (in English) used frequently in daily communications. The concepts are interlinked through semantic relations inherited from WordNet [10]. Each data unit in SoundNet (structure shown in Figure 1) has three components: a concept represented as a synonym set (synset) with its definition, an **audioability** (we define as "the ability for a concept to be conveyed by an environmental sound") rating, and a soundnail (a five-second non-speech sound) if audioable.

3.1. Vocabulary Generation

The SoundNet vocabulary consisting of commonly used concepts is based on the glossary of Lingraphica [22], a communication support device for people with aphasia. We extracted 1376 words in base form from the Lingraphica vocabulary.

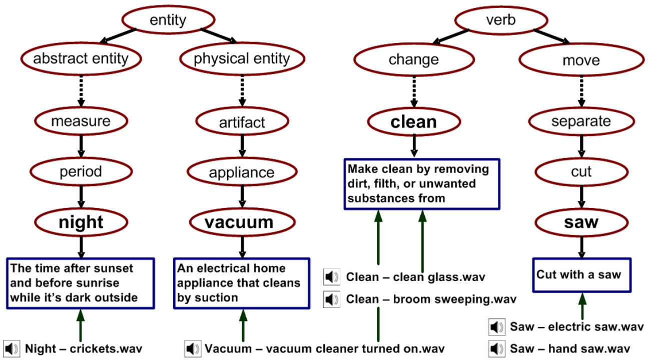

entity
abstract entity
physical entity
measure
artifact
period
appliance
night
vacuum
The time after sunset and before sunrise while it's dark outside
An electrical home appliance that cleans by suction
Night – crickets.wav
Vacuum – vacuum cleaner turned on.wav
verb
change
move
clean
separate
Make clean by removing dirt, filth, or unwanted substances from
cut
saw
Cut with a saw
Clean – clean glass.wav
Clean – broom sweeping.wav
Saw – electric saw.wav
Saw – hand saw.wav

Figure 1. Structure of SoundNet

Score	Justification	e.g.	Script
0	cannot make sound or be used to produce sound and cannot be evoked by sound	"month"	N/A
1	can make sound or be used to produce sound, but cannot be evoked by sound	"fruit"	biting an apple
2	can make sound or be used to produce sound, and may be able to be evoked by sound, meaning the sound could be ambiguous	"pen"	pen writing on paper
3	can make sound or be used to produce sound, and can be evoked by sound (the sound is distinctive)	"dog"	dog barking

Table 1. Audioability four-point rating scale and examples.

However, not every word on initial Lingraphica list could be illustrated by a sound. As a second step, we brought in sound track labels from the BBC Sound Effects Library [4] to pull out words with potential good sound-concept correspondence, since the BBC library is the major environmental sound provider for SoundNet. All the BBC sound captions were decomposed into individual words. The same process was applied to the raw BBC vocabulary as to that of Lingraphica. A list of 1368 words was generated. Its overlap with the Lingraphica list contained 211 nouns, 68 verbs, 27 adjectives, and 16 adverbs. This became the core vocabulary of SoundNet, with each word assigned to its most frequent sense and part of speech according to WordNet.

3.2. Audioability Ratings

Before attempting to create auditory representations for each concept in SoundNet vocabulary, their **audioability** was assessed on a four-point scale (Table 1). Five raters assigned each concept an audioabilty score, and for the ones with a non-zero rating, wrote a script of sound scene that could be used to evoke the intended concept. Two additional raters helped judge and finalize the audioability ratings and scripts. Overall, 184 out of 322 words were considered audioable (score > 1), and their associated sounds were selected based on the scripts.

3.3. Soundnail Creation

The three sources of the environmental audio clips employed in SoundNet include the BBC Sound Effect library (about 2/3 of the representations), Freesound [12] and FindSounds [11]. For three practical reasons, we constructed 5-second auditory illustrations called **soundnails** from the original sound files. First, most of the original sounds are dozens of seconds to several minutes in length, requiring significant listening/processing time. Second, the sound scenes with multiple events are often too complex to evoke individual concept. Third, the sounds, especially the BBC high quality stereo clips, are too large to store in mobile devices and to play on the Internet (the main interfaces to our database).

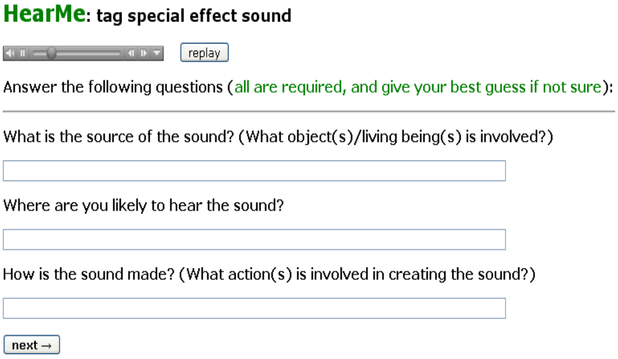
We first down-sampled all selected clips to 16kHz, 16 bit mono. We picked the 16kHz sample rate because it has been conventionally used in speech recognition, and the sample rate used in many games (especially mobile games) is between 8kHZ and 22.05kHz. A pilot study [31] also verified that people can recognize sound events at the 16kHz sample rate.

Each down-sampled sound clip was then randomly divided into five to over a hundred 5-second fragments in proportion to the length of the original clip. To pick out a representative soundnail for the given concept, the fragments were grouped into three to four clusters (based on sound scene complexity) using K-Means algorithms using six audio features, including means and standard deviations of RMS Energy, Spectral Centroid, Spectral Flux, 50% and 80% rolloff, and MFCCs [33]. The fragments closest to the center of each cluster automatically became soundnail candidates. We review all candidates which captured different distinctive parts of the original sound scene and picked out the most appropriate one to illustrate concepts in SoundNet. For example, 5-second fragments from the sound "Lines AND Tones, 3 STD Rings, Phone Answered With Pip" were clustered into "connecting," "ringing," and "ringing and picked up." The representative from "ringing and picked up" was assigned to the concept "call: get or try to get into communication by telephone."

A total of 327 soundnails were generated for the 184 audioable concepts in SoundNet. It is not a one to one mapping (Figure 1). Certain concepts are associated with more than one sound. For example, two sounds "electric saw" and "hand saw" are assigned to the verb "saw (cut with a saw)." On the other hand, some soundnails are used to depict multiple concepts. For instance, the soundnail "vacuum cleaner turned on" is assigned to both "vacuum" (noun) and "clean" (verb). As suggested by previous research [2][3], the number of options and the ease of mental image generation may affect people's performance on sound naming. Most of the soundnails were normalized in volume, except for those that explicitly needed to have higher or lower volume, such as the soundnail for "distance".

4. STUDY: SOUNDNAIL COMPREHENSION

Before applying SoundNet to assistive communication systems, we need to investigate if the soundnails effectively convey the pre-assigned concepts or cause confusions, and try to determine guidelines to generate more evocative auditory representations. This can be extended to more general research questions: what kinds of concepts can be evoked by a natural sound? What kinds of sounds are distinctive enough to evoke a concept? What kinds of miscomprehension may appear in everyday listening and what introduces the confusion? To address these questions, we designed and conducted a large-scale study to collect human-generated semantic labels for the nonspeech soundnails on the Amazon Mechanic Turk (AMT) platform [1]. Compared to a well-controlled lab experiment, an online study is faster, less expensive, and can access a larger number of participants more easily, despite the lack of knowledge of participants' background and behaviours. We inserted several safeguarding methods to ensure the quality of the online study.



HearMe: tag special effect sound

Answer the following questions (all are required, and give your best guess if not sure):

What is the source of the sound? (What object(s)/living being(s) is involved?)

Where are you likely to hear the sound?

How is the sound made? (What action(s) is involved in creating the sound?)

Figure 2. Sound labeling experiment interface.

4.1. Study Design

Our goal was to determine whether, and in which cases, specific responses (nouns, verbs, adjectives, and adverbs) can be generated from auditory perception of a soundnail. Since people tended to label a sound with its source(s) in a free tagging study [14][36], we collected answers to three questions about each soundnail, so as to encourage people to generate as much information across different parts of speech as possible:

1) What is the source of the sound? (What object(s)/living being(s) is/are involved?)
2) Where are you likely to hear the sound?
3) How is the sound made? (What action(s) is/are involved in creating the sound?)

4.2. Study Environment and Participants

Figure 2 shows the web-based experiment interface. The sound automatically starts to play once the page is loaded. Subjects could replay the sound as desired. They need to submit responses regarding the source(s), location(s), and interaction(s) involved in the sound production. The study was posted on Amazon Mechanical Turk (AMT), a web service provided by Amazon, where people all over the world can post or take part in online surveys with an Amazon account.

In our sound labelling study, the 327 soundnails were randomly divided and grouped into 32 Human Intelligence Tasks (HITs), with 10 to 11 sounds in each. A HIT is the basic unit for task submission and payment. The average completion time of a HIT of tagging 10 to 11 soundnails is 14.64 minutes, not too long to get tired and lose focus. We requested at least 100 people to label each HIT, and no participants could work on the same HIT twice. It took 97 days to complete the experiment. Although we have no access to participants' identity and demographic information, we were able to record their geographic locations. Over 2,000 people from 46 countries took part in the study, which implies that our results had universal validity. Individual responses and completion time was logged.

4.3. Quality Control

A pilot study was carried out to test the experiment interface with 22 undergraduate students. Each soundnail was tagged by five to eight students, and a post-study questionnaire gathered feedback on the design of the study. Adjustments such as auto-playing of the sound and phrasing of the questions were made accordingly.

Since we have no control over participants' behavior in the AMT study, quality-guarding schemes were applied the study:

1) **Hardware/software preparation**: The hardware (speakers or a headphone) and software (proper plugin to play the sounds) requirements were specified on the welcome page of the experimental interface. Instructions and links were provided to help with the study setup.
2) **Embedded checks**: Participants needed to correctly fill out a sequence of letters and numbers presented in an auditory **"captcha"** to login the actual study. A **training sound** was played at the beginning of each HIT. It demonstrated what kinds of sound would be played and how to answer the three questions. Participants were asked to fill out corresponding text fields as instructed as a practice. These mechanisms checked the quality of the sound system, and ensure that it was a human listening, not a computer script. Participants also get a chance to learn about the interface and tasks.
3) **Label validation**: Once the answers were submitted, non-lexical responses such as "09j1h" were automatically eliminated. To further filter out irrelevant words like "hello," the responses were compared to the labels collected from the undergraduate student pilot study. Responses with less than 50% overlap were rejected. Finally, we manually reviewed the remaining responses and kept the valid ones.

5. RESULTS AND ANALYSIS

Over 100 (up to 174) tags were collected for each soundnail in the AMT study. They are mostly in the form of sentences or short phrases. We extracted concepts out of the raw answers following the process described in Section 5.1, and a quantitative measure called sense score was computed to assess people's agreement. Section 5.2 presents the validation of concept audioability. Section 5.3 explores the influence on audioability of two linguistic properties, concreteness and parts of speech. Section 5.4 looks at three main aspects of everyday listening: source, location, and action. Section 5.5 provides a detailed discussion of confusion errors in soundnail perception.

5.1. Data Processing

The processing procedure of raw responses collected in the AMT study was similar to that for the BBC sound captions. Sentence and short phrase were broken up into "bags of words," with function words such as "a" and "or" removed. Remaining content words were checked in WordNet [10] for validity. If not found, they were transformed back to the base form using a Natural Language Toolkit stemmer [28] and then assigned to proper sense. For example, "woods" were kept while "pens" was changed to "pen." All misspellings were corrected manually.

For each soundnail, we counted how often each word appeared across all labellers. This number is referred as the **word count.** Because people may use different words to express the same idea, we further group lexicons with same or very similar meanings into units called **sense set**. By its nature, words from the same synonym set were always in the same sense set (e.g. "child" and "kid"). Other relations between words in sense set include hypernym (superordinate), hyponym (subordinate), meronym (part), holonym (whole), instance, etc. Words in a sense set could have different parts of speech. For instance, the "rain" sense set includes "rain" (n.), "rain" (v.), and "rainy" (adj.). To be distinguished from individual **words**, a sense set is referred as a **label** in the following sections. If not specified, the evaluations described below are all label (sense set) instead of word-based. The most frequent word within a sense set (from WordNet) was used as the representative for reference.

The word count of a sense set is the sum over all word counts of its members. Since a word count depends on the number of labellers and thus cannot be compared across sounds, a relative score, referred as **sense score** is calculated for each sense set per sound. It is the average number of times a sense set (label) is generated for a soundnail across all labellers.

sense score = word count of a sense set / number of labelers

The sense score indicates the strength of people's agreement on a label. The estimate of the highest sense score is 3, meaning that every labeller used the label in answers to all three questions. A sense score of 0.5 means 50% of the participants generate the label (sense set) once, and a score of 2 means each person entered the label twice on average. For each soundnail, the sense set receiving the highest sense score (**top sense score**) is considered as the **most agreed-on label**.

5.2. Audioability

For each soundnail, we compared its pre-assigned concepts in SoundNet to the most agreed-on label obtained in the AMT study. Table 2 shows the top five and bottom five soundnails based on the sense scores of intended concepts. A test for homogeneity of variances showed that sense scores for intended concepts and most agreed-on labels came from the same normal distribution. It suggests that if the pre-assigned concepts are strongly audioable (with a rating 3 in the parentheses), they are likely to be agreed-on by labelers. In contrast, people tend to generate a different more audioable concept if the intended one is less evocative.

Sound	Assigned	S.S.	Agreed-on	S.S.
cat_meowing	cat (3)	2.53	cat (3)	2.53
train_choochoo	train (3)	2.46	train (3)	2.46
telephone_ring	phone (3)	2.43	phone (3)	2.43
horm_carHorn	horn (3)	2.42	horn (3)	2.42
baby_happy	baby (3)	2.36	baby (3)	2.36
empty_waterOut	empty (2)	0	water (3)	1.68
teapot_waterFill	teapot (1)	0	water (3)	1.71
speed_carTurnFast	speed (2)	0	car (3)	1.71
skip_tapeForward	skip (1)	0	projector(3)	1.81
cracker_eatCrunch	cracker(2)	0	eat (3)	1.91

Table 2. The five most and least effective soundnails with audioability ratings and sense score (S.S.) for their pre-assigned concept and the most agreed-on labels.

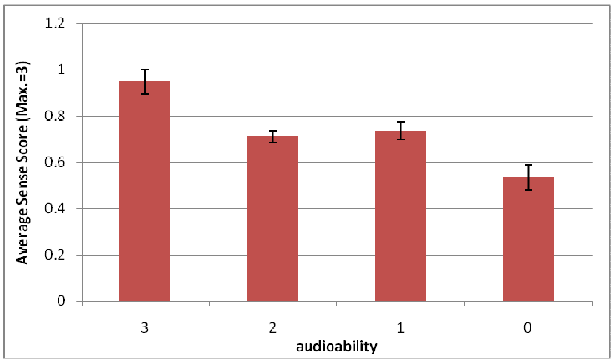


Figure 3. Comparison of audioability ratings and sense score.

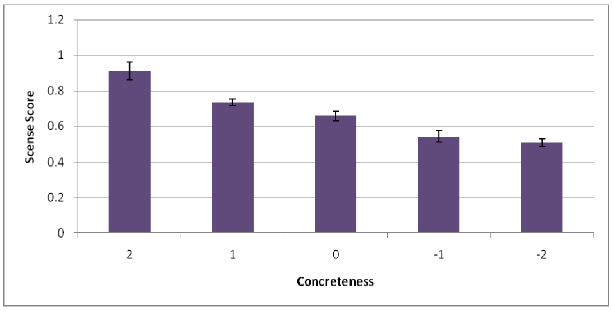


Figure 4. Comparison of concreteness and sense score.

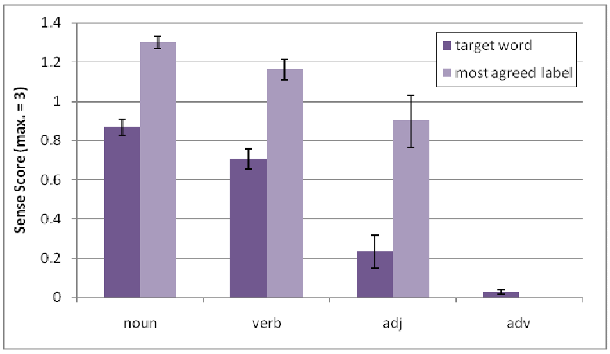


Figure 5. Comparison of sense score of target words and most agreed-on labels from different parts of speech.

Comparison of the audioability ratings and sense scores of the target concepts is shown in Figure 3. ANOVA shows that strongly audioable (rating 3) concepts received a significantly higher sense score, and scores for non-audioable concepts were significantly lower (F(1, 206) = 19.941, p < 0.01).

5.3. Relevant Linguistic Properties

How likely a concept can be evoked by an environmental sound may be affected by its linguistic properties such as concreteness and part of speech. We collected all labels with a sense score no less than 0.25 (meaning that at least 25% of the participants generated the label once), and explored the impact of two lexical properties on their sense scores.

Concreteness: Figure 4 shows that concrete words are easier to name and categorize based on nonspeech sounds, similar to the conclusions for pictures [23][24][35]. Sense scores dropped significantly as the concreteness (based on the MRC Database [34]) went down (F(1,702) = 33.596, p < 0.01).

POS	What	Where	How
Noun	313	323	256
Verb	56	15	134
Adj.	3	2	2
Adv.	0	8	0

Table 3. Comparison of numbers of labels in different parts of speech among answers to the three questions.

	Word Count		Label Count		Top Sense Score	
	Mean	SD	Mean	SD	Mean	SD
What	49.57	1.31	39.99	1.20	0.535	0.013
Where	56.46	0.93	47.64	0.90	0.414	0.010
How	104.48	1.93	86.43	1.83	0.624	0.014

Table 4. Word count, label count, and sense score comparison.

	Source	Location	Action
1	phone	playground	start (a car)
2	baby	road	type
3	chicken/rooster	farm	zip
4	car horn	house/home	honk
5	doorbell	station	ring (phone)
6	cat	school	print
7	bird	office	eat
8	train	store	uncork
9	dog	swimming pool	knock
10	typewriter	kitchen	talk

Table 5. Top 10 recognizable sources, locations, and actions.

Parts of speech: Figure 5 shows the sense score for intended (target) concepts and the most agreed-on labels for different parts of speech. Results showed that it was significantly more likely for people to generate a noun than a verb, and even more than an adjective or adverb (for target words: F(3,204) = 3.296, p = 0.0215, $\eta^2 = 0.7673$). About 80% of soundnails intended for a noun, half of those for a verb, and almost all for adjectives and adverbs were most agreed upon as nouns.

However, the parts of speech of responses varied for the different questions (Table 3). Predominantly nouns and some verbs were used for sound sources, since these are usually a person, a thing, or an action/event. Answers to "how the sound was made," mainly about sound production actions, contained many more verbs. On the contrary, fewer verbs and some adverbs (e.g. "outside") appeared in the descriptions of the location(s).

5.4. Sources, Locations, and (Inter)actions

Previous sections discussed at word level which concepts can be evoked by a sound. In this section the data are investigated from the sound perspective: what kinds of sounds are distinctive, and what aspects of the sounds can people perceive.

Table 4 shows the comparison on word counts, sense set (label) counts, and top sense scores among responses to the three questions "what," "where," and "how." In general, significantly more words (F(2,978) = 424.85, p < 0.01) and sense sets (F(2,978) = 331.84, p < 0.01) were generated to describe an interaction than to name a source or location. People are significantly more likely (F(2,978) = 67.668, p < 0.01) to agree on what kinds of actions create sound (60% soundnails) than what object(s) it was (23% soundnails) and where it took place (17% soundnails). Table 5 lists the top 10 sources, locations, and actions that people correctly recognized given the soundnails.

Assigned Type	Resp.Type (by S.S.)	Resp.Type (by %)
source	sound source (1.31)	location (21%)
	sound action (0.76)	similar source (15%)
	similar source (0.59)	sound source (12%)
	location (0.49)	sound action (9%)
source indirect	source indirect (1.87)	location (17%)
	source partial (1.37)	action partial (11%)
	action partial (0.57)	source partial (9%)
source active	source active (1.10)	source active (19%)
	source passive (0.97)	location (18%)
	sound action (0.68)	sound action (12%)
source passive	source passive (1.21)	location (14%)
	source active (0.99)	source active (10%)
	sound action (0.76)	sound action (10%)
location	sound source (1.04)	action partial (17%)
	location (0.74)	location (11%)
	action partial (0.67)	source partial (9%)
action	sound source (1.15)	location (16%)
	sound action (0.87)	sound action (15%)
	similar source (0.63)	similar source (7%)
attribute	sound source (1.17)	similar source (36%)
	sound action (0.99)	sound action (9%)
	location (0.72)	location (8%)
scene	sound source (0.83)	source partial (18%)
	scene (0.81)	location (13%)
	source partial (0.71)	action partial (11%)
	action partial (0.68)	sound source (10%)
time	sound source (1.44)	sound source (20%)
	source partial (0.95)	source partial (16%)
	location (0.69)	location (16%)
	time (0.56)	time (12%)

Table 6. Types of pre-assigned concepts and agreed-on sense sets (Resp.) ranked by sense score (S.S.) and by percentage (%).

Case	Sound	Intended	Agreed
1)	Phone, ring and pick up	phone	phone
2a)	Knock, on the door	knock	door
2b)	Bag, zipping	bag	zipper
3)	Turn, right turn signal	turn	clock
4)	Umbrella, open umbrella	umbrella	match

Table 7. Examples of how well sounds convey target concepts.

1) **Source**. Results suggest that human and animal sounds are relatively easy to name. However, how fine the distinction is depends on the sound characteristics. For example, most people can identify sea gulls but not chaffinches. For non-living objects and devices, those which are an auditory system themselves like doorbells and those which produce sounds with special temporal patterns are easier to tell.
2) **Location**. Environments in which sounds are made can be identified by the sources and events detected in the scene. For example, traffic sounds may suggest "road" whereas dish clicking sounds may suggest "kitchen."
3) **Action**. Sound-creating actions that people can name include ones that aim to make a sound, like "honk," and ones that represent an operation or a process with unique sounds generated, like "zip" and "start (a car)."

We assigned what aspect of a sound is described (type) to both the intended concepts and the sense sets (score >= 0.25). The comparison results are listed in Table 6, ranked by both sense score and frequency. If a sound is produced by a single object or living being, the source is denoted as "source" (e.g. "bird" for the bird chirping sound). If the sound is created by interaction between two objects (e.g. footsteps on the wooden floor), the one that initiates the action (e.g. feet and shoes) is called "source active," while the other one (e.g. floor) is called "source passive." If the object is not directly related to the sound scene, it is called "source indirect" (e.g. "bag" for the sound of "zipping up"). "sound source" and "sound action" refer to the actual source and action; "source partial" and "action partial" refer to part of the source/action; and "similar source" and "similar action" refer to those generating similar sounds to what were given. Results show that, regardless of what information was expected (whether it is source, location, action, or attribute), many sense sets were related to locations but all with relatively low scores. It suggests that the location information is usually more ambiguous because some sounds can appear in different places. For example, the dish clinking sound occurs in the kitchen or on the dining room table. The sound of someone coughing can happen nearly everywhere with a person. On the contrary, less labels about the source of the sounds were generated, but with high agreement.

5.5. Confusion Errors in Soundnail Perception

We compared for each soundnail the intended concept assigned in SoundNet and the most agreed-on sense set in the AMT study. The results can be categorized in four cases (Table 7):

1) The target concept appeared in the most agreed-on sense set. Soundnails in this category (90 of them) succeeded in conveying the intended concept and has the potential to enhance language comprehension and communication.
2) People agreed on a concept related to certain aspect of the sound, though not the one given in SoundNet. It indicates that the sound is distinctive but people have different focus: 2a) a different object or action; 2b) concrete content in the sound scene while the assigned concept is abstract or not directly reflected. This category has 150 soundnails.
3) Label with highest agreement was completely unrelated to the sound scene (52 sounds in this category). It suggested that those soundnails have some characteristics, but not fine enough to be told apart from similar sound events.
4) People showed no agreement on 35 soundnails, meaning that these sounds are too ambiguous to illustrate a concept.

We further looked into the semantic relations (based on WordNet) between sense set members for each question. This gives us an insight on the causes for confusion, including synonyms (e.g. car-auto mobile), hypernyms (e.g. vehicle-car), hyponyms (e.g. sports car-car), meronyms (e.g. car window-car), holonyms (e.g. window-windowpane), sisters (e.g. truck-car), nephews (e.g. fire truck-car), and instances (e.g. Ford-car). Table 8 shows that over 1/3 of the words in the responses to each question are synonyms to the representative word for the sense set they belong to, around 10% are hyponyms. However, hypernyms and meronyms got relatively higher scores (bold in Table 8). This suggests that people are more likely to recognize a more generic scope of the actual source, location, and action in the sound, or detect part of them. People usually got confused with objects or interactions in the sister or nephew categories, and even with completely unrelated events that cause similar effects or generate similar sounds.

Question	Resp. Type	Percentage	Sense Score
what	**synonym**	**0.3995**	**0.2085**
	hyponym	0.1026	0.0354
	sister	0.0691	0.0411
	hypernym	0.0676	0.0532
	similar sound	0.0386	0.0358
	nephew	0.0338	0.0270
	meronym	**0.0331**	**0.0829**
	instance	0.0286	0.0424
	holonym	0.0193	0.0324
where	**synonym**	**0.3386**	**0.1817**
	hyponym	0.0974	0.0506
	hypernym	**0.0876**	**0.0933**
	meronym	0.0788	0.0448
	nephew	0.0479	0.0429
	sister	0.0435	0.0487
	similar place	0.0411	0.0392
	instance	0.0240	0.0307
	holonym	0.0210	0.0588
how	**synonym**	**0.3490**	**0.2412**
	hyponym	0.0900	0.0430
	sister	0.0638	0.0478
	hypernym	**0.0517**	**0.0637**
	similar sound	0.0459	0.0521
	nephew	0.0398	0.0396
	meronym	**0.0391**	**0.1053**
	instance	0.0315	0.0378
	similar effect	0.0237	0.0444
	holonym	0.0204	0.0331

Table 8. Semantic relations between sense set members.

Intended Concept (bold) and Different Responses	
source	**alarm**: siren, alert, warning, doorbell, clock
	baby: infant, newborn, child, kid, toddler, little
	bottle: container, jar, can, dish, plate, glass
	car: vehicle, engine, motor, truck, bus, motorcycle
	floor: ground, stairs, porch, patio, surface
	movie: film, TV, radio, stereo, videogame
	plastic: wrapper, cellophane, polyethylene, paper
	rain: droplet, storm, hail, downpour, waterfall
	snow: dirt, dry leaves, ice, gravel, mud, twig
	typewriter: copier, fax, printer, computer, xerox
location	**farm**: barn, livestock, ranch, yard, garden, zoo
	hospital: clinic, nursery, daycare, medical center
	kitchen: restaurant, bar, café, cafeteria,lunchroom
	playground: park, court, gym, yard, stadium
	road: street, highway, race track, driveway
	school: class, classroom, college
	store: shop, supermarket, market, mall, retail
	swimming pool: lake, pond, river, ocean, beach
	train station: airport, terminal, platform, bus stop
	workshop: factory, garage, construction site
action	**break**: crack, creak, crush, shatter, smash, crash
	chirp: call, crow, sing, whistle, cackle
	clink: clank, jingle, tinkle, click, chime
	crunch: crackle, crisp, rack, scrap, scratch, break
	eat: bite, chew, munch, masticate, crunch
	jingle: rattle, rustle, fiddle, tinkle, shake
	knock: beat, kick, bang, strike, clap, hit, punch
	rub: scratch, scrub, rip, stretch, twist, squeeze
	pour: drip, fill, leak, trickle, splash, drop
	walk: gallop, run, jump, stomp, climb, jog, trot

Table 9. Examples of confusions generated for the sounds

Table 9 summarizes examples of words (some are confusion errors) people generated for the pre-assigned concepts given soundnails. The bold words are the actual sound source, location, or action. The confusions for sound sources may come from similar materials (e.g. bottle and jar) or textures (e.g. snow and gravel), and similar functions/interactions (e.g. typewriter and computer). The confusions for sound locations can be caused by similar content (e.g. farm and zoo), and similar events (e.g. playground and gym). The confusions for sound-producing actions can result from similar objects involved (e.g. knock and kick) and similar effects they lead to (e.g. crumple and squeeze).

6. DISCUSSION

The results from the soundnail labeling study may guide us towards better creation of nonspeech auditory representations.

It seems while people are focusing on everyday listening (as expected for our purposes), less information from musical listening is utilized. For example, a soundnails (far away foghorn) were used to illustrate "distance." The volume of the sound is much lower than average, but people still tried to indentify the source instead of describing the distance. In another example, the "power down" sound is used to evoke "down." People mostly wrote "videogames," "Sci-Fi," or "synthesized," rather than saying that the pitch and loudness went down.

Abstract concepts are hard to evoke. We have tried to use sounds of a special instance (e.g. the sleigh bell sound for "winter") and the combination of sounds for several concrete components of the abstract event (e.g. a sequence of rooster crowing – clock ticking – crickets chirping to depict "day" (a 24-hour period)), but none was successful. People almost always identify the concrete objects and actions, such as "bell," "rooster," and "crowing."

The effectiveness of different sounds from similar source(s) may vary greatly. For instance, the top sense score for the "saw – hand saw" soundnail is 1.78, while that for "saw – electric saw" is 0.65; the "train – choochoo.wav" sound (steam train whistling) receives a top sense score of 2.48 while the soundnail "train – arriving.wav" gets a score of 1.41. It implies certain sounds are more distinctive and should be selected as the representation.

People's familiarity with the sounds has great impact on their interpretation. This difference may come from 1) Age: younger generations have little exposure to old fashioned devices, and thus have more trouble recognize them. For example, the sense score for the "call – rotary dial.wav" soundnail is much lower with the 25 undergraduate students in the pilot study than in the AMT study. 2) Culture: people from different cultures may associate completely different sounds with the same event/scene. For example, for labelers from China, the "NBC news theme" sound used for the concept "news" may just be a piece of music. 3) Personal experience: people who have never heard an elephant trumpeting are less likely to name the sound correctly.

7. CONCLUSIONS

We presented an attempt to use short environmental sounds to convey concepts for facilitating communication across language barriers. SoundNet, a lexical semantic network enhanced with nonspeech sounds was constructed and evaluated via a large scale sound labeling study conducted on Amazon Mechanical Turk.

Results showed that over 73% of the soundnails evoked a concept that people agreed upon, 37.5% of which matched what was assigned in SoundNet. It was suggested that concrete concepts are easier to name from a sound, and many more nouns were generated than other parts of speech. As perceived in everyday listening, location(s) of a sound is the hardest to specify, whereas actions involved in the sound production are easier to distinguish. Similar materials or textures of the sound sources, similar effects of the interactions, and similar events that take place can all be the cause of confusion. Furthermore, people are more likely to agree on a more generic concept or a specific part of a complex source or action involved in the sound creation.

Overall, distinctive environmental sounds can effectively evoke concepts (nouns and verbs) commonly used in everyday communication, indicating that SoundNet has the potential of assisting communication across language barriers.

8. REFERENCES

[1] Amazon Mechanical Turk. https://www.mturk.com/ 2009.

[2] Ballas, J.A. Common factors in the identification of an assortment of brief everyday sounds. *J. of Experimental Psychology*, 19(2):250–267, 1993.

[3] Ballas, J.A. and Sliwinsky, M.J. Causal uncertainty in the identification of environmental sounds. *Tech Report ONR-86-1*, Office of Naval Research, Dept. of Psychology, Georgetown University, Washington, D. C., 1986.

[4] BBC Sound Effects Library. Original CD series. 2009.

[5] Begault, D., Wenzel, E., Shrum, R., and Miller, Joel. A Virtual Audio Guidance and Alert System for Commercial Aircraft Operations. *ICAD'96* 1996.

[6] Blattner, M., Sumikawa, D., and Greenberg, R. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*. 4(1). 1989.

[7] Brewster, S. Using nonspeech sounds to provide navigation cues. *ACM Transaction on Computer-Human Interactions*. 5(3), pp. 224-259. ACM Press. 1998.

[8] Clarke, S., Bellmann, A., De Ribaupierre, F., and Assal, G. Non-verbal auditory recognition in normal subjects and brain-damaged patients: Evidence for parallel processing. *Neuropsychologia*. 34 (6), 587-603. 1996.

[9] Dick, F., Bussiere, J., and Saygm, A. The Effects of Linguistic Mediation on the Identification of Environmental Sounds. Center for Research in Language. 14 (3). 2002.

[10] Fellbaum, C. WordNet: Electronic Lexical Database, A semantic network of English verbs. MIT Press, 1998.

[11] FindSounds. http://www.findsounds.com/. 2008

[12] Freesound Project. http://www.freesound.org/. 2008

[13] Garzonis, S., Jones, S., Jay, T., and O'Neill, E. Auditory Icon and Earcon Mobile Service Notifications: Intuitiveness, Learnability, Memorability and Preferences. In Proc. *CHI'09*. pp. 1513-1522. 2009.

[14] Gaver, W. Everyday listening and auditory icons. Doctoral Dissertation, University of California, San Diego. 1988.

[15] Gaver, W. The SonicFinder: An Interface That Uses Auditory Icons. *Human-Computer Interaction*. 4. 1989.

[16] Gaver, W. What in the World Do We Hear? An Ecological Approach to Auditory Event Perception. *Ecological Psychology*, 5 (1): 1-29. 1993.

[17] Gaver, W., Smith, R., and O'Shea, T. Effective Sounds in Complex Systems: The ARKola Simulation. In Proc. *CHI'91*. pp. 85-90. ACM Press, 1991.

[18] Handel, S. *Listening: An introduction to the perception of auditory events*. Cambridge, MA. MIT Press. 1989.

[19] Hartmann, W. M. Sounds, signals, and sensation: Modern acoustics and signal processing. Springer Verlag. 1997.

[20] Jenkins, J. J. Acoustic information for objects, places, and events. *Persistence and change: Proceedings of the first international conference on event perception*. Hillsdale, NJ: Lawrence Erlbaum Associates. 1985.

[21] Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N. and Neuhoff, J. "Sonification Report: Status of the field and research agenda", ICAD technical report, 1999.

[22] Lingraphica. http://www.lingraphicare.com/. 2005

[23] Ma, X., Boy-Graber, J., Nikolova, S., and Cook, P. Speaking Through Pictures: Images vs. Icons. *Proc. ASSETS09*. 2009.

[24] Ma, X. and Cook, P. How Well do Visual Verbs Work in Daily Communication for Young and Old Adults? In *Proc. CHI 2009*, ACM Press, 2009.

[25] Moore, B. C. J. (ed.). *Handbook of perception and cognition*: Vol. 6. Hearing. 1995.

[26] Moore, B. C. J. An introduction to the psychology of hearing. 4th ed. Orlando, FL: Academic Press. 1997.

[27] Mynatt, J. Designing with Auditory Icons: How Well do We Identify Auditory Cues? Proc. *CHI'94*. 269-270. 1994.

[28] Natural Language Toolkit. http://www.nltk.org/. 2009.

[29] Patterson, R, and Milroy, R. Auditory warnings on civil aircraft: The learning and retention of warnings. *MRC Applied Psychology Unit*. Cambridge, England. 1980.

[30] Pereverzev, S. V., Loshak, A., Backhaus, S., Davis, J. C., and Packard, R. E. Quantum oscillations between two weakly coupled reservoirs of superfluid 3He, *Nature* 388, 449-451. 1997.

[31] Scavone, G., Lakatos, S., Cook, P., and Harbke, C. Perceptual Spaces for Sound Effects Obtained with an Interactive Similarity Rating Program. Intl. Symposium on Musical Acoustics, Perugia, Italy. 2001.

[32] Takasaki, T. PictNet: Semantic Infrastructure for Pictogram Communication. In Proc. *Global WordNet Conference 2006*. pp. 279-284. 2006

[33] Tzanetakis, G. and Cook, P. Musical Genre Classification of Audio Signals. In *Proc. IEEE Transaction of Speech and Audio Processing*. 10 (5), 293-302. IEEE Press, 2002.

[34] UWA Psychology. MRC Psycholinguistic Database.2009. http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm.

[35] Van Hell, J. and De Groot, A. Conceptual Representation in Bilingual Memory: Effects of Concreteness and Cognate Status in Word Association. *Bilingualism*, 1(3):193-211. 1998.

[36] Vanderveer, N. J. Ecological acoustics: Human perception of environmental sounds. Dissertation Abstracts International. 40/09B, 4543. 1979.

[37] Visuri, P. J. Multi-variate alarm handling and display. In Proc. *the International Meeting on Thermal Nuclear Reactor Safety*. National Technical Information Service. 1983.

[38] Yost, W., Popper, A, and Fay, R. Auditory Perception of Sound Sources. Springer. 2007

SONIFICATION: CELESTIAL DATA AND POETIC INQUIRY

Neil Leonard

Berklee College of Music,
Electronic Production and Design,
FB #71, 1140 Boylston St. Boston, MA 02445
nleonard@berklee.edu

ABSTRACT

This paper describes a composition/sonification project to be realized by faculty and students from the Electronic Production and Design department (EP/D) at Berklee College of Music in Boston. The goal of the project is compose music for a 30-minute interdisciplinary-networked performance to be premiered in Boston, Lyon and Havana involving artists from each city. In the process, artists are examining new modes of expression and the construction of knowledge and artistic dialog. Kelly Snook, Ph.D. Astrophysicist, Division of Solar System Exploration, NASA Goddard Spaceflight Center is working with the group to choose scientific data for sonification including compelling new planetary science and solar system data.

1. INTRODUCTION

The composition explores ancient questions related to the poetic use of pattern, symmetry and ratios. The piece leverages our study of the works of Pythagoras, Hildegard of Bingen, Bartok, Coltrane and Ikeda, all of whom explored these ideas. This resultant composition furthers my work in the sonification of naturally occurring patterns that I started in 1986 while working with Hubert Hohn, Director of the Computer Arts Center at Massachusetts College of Art.

An example composition is *Nocturnal Sounds from Hohle Fels* (2009) is for alto saxophone and laptop. The computer performs real-time signal processing and executes computer-driven models for improvisation, based in the Max/MSP programming environment. The audio excerpt provided for this conference uses a vocal performance of Vincenzo Galilei's madrigal *In Exitu Israel* (sung by Alessandro Carmignani) that was reworked using the partial editing resources of SPEAR, time-stretching algorithms of MetaSynth and the experimental use of convolution reverbs. In addition, other sections of *Nocturnal Sounds from Hohle Fels* feature 'glitch' sounds derived from my composition *M87* (1995) from my solo CD *Timaeus*. *M87* is named after a giant elliptical galaxy photographed by the NASA Hubble Space Telescope, and it is believed by some to be a supermassive black hole. The Hubble photos shows a 5000 light-year long jet stream made up of electrons being ejected outward at near light-speed. This 'mash-up' of *M87* was made using Max/MSP patches created by myself and EP/D alumni Gadi Sassoon and Edward Loveall.

In spring 2010, I formed the Global Sonification Network Ensemble (GSNE), to explore sonification and aesthetic issues with students. The GSNE made its debut performance for the live broadcast to the 25th anniversary of VideoFormes international video art and digital culture festival; Clermont-Ferrand, France. The ensemble included Berklee faculty, students and alumni: John Hull, Jinku Kim, Neil Leonard, Daniel Piccione, Enrico de Trizio, Pierce Warnecke and School of the Museum of Fine Arts students: Daniel Cevallos, Shane Butler, Merideth Hillbrand.

2. COMPOSITION

The present task is to expand on selected poetic ideas using scientific models with special attention to data that is uniquely valuable to scientists working in sonification research.

2.1 Spatialization

One immediate undertaking is to expand on the celestial motion model for mixing that was explored in *Partita Tripla con Galilei* composed by composers Maura Capuzzo, Marco Braggion and myself and coordinated by Nicola Bernardini for the "Giornata dell' Ascolto" event in Padova, Italy, 2009. For this installation, Bernardini created an automated mix in Csound that featured a spiraling effect among the six-speaker array that was installed along the perimeter the main Piazza in Padova.

The vocal except presented here is a dense collage of processed samples, rendered in stereo for the conference proceedings. In fact, the piece demands a more choral-like sound dispersion model, where each voice can emanate from a discreet source. With the use of 12.2 audio or a similar 360 diffusions system, the group will explore the automated localization of samples, or even partials, based on solar system, galaxy and other celestial motion.

2.2 Pitch material

In 1993, using non-linear iterative functions almost exclusively, I composed weekly theme music for Hubert Hohn's program "Chaos and Order" (60 computer generated pieces in total), broadcast on MCET interactive educational television. My interest in the sonification of non-linear iterative functions was renewed when Dr. Snook introduced me to Joachim Goßmann's program "Audio Fraktal"[1]. Of particular interest is the use of "Audio Fraktal" to dynamically construct pitched sequences and harmonies outside of equal tempered tuning. The real-time construction of non-equal tempered harmonies has long been an integral part of how I process my saxophone in concert. The creative use of fractal data to synthesize related harmonies in real-time, or to provide a control source for real-time audio pitch shifting potentially expands these new modes of expression.

2.3 Noise

Noise has long been a key artistic resource for musicians. In his article "Noh and Transcendence," composer Toru Takemitsu points out "On examination, we find that the Japanese prefer an artistic expression close to nature while the Westerner treasures an artificial expression that is not part of nature. This is true of the Japanese preferences in sound. Historically Western music has striven to eliminate noise. On the periphery of Western music we find folk and tribal music, which creates unusual sounds that include noise[2]." In short, noise has played an increasingly important role in Western music as well. It was formally embraced by futurists, Dadaists, mid-20th century avant-garde composers and present schools of pop-musicians[3].

Sonification of scientific and solar system data provides new and rich artistic resources for the exploration of noise as a critical resource. Perhaps the "music of the spheres" or "celestial harmonies" can be creatively coupled with "noise of the spheres" as artists Tony Oursler,Constance DeJong and I suggested in our collaborative piece *Relatives* (1989) that used audio/video noise to make a poetic reference to the big bang - the most ancient sonic icon[4].

3. CONCLUSION

Artists evolve through continued experimentation, curiosity and exposure to new tools. The field of sonification provides new and powerful resources for artistic use. Poetic exploration of these sources calls for an ongoing dialog between artists and scientists. Innovations in sonic arts require new ways of listening and artist working at this intersection of disciplines are exposed not only to new modes of expression, but can also learn to listen to nature from the perspective of specialists from outside of the music guild, thus transforming their work at the most fundamental level.

4. ACKNOWLEDGMENT

Special thanks to John Hull for mixing and logistical support for this presentation and Julia Easterlin for performing additional vocal excerpts. I am indebted to Hubert Hohn for introducing me to the sonification of non-linear iterative functions when we worked together at Massachusetts College of Art. A very special thanks to Nicola Bernardini and Sergio Durante for their role in the realization of *Partita Tripla con Galilei* that provided a framework to test some of the compositional and conceptual ideas, presented in this paper. I am extremely grateful for the ongoing support of the EP/D department at Berklee College of Music and support for this project from the office of Faculty Development in the form of a Faculty Led Initiative in Innovation (Fly) grant.

5. REFERENCES

[1] J. Goßmann: *“Towards An Auditory Representation of Complexity,”* in Proc. of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, Limerick, Ireland, July, 2005.

[2] T. Takemitsu: *“Confronting Silence,”* Fallen Leaf Press, Saint Paul MN, USA, 1995.

[3] D. Kahn, G. Whitehead, “*Wireless Imagination: Sound, Radio, and the Avant-Garde*,” The MIT Press, Boston, MA, USA,1994.

[4] R. Schafer, *"The Soundscape. Our Sonic Environment and the Tuning of the World,"* Destiny Books, Rochester, VT, USA,1993.

SONIFICATION: CELESTIAL DATA AND POETIC INQUIRY

Neil Leonard

Berklee College of Music,
Electronic Production and Design,
FB #71, 1140 Boylston St. Boston, MA 02445
nleonard@berklee.edu

ABSTRACT

This paper describes a composition/sonification project to be realized by faculty and students from the Electronic Production and Design department (EP/D) at Berklee College of Music in Boston. The goal of the project is compose music for a 30-minute interdisciplinary-networked performance to be premiered in Boston, Lyon and Havana involving artists from each city. In the process, artists are examining new modes of expression and the construction of knowledge and artistic dialog. Kelly Snook, Ph.D. Astrophysicist, Division of Solar System Exploration, NASA Goddard Spaceflight Center is working with the group to choose scientific data for sonification including compelling new planetary science and solar system data.

1. INTRODUCTION

The composition explores ancient questions related to the poetic use of pattern, symmetry and ratios. The piece leverages our study of the works of Pythagoras, Hildegard of Bingen, Bartok, Coltrane and Ikeda, all of whom explored these ideas. This resultant composition furthers my work in the sonification of naturally occurring patterns that I started in 1986 while working with Hubert Hohn, Director of the Computer Arts Center at Massachusetts College of Art.

An example composition is *Nocturnal Sounds from Hohle Fels* (2009) is for alto saxophone and laptop. The computer performs real-time signal processing and executes computer-driven models for improvisation, based in the Max/MSP programming environment. The audio excerpt provided for this conference uses a vocal performance of Vincenzo Galilei's madrigal *In Exitu Israel* (sung by Alessandro Carmignani) that was reworked using the partial editing resources of SPEAR, time-stretching algorithms of MetaSynth and the experimental use of convolution reverbs. In addition, other sections of *Nocturnal Sounds from Hohle Fels* feature 'glitch' sounds derived from my composition *M87* (1995) from my solo CD *Timaeus*. *M87* is named after a giant elliptical galaxy photographed by the NASA Hubble Space Telescope, and it is believed by some to be a supermassive black hole. The Hubble photos shows a 5000 light-year long jet stream made up of electrons being ejected outward at near light-speed. This 'mash-up' of *M87* was made using Max/MSP patches created by myself and EP/D alumni Gadi Sassoon and Edward Loveall.

In spring 2010, I formed the Global Sonification Network Ensemble (GSNE), to explore sonification and aesthetic issues with students. The GSNE made its debut performance for the live broadcast to the 25th anniversary of VideoFormes international video art and digital culture festival; Clermont-Ferrand, France. The ensemble included Berklee faculty, students and alumni: John Hull, Jinku Kim, Neil Leonard, Daniel Piccione, Enrico de Trizio, Pierce Warnecke and School of the Museum of Fine Arts students: Daniel Cevallos, Shane Butler, Merideth Hillbrand.

2. COMPOSITION

The present task is to expand on selected poetic ideas using scientific models with special attention to data that is uniquely valuable to scientists working in sonification research.

2.1 Spatialization

One immediate undertaking is to expand on the celestial motion model for mixing that was explored in *Partita Tripla con Galilei* composed by composers Maura Capuzzo, Marco Braggion and myself and coordinated by Nicola Bernardini for the “Giornata dell' Ascolto” event in Padova, Italy, 2009. For this installation, Bernardini created an automated mix in Csound that featured a spiraling effect among the six-speaker array that was installed along the perimeter the main Piazza in Padova.

The vocal except presented here is a dense collage of processed samples, rendered in stereo for the conference proceedings. In fact, the piece demands a more choral-like sound dispersion model, where each voice can emanate from a discreet source. With the use of 12.2 audio or a similar 360 diffusions system, the group will explore the automated localization of samples, or even partials, based on solar system, galaxy and other celestial motion.

2.2 Pitch material

In 1993, using non-linear iterative functions almost exclusively, I composed weekly theme music for Hubert Hohn's program “Chaos and Order” (60 computer generated pieces in total), broadcast on MCET interactive educational television. My interest in the sonification of non-linear iterative functions was renewed when Dr. Snook introduced me to Joachim Goßmann's program "Audio Fraktal”[1]. Of particular interest is the use of "Audio Fraktal" to dynamically construct pitched sequences and harmonies outside of equal tempered tuning. The real-time construction of non-equal tempered harmonies has long been an integral part of how I process my saxophone in concert. The creative use of fractal data to synthesize related harmonies in real-time, or to provide a control source for real-time audio pitch shifting potentially expands these new modes of expression.

2.3 Noise

Noise has long been a key artistic resource for musicians. In his article "Noh and Transcendence," composer Toru Takemitsu points out "On examination, we find that the Japanese prefer an artistic expression close to nature while the Westerner treasures an artificial expression that is not part of nature. This is true of the Japanese preferences in sound. Historically Western music has striven to eliminate noise. On the periphery of Western music we find folk and tribal music, which creates unusual sounds that include noise[2]." In short, noise has played an increasingly important role in Western music as well. It was formally embraced by futurists, Dadaists, mid-20th century avant-garde composers and present schools of pop-musicians[3].

Sonification of scientific and solar system data provides new and rich artistic resources for the exploration of noise as a critical resource. Perhaps the "music of the spheres" or "celestial harmonies" can be creatively coupled with "noise of the spheres" as artists Tony Oursler,Constance DeJong and I suggested in our collaborative piece *Relatives* (1989) that used audio/video noise to make a poetic reference to the big bang - the most ancient sonic icon[4].

3. CONCLUSION

Artists evolve through continued experimentation, curiosity and exposure to new tools. The field of sonification provides new and powerful resources for artistic use. Poetic exploration of these sources calls for an ongoing dialog between artists and scientists. Innovations in sonic arts require new ways of listening and artist working at this intersection of disciplines are exposed not only to new modes of expression, but can also learn to listen to nature from the perspective of specialists from outside of the music guild, thus transforming their work at the most fundamental level.

4. ACKNOWLEDGMENT

Special thanks to John Hull for mixing and logistical support for this presentation and Julia Easterlin for performing additional vocal excerpts. I am indebted to Hubert Hohn for introducing me to the sonification of non-linear iterative functions when we worked together at Massachusetts College of Art. A very special thanks to Nicola Bernardini and Sergio Durante for their role in the realization of *Partita Tripla con Galilei* that provided a framework to test some of the compositional and conceptual ideas, presented in this paper. I am extremely grateful for the ongoing support of the EP/D department at Berklee College of Music and support for this project from the office of Faculty Development in the form of a Faculty Led Initiative in Innovation (Fly) grant.

5. REFERENCES

[1] J. Goßmann: *"Towards An Auditory Representation of Complexity,"* in Proc. of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, Limerick, Ireland, July, 2005.

[2] T. Takemitsu: *"Confronting Silence,"* Fallen Leaf Press, Saint Paul MN, USA, 1995.

[3] D. Kahn, G. Whitehead, "*Wireless Imagination: Sound, Radio, and the Avant-Garde*," The MIT Press, Boston, MA, USA,1994.

[4] R. Schafer, *"The Soundscape. Our Sonic Environment and the Tuning of the World,"* Destiny Books, Rochester, VT, USA,1993.

TURBULENT FLOW: AN ELECTROACOUSTIC WORK COMPOSED WITH SONIFIED DATA

Brent Lee

Noiseborder Multimedia Performance Lab,
University of Windsor,
Windsor, Ontario, Canada N9B 3P4
brentlee@uwindsor.ca

ABSTRACT

Turbulent Flow is a 12-minute electroacoustic work incorporating various strategies for data sonification in its composition. A data set derived from measuring the fluctuating wind velocity of a turbulent airflow was mapped onto a variety of parameters in the design of timbres and in the sequencing of musical events. This mapping includes pitch, rhythm, amplitude, panning, filter frequencies, and various parameters of FM and granular synthesis instruments. The translation of raw data to usable musical information was done using Csound software, while the multi-tracking and triggering of events as well as the processing of the sonic material was achieved with Ableton Live. While this work is part of a larger project to investigate the aural perceptibility of patterns in complex data sets, Turbulent Flow uses the interesting sonic results of these experiments as source material for an artistic work, freely ordering and combining fragments into rich electroacoustic textures.

Index of Authors