



THE 22nd ANNUAL INTERNATIONAL CONFERENCE ON AUDITORY DISPLAY
SONIC INFORMATION DESIGN
Australian National University, Canberra 3-7 July 2016

SONIC INFORMATION DESIGN

PROCEEDINGS

OF THE 22nd ANNUAL

INTERNATIONAL CONFERENCE ON AUDITORY DISPLAY

David Worrall (Editor)

School of Music, Australian National University
Canberra 3-7 July 2016

<http://www.icad.org/icad2016/>

DOI: 10.21785/icad2016.00

ISBN 0-9670904-3-1

SONIC INFORMATION DESIGN

PROCEEDINGS

OF THE 22nd ANNUAL

INTERNATIONAL CONFERENCE ON AUDITORY DISPLAY

David Worrall (Editor)

School of Music, Australian National University
Canberra 3-7 July 2016

<http://www.icad.org/icad2016/>

DOI: 10.21785/icad2016.00

ISBN 0-9670904-3-1

ICAD2016

ORGANIZING COMMITTEE

Co-Chairs	David Worrall and Stephen Barrass
Papers Reviewing Chairs	David Worrall and Stephen Barrass
Student ThinkTank	Hiroko Terasawa
Workshops	Derek Brock
Installations	Stephen Barrass
Concert	Charles Martin
Diversity Workshop	Hiroko Terasawa, Visda Goudarzi, Areti Andreopoulou
Finances & Social Events	Stephen Barrass
Volunteer Coordination	Ellen Falconer
Volunteers	Millie Watson, Ben Harb, Benjamin Drury, Charles Martin
Proceedings Editor	David Worrall
Proceedings Publishing	Rebekah Teo, Derek Brock, Matti Gröhn
Website	David Worrall

ICAD BOARD (2016)

EXECUTIVE MEMBERS

David Worrall	President, Columbia College Chicago, USA
Derek Brock	Treasurer, Naval Research Laboratory, USA
Areti Andreopoulou	Secretary, LIMSI-CNRS, France

EMERITUS MEMBERS

Gregory Kramer	Founder, Clarity and Metta Foundation, USA
Matti Gröhn	Professional Cooperative Polkuverkosto, Finland

ELECTED MEMEBERS

Brian Katz	LIMSI-CNRS, France
Michael Musick	New York University, USA (Student Representative)
Paul Vickers	Northumbria University, United Kingdom

EX-OFFICIO MEMEBERS

Mark Ballora	The Pennsylvania State University, USA (2017 Conference Chair)
Visda Goudarzi	Institute of Electronic Music and Acoustics, Austria
S. Camille Peres	Texas A&M University, USA
Hiroko Terasawa	University of Tsukuba, Japan

TABLE OF CONTENTS

PRELIMINARIES

Organising Committee ICAD 2016	ii
Welcome: ICAD 2016 Co-Chairs - David Worrall and Stephen Barrass	3
Programme	5
Keynotes	7

ORAL PAPERS

Sonic Information Design

AUTHOR	TITLE	PAGE
Daniel Verona and Camille Peres	<i>Exploring a Task-Analysis-Based Approach to Sonification Design</i>	11
Takuya Yamauchi	<i>Designing Sound Representations for Responsive Environments</i>	15
William Martens, Philip Poronnik and Darren Saunders	<i>Hypothesis-Driven Sonification of Proteomic Data Distributions Indicating Neurodegradation in Amyotrophic Lateral Sclerosis</i>	21
Michael Quinton, Iain McGregor and David Benyon	<i>Sonifying the Solar System</i>	28
Ivica Bukvic	<i>3D Time-Based Aural Data Representation Using D4 Library's Layer Based Amplitude Panning Algorithm</i>	36
Keith Nesbitt, Paul Williams, Patrick Ng, Karen Blackmore and Ami Eidels	<i>Designing Informative Sound to Enhance a Simple Decision Task</i>	46

Aesthetics, Philosophy, and Culture of Auditory Displays

AUTHOR	TITLE	PAGE
Teresa Connors	<i>The Aesthetics of Causality: A Descriptive Account into Ecological Performativity: A Creative Research Practice</i>	55
Marc St. Pierre and Milena Droumeva	<i>Sonifying for Public Engagement: A Context-Based Model for Sonifying Air Pollution Data</i>	60
Dugal McKinnon, James Murphy and Mo Zareei	<i>Lost Oscillations: Exploring a City's Space and Time with an Interactive Auditory Art Installation</i>	65

Body and Mind

AUTHOR	TITLE	PAGE
Steven Landry, Yuanjing Sun, Darnishia Slade and Myoungsoon Jeon	<i>Tempo-Fit Heart Rate App: Using Heart Rate Sonification as Exercise Performance Feedback</i>	73
Letizia Gionfrida, Agnieszka Roginska and Kent Friedman	<i>The Triple Tone Sonification Method to Enhance the Diagnosis of Alzheimers and Dementia</i>	77
Ruimin Zhang, Jaclyn Barnes, Joseph Ryan, Myoungsoon Jeon, Chung Hyuk Park and Ayanna Howard	<i>Musical Robots for Children with ASD Using a Client-Server Architecture</i>	83
Ridwan Ahmed Khan, Ram Avvari, Katherine Wykovich, Pooja Ranay and Myoungsoon Jeon	<i>Lifemusic: Reflection of Life Memories by Data Sonification</i>	90

TABLE OF CONTENTS (continued)

ORAL PAPERS

Mobiles and Wearables

AUTHOR	TITLE	PAGE
Jason Sterkenburg, Steven Landry, Myoungsoon Jeon and Joshua Johnson	<i>Towards An In-Vehicle Sonically-Enhanced Gesture Control Interface: A Pilot Study</i>	95
John Dyer, Paul Stapleton and Matthew Rodger	<i>Sonification of Movement for Motor Skill Learning in a Novel Bimanual Task: Aesthetics and Retention Strategies</i>	99
Greg Schiemer	<i>Satellite Gamelan: microtonal sonification using a large consort of mobile phones</i>	103

Tasks and Attention

AUTHOR	TITLE	PAGE
Anna Bramwell-Dicks, Helen Petrie and Alistair Edwards	<i>Can Listening to Music Make You Type Better? The Effect Of Music Style, Vocals And Volume On Typing Performance</i>	109
Toby Gifford	<i>Tuning into the Task: Sonic Environmental Cues and Mental Task Switching</i>	117
Clayton Rothwell, Griffin Romigh and Brian Simpson	<i>Aurally Aided Visual Search on a Monitor: Spatialized Virtual Audio Cues for Desktop Computers</i>	123
Derek Brock, Christina Wasylshyn, and Brian McClimens	<i>Word Spotting in a Multichannel Virtual Auditory Display at Normal and Accelerated Rates of Speech</i>	130

3D Audio and Spatial Sound

AUTHOR	TITLE	PAGE
Amit Barde, Matt Ward, William Helton, Mark Billinghamurst and Gun Lee	<i>A Bone Conduction Based Spatial Auditory Display As Part of a Wearable Hybrid Interface</i>	139
Nandini Iyer, Eric Thompson and Brian Simpson	<i>Response Techniques And Auditory Localization Accuracy</i>	147
Griffin Romigh, Brian Simpson and Nandini Iyer	<i>In Ear to Out There: A Magnitude Based Parameterization Scheme for Sound Source Externalization</i>	153
Ziqi Fan, Yunhao Wan and Kyla McMullen	<i>Quantitatively Validating Subjectively Selected HRTFs for Elevation and Front-Back Distinction</i>	158

POSTERS

AUTHOR	TITLE	PAGE
Mark Ballora	<i>Music of Migration and Phenology: Listening to Counterpoints of Musk Ox and Caribou Migrations, and Cycles Of Plant Growth</i>	169
Amit Barde, William Helton, Gun Lee and Mark Billinghamurst	<i>Binaural Spatialisation Over a Bone Conduction Headset: Elevation Perception</i>	173
Allan Coop	<i>Sonification, Musification and Synthesis of Absolute Program Music</i>	177
Andrea Genovese, Jordan Juras, Christopher Miller and Agnieszka Roginska	<i>Investigation of ITD Symmetry in Measured HRIRs</i>	184

TABLE OF CONTENTS (continued)

POSTERS

AUTHOR	TITLE	PAGE
Daeyoung Jang, Jae-Hyoun Yoo and Taejin Lee	<i>Implementation and Evaluation of 10.2 channel Microphone for UHDTV Audio</i>	191
Steven Landry, David Tascarella, Myoungsoon Jeon and Maryam Fakhrhosseini	<i>Listen to Your Drive: Sonification Architecture and Strategies for Driver State and Performance</i>	196
Fiore Martin, Oussama Metatla, Tony Stockman and Nick Bryan-Kinns	<i>Spectrum Analyser</i>	199
Ryan McGee and David Rogers	<i>Musification of Seismic Data</i>	201
Katieanna Wolf and Reid Oda	<i>MalLo March: A Live Sonified Performance with User Interaction</i>	205

INSTALLATIONS

TITLE	PRESENTER	DESCRIPTION	PAGE
<i>Flight Variant</i>	Teresa Connors and Andrew Denton	Audiovisual installation.	211
<i>Terramomentum</i>	Ryan McGee and David Rogers	A low-frequency sound installation for experiencing earthquake tremors.	214
<i>Lux Mix</i>	Michael Norris	An interactive sonification installation that features modulated light as a medium for carrying audio signals.	215
<i>Native Modulations</i>	Andrew Zylstra	Scribbly-gum sonifications.	217
<i>Acoustic Sonifications</i>	Stephen Barrass	Research Through Design into Acoustic Sonification. A series of 3D printed data forms that map shape to acoustics.	218

CONCERT

TITLE	DESIGNERS	PERFORMERS	DESCRIPTION	PAGE
<i>Transposed Dekany</i>	Greg Schiemer	ICAD participants and the ANU EMS	A microtonal performance using the Satellite Gamelan app.	223
<i>BioLogging Retrofit</i>	Nigel Helyer	ANU EMS	A mechanical/performative sonification of environmental knowledge from Antarctic bio-logging data.	225
<i>MalLo March</i>	Katie Wolf and Reid Oda	Katie Wolf, Steven Landry and Daniel Verona	An audience-customised performance with MalLo, a predictive percussion instrument.	228
<i>Hearing A Gene of Hearing</i>	Stephen Barrass	ANU EMS	Percussion Performance of a Gene of Hearing for 4 species on 4 Ubang clay drums.	231
<i>PhaseRings for 6 iPads and Ensemble Director Agent</i>	Charles Martin	ANU EMS	An improvised iPad performance where performers are tracked and guided by an Ensemble Director Agent.	232
<i>Atom Tone</i>	Jirí Suchánek	Jirí Suchánek	A live sonification and aesthetic exploration of atomic spectra.	234

TABLE OF CONTENTS (continued)

WORKSHOPS

PRESENTER	TITLE	PAGE
Assoc Prof Ivica Ico Bukvic Virginia Tech	<i>Using A Layer-Based Amplitude Panning (LBAP) Audio Spatialization Algorithm</i>	239
Prof Simon Carlile, Starkey Hearing Technologies	<i>Dare to Design Hearables</i>	239
Prof Angelina Russo University of Canberra	<i>Computer Knitted Data Scarf</i>	240
Prof David Worrall Audio Arts and Acoustics Department Columbia College Chicago	<i>Data Sonification using Python and Csound</i>	240
Dr George Poonkhin Khut UNSW Australia Art & Design	<i>Neurofeedback and Contemplative Interaction</i>	241
Tim Barrass, Mozzi	<i>Hack Your ICAD NameBadge</i>	241
Dr Greg Schiemer	<i>Biologging Retrofit</i>	241

PRELIMINARIES

Welcome to ICAD from the Conference Co-Chairs

Stephen Barrass and David Worrall, Monday 4 July 2016

We would like to acknowledge the Ngunnawal people who are the traditional custodians of this land on which we are meeting, and pay respect to the Elders of the Ngunnawal Nation both past and present. We extend this respect to all Aboriginal and Torres Strait Islander peoples in attendance today.

For the past 8 weeks in Canberra we have been counting down to this date, not only because it is the Australian Federal Election, but because it is ICAD 2016. Welcome to Canberra, the Capital City of Australia, which is only just 100 years old. It is one of a handful of “designed cities” in the world, alongside Washington and Brasilia. The designer, Sir Walter Burley Griffin, was a modernist who believed that the geometry of the city shaped the lives of its inhabitants, and the many roundabouts may also explain all the going round in circles we have experienced during the election campaign. When the winning design was announced the public were invited to suggest names for this new city, and some responses included Commonwealth Circular City, Federalia, Eden, Paradise, Harmony, Unison, Pacifica, Frazer Roo, Wheatwoolgold and Sydmeladperbriso (which is an amalgam of every other Australian capital). Finally, Canberra was declared the winner, which means ‘meeting place’ in the local indigenous Ngunnawal language. ICAD visitors will be welcomed with a “smoking ceremony” which is an indigenous Welcome to Ngunnawal Country on Monday evening. This welcome will be reiterated by an Aussie pub welcome at the Wig and Pen microbrewery with custom winter brews accompanied by Kangaroo pies and vegetarian pasties made from native Warrigal greens and Bush Tomatoes.

Since Canberra is a Design City it is very appropriate that the theme for ICAD this year is Sonic Information Design. This theme aims to integrate design as a method for research and discovery in the field of Auditory Display. The theme has grown out of a workshop at ICAD 2014 in Atlanta, and was developed further in a keynote at ICAD 2015 in Graz. Like other design disciplines, Sonic Information Design has the aspiration that artificial sounds may be designed to make the world a better place. Sonic Information Design takes a user-centred view of the relationship between artefacts, those that are affected by them, and the social contexts in which they occur. A Design orientation pays particular attention to user experience—including physical, cognitive, emotional, and aesthetic issues; the relationship between form, function, and content; and emerging concepts such as fun, playfulness and design futures. Practice-based research is considered as a generative process of exploration, speculation and discovery, with outcomes that can be provisional, contingent and aspirational, while aiming for richer, more situated understandings that lead to the advancement of knowledge and the proliferation of new realities.

The Keynotes for ICAD 2016 have been chosen to expand on this theme within the context of Auditory Display. Dr. Viveka Turnbull Hocking's design-led research engages cross-disciplinary conversations and collaborations through the area of design and the built environment. Dr. George Poonkhin Khut 's body-focussed artworks use biofeedback technologies to re-frame experiences of embodiment, health and subjectivity. A/Prof. Adrian KC Lee's research investigates the neural network involved in auditory scene analysis that provides a basis for Stream-based

Sonification where the theory of Auditory Scene Analysis is used to design Auditory Displays. Prof. Simon Carlile, Senior Director of Research at Starkey Hearing Technologies, will present Industry Insight into innovations in Hearables which extend hearing aid technology to new applications.

The papers programme includes 24 peer reviewed papers which will be presented beginning on Monday, and a Posters session on Tuesday afternoon. On Tuesday evening social activities continue with the ICAD jam session at Smith's Alternative Book Store.

On Wednesday we have introduced a recreational break, to foster social networking while exploring the recreational and cultural aspects of Canberra, or to take a collaborative workshop. The ICAD Concert on Wednesday evening will bring everyone back together to swap experiences. Local attractions include Old and New Parliament House, the National Museum of Australia, the National Gallery of Australia, or the National Portrait Gallery where keynote George Khut will present a workshop on Neurofeedback and Contemplative Interaction. Other workshops include collaborating in a proposal for a Hearables application, designing your own Computer Knitted Scarf from data patterns, programming the Mozzi wearable sonification synthesiser, learning Csound+Python, and 3D panning software for multi-speaker arrays.

The concert will be open to the general public, with pieces that explore emerging themes of mobile platforms and collaborative sonification through works for an iPad Ensemble, mobile phones, audience interaction on their own mobile devices, a sonification of Antarctic datasets on four Music Boxes, hearing genes of hearing on four clay ubang drums, and a live electronic performance of atomlc data. A public exhibition of sound installations around the School of Music includes the Pacific Bell Tower which chimes to resonate the seismic activity around the Pacific on the hour, Terramomentum where blindfolded listeners lie on the floor to hear 5 subwoofers emit audifications of earthquakes, an interactive sonification of astronomical data encoded in light variations, a sonification of scribbly-gum markings, and 3D printed acoustic objects constructed from blood pressure data.

The Banquet will provide a social space to discuss the days activities and the concert. The venue is A. Baker which is a highly regarded restaurant in walking distance from the School of Music. The seasonal menu is sourced from the Canberra district and includes local wines. It is the Truffle Season in Canberra, and who knows, we may get lucky.

The conference papers programme will continue on Thursday, and conclude with the traditional open-mic session that afternoon. Then, if you are feeling fit you can join the ICAD snow bus for a one day return trip to the Australian Alps, to ski through the snow-gums. Your ICAD registration includes a 'data beanie' knitted from possum fur and designed to keep your head warm in the frosty mid winter mornings in Canberra. This beanie is computer knitted from punch-cards that encode data logged from seals diving under the antarctic. The ski trip will be your opportunity to test the ICAD data beanie in earnest as protection against the cold and snow in downhill skiing conditions ! We hope you have a great time at ICAD 2016 in Canberra.



ICAD2016 PROGRAMME

v0.11 SUNDAY 3 July 2016						
Nr	TIME	PLACE	EVENT	DETAILS	DESCRIPTION	
0	9-17	LT2	STUDENT THINK TANK Chair:Hiroko Teresawa	Presentations and Discussions by Research Students with a faculty panel of international researchers.		
MONDAY 4 July 2016						
1	8.30	SofM L5	REGISTRATION			
2	9.00	LT3	WELCOME	ICAD 2016 Co-Chairs : David Worrall and Stephen Barrass		
3	9.30	LT3	ORAL PAPERS 1	Sonic Information Design		
7	11.30	LT3	KEYNOTE ADDRESS 1	Viveka Turnbull Hocking	<i>A Design-led Approach: opening up a Cross-disciplinary Discourse into Design research</i>	
8	12.30	YouChoose	LUNCH			
9	14.00	LT3	ORAL PAPERS 2	Aesthetics, Philosophy, and Culture of Auditory Displays		
12	15.30	Café	COFFEE			
13	16.00	LT3	ORAL PAPERS 3	Sonic Information Design (cont'd)		
16	17.30	Wig & Pen Pub (School of Music)	RECEPTION - Free to conference attendees!			
TUESDAY 5 July 2016						
17	9.00	LT3	ORAL PAPERS 4	Body and Mind		
21	11.00	Café	COFFEE			
22	11.30	LT3	KEYNOTE ADDRESS 2	George Khut	<i>Biofeedback and Beyond</i>	
23	12.30	YouChoose	LUNCH			
24	13.30	Foyer	POSTERS			
25	15.00	Café	COFFEE			
26	15.30	LT3	INDUSTRY REPORT	Simon Carlile	<i>The What and Why of Hearables</i>	
27	16.00	LT3	ORAL PAPERS 5	Mobiles and Wearables		
30	18.00+		JAM SESSION at Smiths Alternative Bookshop			
WEDNESDAY 6 July 2016						
Level5			WORKSHOPS 1 (&/or tourism)			
52	11.00	Café	COFFEE			
36	11.30		WORKSHOPS 1 (cont'd)			
37	12.30	YouChoose	LUNCH			
38	13.30	L6	ICAD BOARD MEETING - Board Room			
			WORKSHOPS 2 (afternoon)(&/or tourism)			
52	11.00	Café	COFFEE			
44	15.30		ICAD BOARD MEETING (cont'd)			
45	15.30	various	CONCERT PREPARATION			
46	17.30	Stage	CONCERT - Free to conference attendees! Llewellyn Hall Stage			
47	19.30	A.Baker	BANQUET Free to conference attendees!			
THURSDAY 7 July 2016						
48	9.00	LT3	ORAL PAPERS 6	Tasks and Attention		
52	11.00	Café	COFFEE			
53	11.30	LT3	KEYNOTE ADDRESS 3	Adrian KC Lee	<i>Auditory Scene Analysis, Object Formation and Selection and their Implications with respect to Stream-Based Sonification</i>	
54	12.30	YouChoose	LUNCH - Gender Diversity Discussion Group (Kingsland Room, School of music, Level 6)			
55	13.30	LT3	ORAL PAPERS 7	3D Audio and Spatial Sound		
59	15.30	Café	COFFEE			
60	16.00	LT3	REPORTS & OPEN MIKE			
61	17.00		CONFERENCE CLOSE			
FRIDAY 8 July 2016						
62		Perisher Valley	Ski through the Gum Trees - Tour led by Stephen Barrass			



ICAD2016 Canberra



Places for ICAD2016 conference attendees



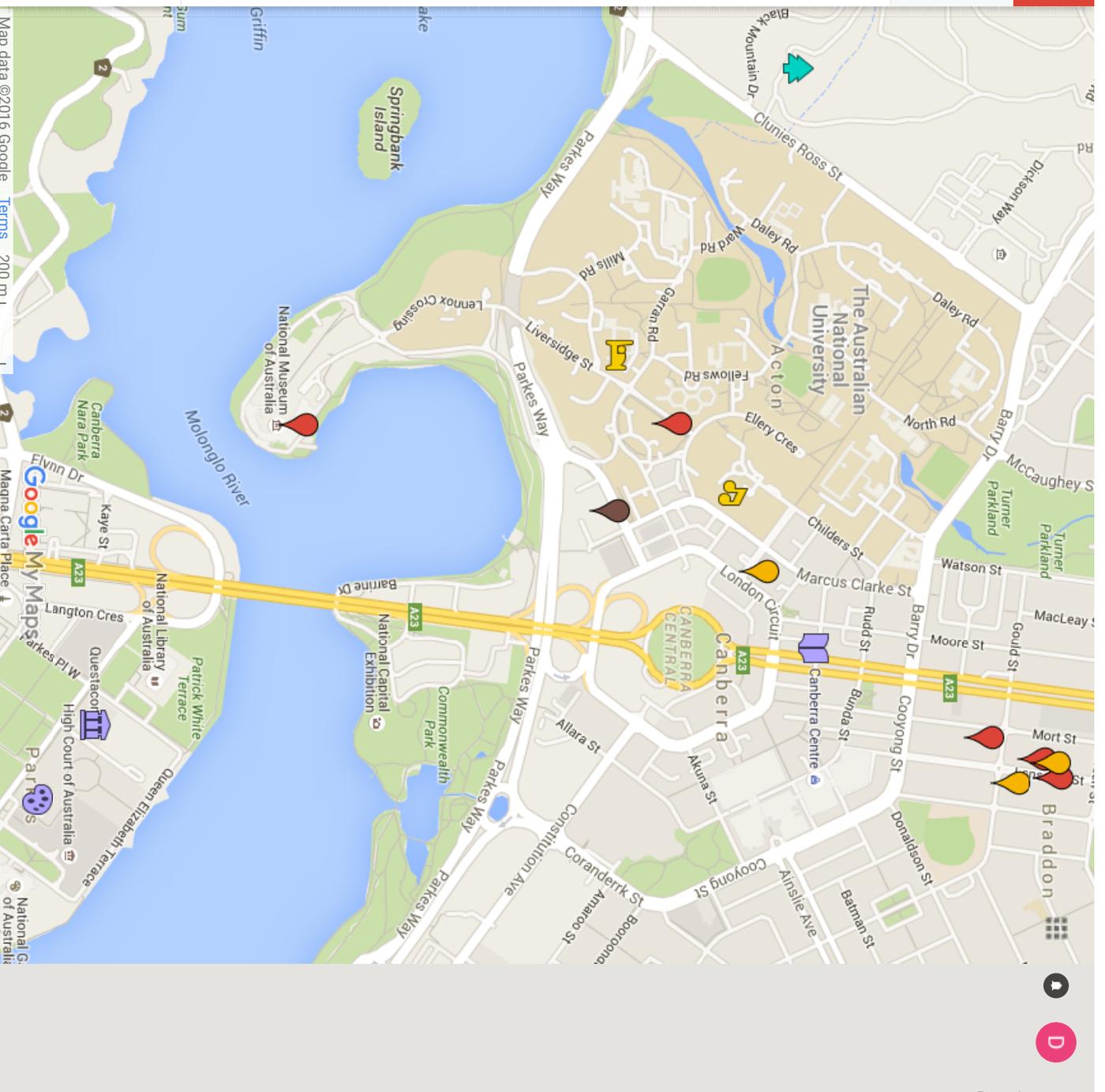
1 view

[SHARE](#) [EDIT](#)



Untitled layer

- ANU School of Music
- Smiths Alternative
- A. BAKER
- Braddon Officeworks
- Debaclé
- Autlyse
- The Cupping Room
- Sweet Bones Bakery and Cafe
- Lonsdale Street Eatery
- National Portrait Gallery
- Questacon
- National Museum of Australia
- National Film and Sound Archive of Austral...
- Australian National Botanic Gardens
- University House



Made with Google My Maps

Map data ©2016 Google



200 m

ICAD2016 KEYNOTE SPEAKERS



Viveka Turnbull Hocking

A Design-led Approach: opening up a cross-disciplinary discourse into design research

Design as a profession emerged out of the industrial revolution. However, the word 'design' has been around much longer and describes an activity that we as humans all do and have done for some time. A clear distinction can be made between uppercase 'Design' as the profession verses lowercase 'design' as the activity in order to highlight the many uses, meanings and applications of the activity of design independent of the fields of Design. Although design may have many forms, there is an overriding commonality in a process concerned primarily with generating 'what could be' rather than 'what is', of which the discipline of Design and our approaches can contribute. This key note address aims to initiate a conversation about the value of Design approaches to the research community and the multi-disciplinary area of sonic information design. To enable such a dialogue this address offers some insights into the characteristics of design: what it is, how it is done and why it might be of significant value; outlining the nature of design-led approaches that are emerging in order to open up a cross-disciplinary discourse into design research.

Dr Viveka Turnbull Hocking's work looks at design-led approaches to research. Her research has included design for social innovation, sustainability and now is exploring design for cross-species cohabitation. Her design-led approach to research engages with cross-disciplinary conversations and collaborations through the area of design and the built environment. She researches through the Fenner School of Environment and Society at the Australian National University and teaches in the School of Design and Architecture at the University of Canberra.



George Khut

Biofeedback and beyond

Affordable wearable bio-sensing technologies and mobile operating systems capable of supporting realtime displays are opening up new possibilities for interactive sound design and composition for biofeedback interactions. George Khut will present examples of his work with heart rate and brainwave interactions, and consider historical precedents for this work in the field of electronic art and computer music, along with more recent developments in the area of commercial "wellness apps", and outline opportunities and challenges for real-time sonification and sonic interaction design.

Dr George Poonkhin Khut is an artist and interaction-designer working across the fields of electronic art, design and health. George's body-focussed artworks use biofeedback technologies to re-frame experiences of embodiment, health and subjectivity. In 2012 he was the winner of National New Media Art Award, at the Queensland Art Gallery, Gallery of Modern Art (GoMA). Recent projects include the BrightHearts (iOS) app (available on iTunes App Store), "Behind Your Eyes, Between Your Ears" a residency and public laboratory at the National Portrait Gallery, Canberra; and "CUSP: Design into the Next Decade", a group exhibition curated by Object Gallery (Australian Design Centre).



Adrian KC Lee

Auditory scene analysis, object formation and selection and their implications with respect to stream-based sonification

Sound arriving at our ears is the sum of acoustical energy from all of the auditory sources in the environment. In order to make sense of our auditory world, we rely on different stimulus features (e.g., pitch, timbre, spatial cues) to segregate one sound source from another. Furthermore, we can flexibly select one sound stream to listen to (e.g., enjoying the violin melody) or switch attention to another (e.g., following the pizzicato of the cello). The process that enables us to listen to one sound out of many is often referred to as auditory scene analysis and, remarkably, while most humans and other animals can do it seamlessly, it is still a difficult challenge for the most sophisticated computer algorithm. In this talk, I will introduce different aspects of auditory scene analysis, as well as discuss a closely related challenge-solving the "Cocktail Party Problem." I will discuss how using an auditory object-based attention framework can help elucidate these perceptual and cognitive processes, thereby shedding light on different aspects of sonification (e.g., stream-based sonification).

Adrian KC Lee is an Associate Professor in the Department of Speech & Hearing Sciences and at the Institute for Learning and Brain Sciences at the University of Washington, Seattle, USA. He obtained his bachelor's degree in electrical engineering at the University of New South Wales and his doctorate at the Harvard-MIT Division in Health Sciences and Technology. Dr. Lee's research focuses on developing multimodal imaging techniques to investigate the cortical network involved in auditory scene analysis and attention, especially through designing novel behavioral paradigms that bridge the gap between psychoacoustics and neuroimaging research.

ORAL PAPERS

Sonic Information Design

A TASK-ANALYSIS-BASED EVALUATION OF SONIFICATION DESIGNS FOR TWO sEMG TASKS

S. Camille Peres

Texas A&M University,
Dept. of Environmental and Occupational Health,
Research on the Interface between Humans
and Machines Laboratory
peres@tamu.edu

Daniel Verona

Texas A&M University,
Dept. of Biomedical Engineering,
Research on the Interface Between Humans
and Machines Laboratory
daniel.j.verona@gmail.com

ABSTRACT

This paper presents a brief description of surface electromyography (sEMG), what it can be used for, as well as some of the problems associated with visual displays of sEMG data. Sonifications of sEMG data have shown potential for certain applications in data monitoring and movement training, however there are still challenges related to the design of these sonifications that need to be addressed. Our previous research has shown that different sonification designs resulted in better listener performance for different sEMG evaluation tasks (e.g. identifying muscle activation time vs. muscle exertion level). Based on this finding, we speculated that sonifications may benefit from being designed to be task-specific, and that integrating a task analysis into the sonification design process may help sonification designers identify intuitive and meaningful sonification designs. This paper presents a brief introduction to what a task analysis is, provides an example of how a task analysis can be used to inform sonification design, and outlines future research into a task-analysis-based approach to sonification design.

1. INTRODUCTION

Surface electromyography (sEMG) is a technique for measuring muscle activation onset, muscle activation duration, and muscle exertion level. It is commonly used by researchers [1, 2] and physical therapists [3] as a tool for biofeedback [4], as an index of muscle fatigue [2], as a strength training tool [5], and as a motor learning and rehabilitation tool [6].

Typically, EMG data are displayed visually on a computer screen, and while this display modality can work well for certain applications, it has its limitations. Visual displays force EMG technicians to focus their visual attention on a screen, which limits their mobility and prevents them from focusing on the movements of the subject [7]. Additionally, in sports applications, visual displays of EMG data can overload an athlete's visual capacities while the athlete is learning a particular movement [8]. To address these limitations of visual displays of EMG data, researchers have begun exploring auditory displays of EMG data, primarily in the form of parameter-mapping sonifications. Researchers have found that EMG sonifications show potential for

improving athletic and exercise performance [8] and identifying musculoskeletal disorders [9]. Additionally, sonification of upper extremity movement within a 3-D space has shown potential for improving motor rehabilitation therapies for stroke patients [10].

These findings suggest that sonification has the potential to be an excellent display modality. Sound is intuitive, the human hearing system has excellent temporal and frequency resolution [7], auditory displays do not restrict a researcher to a computer screen, and sound can have a communal aspect as well – that is to say, an auditory display can afford everyone in a room immediate and simultaneous access to the display (a luxury not common to visual displays). However, despite these advantages, sonification presents its own unique challenges – chief among them is the challenge of display design.

Designing sonifications and choosing mappings that are intuitive, meaningful, and listenable is difficult; thus many different parameters of sound (pitch, loudness, tempo, attack time, spatial location, tremolo, harmonic content, etc.) have been evaluated for use in sonification [11, 12, 13]. Despite this research however, there has been a lack of empirical evaluation and comparison of different sonification designs [14].

We have begun to address this in a recent study (Peres et al., under review) by empirically comparing listener performance between six different sonification designs for sEMG data. We found that different sonification designs resulted in better listener performance for two different sEMG evaluation tasks (results from this study are discussed in further detail below in Section 3). These results indicated that different sonification designs may be better suited to different tasks, which led us to speculate that:

1. Task-specific sonification designs may be helpful in creating effective and meaningful sonifications
2. Basing sonification design on a task analysis (a design tool used in Human Factors in HCI) may help sonification designers identify effective and meaningful mappings for a given task

Applying a task analysis to sonification design is not a new concept [15], however it does not seem to be a well-represented approach to sonification design. In order to better understand the effects that a task analysis could have on sonification design and listener performance, we believe it would be beneficial to continue performing empirical comparisons of different sonification designs by comparing task-analysis-based designs (i.e. designs that are tailored to a specific task or function) to classic parameter-based designs that are typically task-agnostic (i.e. designs that map a data



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

set to some set of auditory parameters and are not tailored to a specific task or function).

This paper offers a brief description of what a task analysis is, describes two previous task analyses that we performed for the two sEMG evaluation tasks used in our previous study, and outlines future research into the effects of task-analysis-based sonification design on listener performance.

2. TASK ANALYSIS

As previously mentioned, a task analysis is an analysis and design tool used in fields such as Human Factors and HCI [16, 17]. The role of a task analysis can be viewed in several ways, ranging from the entire front-end pre-design process, to one element of the front end process, to a range of techniques that come into play at different times during design and development [18]. Common to each of these perspectives however is that a task analysis is meant to provide knowledge about the users, their goals in accomplishing the task, their environment, the manual elements of the task, the cognitive elements of the task, the tools used to perform the task, the duration, order, and complexity of the task, as well as any other unique factors pertaining to the task [19]. Task analysis methods were developed primarily as a means for assessing and reducing human error, though the use of these methods has expanded over time [20].

There are many different types of task analysis methods available and one simple way to categorize them is to divide them into action oriented methods and cognitive methods [21]. Action oriented methods (such as the commonly used hierarchical task analysis) focus on observable actions, or identifying, in top down fashion, the goal of the task, as well as the various subtasks and conditions under which those subtasks must be performed in order to achieve the goal. Cognitive methods, on the other hand, focus on analyzing and outlining the unseen mental processes – diagnosis, decision making, problem solving, etc. – that can give rise to human error [21].

Listening to sonifications for the purpose of data monitoring, analysis, or exploration requires that the listener be able to identify certain characteristics of the sounds that are heard and relate those characteristics to various features of the data in order to make judgments about the data. Since this is primarily a cognitive task (not manual), we advocate using cognitive task analysis methods as tools to understand the listener's task. Section 3 below provides an example of how this can be done and what can be learned from applying this tool to sonification design.

3. TASK ANALYSIS OF TWO sEMG EVALUATION TASKS

As previously mentioned, we found in our prior research that different sonification mappings resulted in better listener performance for different tasks. In this study, each participant listened to sonifications of two simultaneous channels of sEMG data. One channel was referred to as Muscle A while the other channel was referred to as Muscle B. Each sonification lasted 10 seconds, and in each sonification, both muscles (A and B) started at rest, contracted at close to the same time, and then returned to rest. We asked each participant to identify two characteristics of the muscle activity represented by the sEMG data: which muscle (A or B) activated first and which muscle (A or B) exhibited a higher exertion level.

For the parameter mapping, the sEMG data being sonified were sampled at 1000 Hz, and then averaged into blocks of 100 data points each, creating 10 averaged data values for each second of sEMG data. These averaged values were then played back using SuperCollider's Triangle wave oscillator at a rate of 10 tones per second to preserve the original timing of the sEMG data (e.g. 10 seconds of sEMG data yielded a 10 second sonification). The Pitch, Loudness, and Attack Time of each tone were mapped to the averaged sEMG data values. Pitch was mapped to a range of 200 – 768 Hz (roughly G3 to G5), such that Pitch increased as sEMG data values increased. Loudness was mapped to a range of 50 – 68 dB(Z), as measured by the SoundMeter app developed for iOS by FaberAcoustical, such that Loudness increased as sEMG data values increased. The Attack Time of each tone was mapped to a range of 0 – 39 ms, such that Attack Time decreased as sEMG data values increased (i.e. the Attack Time of each tone was longest during muscle relaxation and the Attack Time for each tone became progressively shorter as sEMG data values increased).

Three of the six designs tested were spatialized, meaning that Muscle A was played in the left audio channel and Muscle B was played in the right audio channel. The other three designs were not spatialized, meaning that both Muscles A and B were played equally in the left and right audio channels (in the center of the stereo field).

We found that, out of the six designs tested, a Pitch/Loudness with spatialization mapping (PL) resulted in the best listener performance for the task of identifying which muscle contracted first. For the task of identifying which muscle exhibited a higher exertion level, we found that a Pitch/Loudness/Attack Time with spatialization mapping (PLA) resulted in the best listener performance. The non-spatialized mappings showed very poor listener performance.

To understand why these different sonification designs resulted in better listener performance for different tasks, we performed a task analysis for both sEMG data evaluation tasks, and we use these below to offer a possible explanation for the observed performance differences.

3.1. Task Analysis #1 - Identifying Which Muscle Contracts First

Goal: To accomplish this task, the listener must be able to:

1. Understand that the task has started
2. Identify, as quickly and accurately as possible, when each muscle changes from a state of rest to a state of activation
3. Compare the two moments of muscle activation onset
4. Quickly and accurately report which muscle activated first

Sonic characteristics that may facilitate this:

- A *distinct* and *temporally accurate* contrast between the sound of a muscle at rest and the sound of muscle activation onset
- A *distinct* contrast in sound could be facilitated by a change in a number of different sound parameters including pitch, loudness, harmonic content, and timbre
- *Temporal accuracy* requires that the sonification present the change in muscle state (from rest to activation) at the precise moment which it happens

Observation: Out of the six designs, the PL mapping resulted in the best listener performance for the task of identifying which muscle contracted first. This mapping used tones with a very short attack time and a fast decay, as depicted in the amplitude envelope diagram shown below in Figure 1:

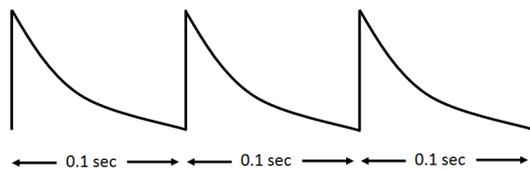


Figure 1 – Shape of the amplitude envelope for three tones in the PL sonification mapping (playing 10 tones per second)

As a result of this amplitude envelope, muscle activation onset was perceived as soon as the tone representing muscle activation onset played. This was not the case in the mappings that incorporated attack time as a parameter (the PLA mapping). For the PLA mapping, the amplitude envelope for each tone at the moment of muscle activation onset looked like that shown below in Figure 2:

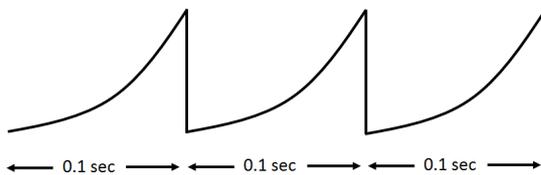


Figure 2 – Amplitude envelope for three tones in the PLA sonification mapping at the point of muscle activation onset

Based on Figure 2, it is clear that, in the PLA mapping, the tone which represented the point of muscle activation onset was not immediately heard by the listener. Thus, while the mappings that included attack time resulted in greater timbral variety (going from a smooth sound while the muscle was at rest to a more percussive sound during muscle activation), this timbral variety may have come at the expense of some temporal accuracy. With the PL mapping being the most temporally accurate out of the six designs, it resulted in the best performance for the task which required temporal accuracy in the sonification (comparing muscle activation times).

3.2. Task Analysis #2 – Identifying Which Muscle Exhibits a Higher Exertion

Goal: To accomplish this task, the listener must be able to:

1. Understand that the task has started
2. Identify when each muscle changes from a state of rest to a state of activation
3. Monitor the level of exertion for each muscle during the muscle activation state
4. Identify when each muscle reverts from a state of activation back to a state of rest
5. Compare the exertion levels for both muscles once the muscle activation state is complete
6. Quickly and accurately report which muscle exhibited a higher exertion level

Sonic characteristics that may facilitate this:

- Easily identifiable sonic differences between the sound of a muscle at rest and the sound of a muscle during contraction
- Easily identifiable sonic differences between various levels of muscle exertion

Observation: Out of the six designs tested, the PLA mapping resulted in the best listener performance for the task of identifying which muscle exhibited a higher exertion level. As discussed previously, including attack time as a parameter of sound in the sonification mapping resulted in greater timbral variety in the sonification (at the expense of some temporal accuracy). With this added timbral variety, the tones in the PLA design sounded smooth and connected while the muscle was at rest, but sounded progressively more percussive and distinct as muscle exertion increased. This not only made it simple to distinguish between the sound of a muscle at rest and the sound of muscle contraction, but also made it easier to distinguish between different levels of muscle exertion, helping to facilitate performance of Step 3 in the above task analysis. The PL mapping (without attack time) did not exhibit this timbral variation between muscle rest and muscle activation, and the listener performance data indicated that this lack impaired the listener's ability to distinguish between various muscle exertion levels. This seems to indicate that variations in timbre, rather than variations in just pitch and loudness, can allow a listener to more easily distinguish between multiple possible data values (in this case multiple different muscle exertion levels).

3.3. Conclusion

This task analysis process demonstrates that there can be tradeoffs for using different parameters of sound in sonification mappings. In this case, the tradeoff was between temporal accuracy and timbral variety, as it was not possible to have both at the same time with the designs used. The PL mapping gave better temporal accuracy than the PLA mapping (and the task analysis showed that temporal accuracy was required for the task of identifying which muscle activated first), but the PL mapping did not have the timbral variety of the PLA mapping. This timbral variety helped the listeners better distinguish between relative muscle exertion levels, which resulted in better performance for the task of identifying which muscle had a higher exertion level. We believe that using task analysis methods in this way as a tool for informing sonification design could help sonification designers identify potential tradeoffs in their designs as well as identify potentially useful and meaningful sonification designs for specific tasks.

4. TASK ANALYSIS IN FUTURE RESEARCH

In order to better understand the role that a task analysis could play in sonification design, we are investigating the space of task-analysis-based sonification design in work that is currently in the planning stages. We will be performing empirical comparisons between task-analysis-based sonification designs and classic parameter-based designs, which are typically task-agnostic. In order to perform an unbiased comparison of these different sonification mappings, it is crucial to ensure that each task-agnostic mapping used faithfully represents the data in the best way that that particular mapping is capable of. It would be easy to create a

poor task-agnostic mapping and compare it to a task-analysis-based mapping and “discover” that a task-analysis-based mapping results in better listener performance (conveniently confirming our hypothesis that task-analysis-based designs will result in better performance). To ensure that this sort of bias does not creep in, we will base the task-agnostic designs on designs found in the literature that have been used by other researchers. The task-agnostic designs will also be designs that pertain to human movement, since this research focuses on sEMG data sonification. We will then compare these task-agnostic mappings with task-analysis-based mappings to see how each mapping affects listener performance.

To perform these comparisons, we will identify characteristics of the sEMG data for our listeners to identify after listening to each sonification and then record the accuracy of their answers. We will then compare the accuracy of the listeners’ responses between sonification designs in order to determine the effects of design on performance and determine whether or not task-analysis-based designs could improve listener performance.

5. LINKS TO SOUND FILES

Pitch/Loudness mapping with spatialization (PL):
<https://soundcloud.com/user-341542684/pitchloudness-mapping-with-spatialization>

Pitch/Loudness/Attack Time mapping with spatialization (PLA):
<https://soundcloud.com/user-341542684/pitchloudnessattack-time-mapping-with-spatialization>

6. REFERENCES

- [1] Mabrouk & Kandil, Surface Multi-Purposes Low Power Wireless Electromyography (EMG) System Design, *Int. J. Comput. Appl.* (2012) 10-16.
- [2] De Luca, The Use of Surface Electromyography in Biomechanics, *J. Appl. Biomech.* 13 (1997) 135-163.
- [3] Kang et al, The Effects of Closed Kinetic Chain Exercise using EMG Biofeedback on PFPS Patients Pain and Muscle Functions, *Int. J. Biosci. Biotech.* (2014) 55-62.
- [4] Steele et al, Electromyography as a Biofeedback Tool for Rehabilitating Swallowing Muscle Function, *Applications of EMG in Clinical and Sports Medicine* (2012) 311-328.
- [5] Croce, The Effects of EMG Biofeedback on Strength Acquisition, *Biofeedback and Self-Regulation* 11.4 (1986) 299-310.
- [6] Giggins et al, *Biofeedback in Rehabilitation*, J. Neuroeng. Rehab. (2013)
- [7] Henkelmann, (2007). *Improving the Aesthetic Quality of Realtime Motion Data Sonification*. (Technical Report CG-2007-4) Bonn, Germany: Universität Bonn.
- [8] Sigrist et al, *Augmented Visual, Auditory, Haptic, and Multimodal Feedback in Motor Learning: a Review*, *Psychon. Bull. Rev.* 20 (2013) 21-53.
- [9] Pauletto & Hunt, *The Sonification of EMG Data*, *Proceedings of the 12th International Conference on Auditory Display* (2006) 152-157.
- [10] Sholz et al, *Moving with Music for Stroke Rehabilitation: a Sonification Feasibility Study*, *Ann. N.Y. Acad. Sci.* 1337 (2015) 69-76.
- [11] Anderson & Sanderson, *Sonification Design for Complex Work Domains: Dimensions and Distractors*, *J. Exp. Psych: Applied* 15.3 (2009) 183-198.
- [12] Baier et al, *Event-Based Sonification of EEG Rhythms in Real Time*, *Clinical Neurophysiology* 118 (2007) 1377-1386.
- [13] Dubus & Bresin, *A Systematic Review of Mapping Strategies for the Sonification of Physical Quantities*, *PLOS ONE* 8.12 (2013) 1-28.
- [14] Dubus, *Evaluation of Four Models for the Sonification of Elite Rowing*, *J. Multimodal User Interfaces* 5 (2012) 143-156.
- [15] Barrass, S. (1997). *Auditory Information Design*. (Doctoral Dissertation)
- [16] Phipps et al, *Human Factors in Anaesthetic Practice: Insights from a Task Analysis*, *British Journal of Anaesthesia* 100.3 (2008) 333-343.
- [17] Van der Veer et al, *GTA: Groupware Task Analysis – Modeling Complexity*, *Acta Psychologica*, 91.3 (1996) 297-322.
- [18] Janice & Dennis, (2003). *Task analysis*. In A. J. Julie & S. Andrew (Eds.), *The human-computer interaction handbook* (pp. 922-940): L. Erlbaum Associates Inc.
- [19] Kirwan & Ainsworth, (1992). *A guide to task analysis*. Philadelphia, PA: Taylor & Francis.
- [20] Berecuartia, (2011). *Compilation of Task Analysis Methods: Practical Approach of Hierarchical Task Analysis Methods, Cognitive Work Analysis and Goals, Operations, Methods and Selection Rules*. (Masters Thesis)
- [21] Embrey, (2000). *Task Analysis Techniques*. Retrieved from <http://www.humanreliability.com/articles/Task%20Analysis%20Techniques.pdf>

Designing Sound Representations for Responsive Environments

Takuya Yamauchi

yamauchi@yinteraction-design.com

ABSTRACT

In this paper, we demonstrate a responsive sound installation consisting of computer-linked thermometers and cameras installed in both interior and exterior locations that detect the states of these spaces based on image and temperature data. The system simultaneously produces and modifies sound pictograms in different spaces in order to convey information on motion or temperature changes. We also propose a sound design method that represents sounds made by physical objects using the above-described installation and sound design method.

1. INTRODUCTION

Herein, we demonstrate a method for creating a sound installation that uses sensors and computers to present computer-generated sound and novel pictogram imagery using information obtained from cameras and thermometers. The computer-generated pictogram sound and image data are presented in a different location from the collection point, and the audience can recognize the changing states by listening to the sound and/or watching the imagery.

The system uses thermometers and cameras that are installed in both interior and exterior spaces to detect the state of those spaces based on the obtained image and temperature data. When the state of the captured images or the temperature changes, such as when a moving object is detected, a signal tone and a pictogram are created and assigned to describe the state based on a temperature and image data-mapping table.

The designed sounds change according to the detected states, and the proposed installation simultaneously produces sounds and pictograms in different spaces in order to convey information related to motion or temperature states. For example, if the temperatures of the interior and exterior spaces are different, different sound pitches are assigned to each space. Thus, individuals walking from outside to inside, or vice versa, will experience a change in sound.

If the installation were demonstrated in both interior and exterior spaces, the temperature in the spaces would also change with the season. For example, if the installation were demonstrated in winter, the inside space would be hot, and outside would be cool, or vice versa in summer.

If the installation is considered from a long-term point of view, individuals can experience sounds of varying pitch, which change with temperature, as the seasons change. Consequently, the assigned sound provides information on both motion concerning a moment and temperature regarding a long period of time.

2. BACKGROUND

2.1. Sound Pictogram

When a subject hears various sample sounds, his or her auditory reaction time will differ for each sound sample. Our proposed auditory system quickly and easily recognizes signal tones, which are characterized by a noticeable sound attack, as compared with environmental sounds. Similarly, sounds that include only a few frequencies are more easily detected than complex sounds that combine several frequencies.

The above-mentioned characteristics are used in various sound design methods in order to provide interaction tones for home appliances, and notification and warning sounds in various environments, as well as navigation information for people with visual impairments. Characteristics such as pitch, tempo, and amplitude can be used to classify whether a sound sample can be easily recognized, and it is important for sound designers to consider these characteristics.

2.2. Sound Design for Interaction

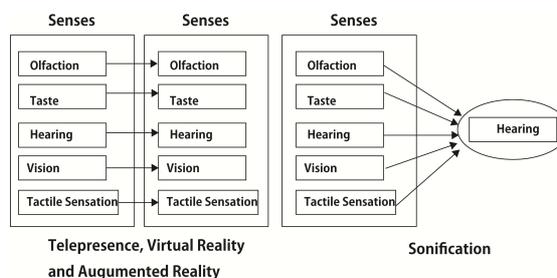


Figure 1 Senses and interaction

Auditory icons are icons that designate sound events on computers, and a number of sound icon design methods, such as SonicFinder, have been proposed for use in mapping sounds to computer events via a graphical user interface. [1] Mapping methods for use between physical objects and sounds have also been presented, and design methods have been applied to a tangible user interface between physical objects and computers. [2][3][4]

An auditory icon assumes that there are several layers that can be used to differentiate sound types. If a sound type is easily discerned, it may be an environmental sound that is generated by some physical phenomenon in the everyday environment. Sounds can evoke concrete images, and we can visualize physical objects based on sounds related to those objects. In



contrast, in the case of simple sounds that consist of only a few frequencies, it may be difficult to imagine where the sounds are generated because of their abstract nature.

Computerized sound engines can generate a number of waveforms in order to produce artificial sounds that do not exist in the everyday environment because of their abstract nature. Both artificial sounds and environmental sounds cover a wide range of frequencies. The abovementioned layers range from concrete to abstract, and auditory icons are based on the concept of layers. Responsive systems analyze signals input from the environment and provide feedback by means of actuators. There are also cases in which feedback is returned as a system.[5] Figure 1 shows the relationship between our senses and demonstrates how the system transforms input into output. In cases involving virtual reality, telepresence, and augmented reality, input data that are detected by sensors are equivalent to output senses.

In other words, the system transforms visual input data to same-sense (visual) output data and auditory data to same-sense (auditory) data. This means that equivalent senses are mapped by the system. On the other hand, there are cases of non-like sense mapping, in which, for example, visual sense data are mapped to auditory sense data. Sonification and auditory display are examples of non-like sense mapping. Such approaches attempt to change visual input data into auditory output data in order to represent the detected data using a sound element, such as pitch or tempo. In the present paper, we focus on auditory representation.

2.3. Sonification timescale

One sound design for navigation systems or home appliances involves sound mapping of an action. In order to represent the target action, a mapped sound corresponding to that action is generated as soon as the action occurs. One such example would be an interaction sound that allows a user to determine whether a button has been correctly pushed based on the interaction sound.

We can begin by assuming that the response system is connected to a camera and that the system enables real-time image analysis. When the camera captures moving objects, the system analyzes their trajectories via image processing and the sum of the vectors of the trajectories is calculated as the state in the image. If we want to express a state via sonification, there are a number of sound mapping methods available.

The interaction of moving objects and sounds must be responsive, and a sound should be generated and expressed as soon as a moving object is detected. This method enables sounds to be recognized effortlessly so that listeners can easily imagine what happened in the space. In other words, the sound is a responsive sign pictogram that expresses the state of the detected space captured by the camera. Since sound feedback from an action is very quick, the response time is very short for the responsive sounds that are used in these methods.

Next, we will consider sonification-related data obtained from the natural environment. Examples of data detected in the natural environment include temperature and humidity over the course a month. The variability of data collected from the natural environment over a short

period of time is lower than that of data collected from interactions between moving objects because such phenomena from the natural environment change very slowly.

Data such as solar wind and cosmic background radiation are measured over long periods in order to reduce data variability. If sonification was designed for these phenomena, sound mappings would be required in order to express changes in phenomena over extended periods of time. Changes in micro-world or space environments would need extended detection periods, and any sonification produced using that data would be difficult to comprehend intuitively. Therefore, it is clear that sound design for sonification should consider both short-period and long-period phenomena.

(1) Sound mapping for short periods of time

Sounds must quickly return feedback in order to report action and motion, and the pitch or tempo in the sound should be assigned to actions and motions.

(2) Sound mapping for a long period of time

Changes associated with the natural environment occur continually and gradually over long periods of time, and sound assignments must report the state of the phenomena. Thus, a new sound method is required.

We investigated the use of thermometers in sound mapping for both the interactions of moving objects and the natural environment, and decided to focus on a design method for interior and exterior spaces.

2.4. Temperature sonification

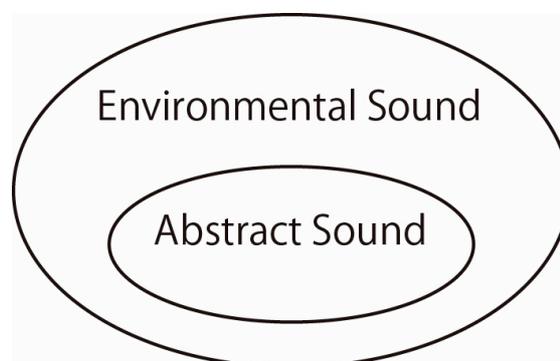


Figure 2 Conceptual diagram of environmental sound and abstract sound

In the following sub-section, we describe our attempt to implement a sound installation system that captures temperature and image data. One purpose of the installation is to create sound pictograms based on images taken by a camera and detecting temperature via a thermometer, from which the system generates a sound that integrates the image and temperature data. In order to present the state and temperature via sound, a new method of designing sound pictograms is required.

The results of psychological experiments using the semantic differential (SD) method indicate that the timbre of the sound of hitting metal provide a cold impression, whereas low tempo and pitch provide a hot impression. High pitch and fast tempo also provide a cold impression.[7][8] We designed the temperature-related sound based on these findings.

Next, we considered how to represent sounds related to sensor data and how sound should be used to present thermometer temperature data. In our installation, a sound pictogram and environmental sounds generated by granular synthesis are considered. Figure 2 shows a conceptual diagram of environmental and abstract sounds. When creating sound pictograms, we assumed that all sounds were composed of physical pitch, rhythm, and amplitude. The abstract sound in Fig. 2 consists of artificial sounds, such as pure tones.

Environmental sounds are composed of numerous abstract sounds. As in the case of environmental sounds, increasing the number of abstract sounds in an artificial design increases the complexity of the resulting sound. Much like an environmental sound, concreteness depends on the number of abstract sounds that are added. We designed environmental sounds and sign pictograms based on the conceptual diagram.

2.5. Environmental and Signal Tone

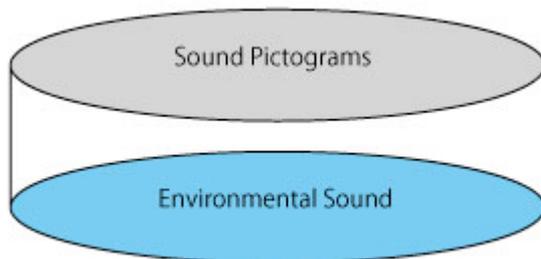


Figure 3 Sound design concept

Our sound design model consists of sound pictograms and environmental sound. Figure 3 shows the concepts behind this sound installation model. As can be seen in the figure, the concept involves the use of abstract sound in the form of sound pictograms and concrete sound as environmental sound.

As described in the previous section, both the abstract sounds used as the sound pictogram and environmental sounds have psychological and physical characteristics that compliment each other. We designed our sound installation based on category of the concept, and the sound pictogram and the environmental sound are used for each purpose.

One of the reasons environmental sound is used is to provide background sounds. In our model, environmental sound was generated by granular synthesis based on 10 sound types. Another purpose of the sound pictogram is to provide notifications and intentions. The sound pictograms represent temperature and provide information on states outside and inside a room, thereby informing listeners in the room by means of the designed sound.

2.6. Background Sound

Figure 4 shows a waveform chart and sound spectrographs based on 10 environmental sound samples

that were generated via granular synthesis. The granular synthesis technique is used for combining sound grains, which are defined as pieces of sound in the sound samples, via numerical calculation. The resulting combined sounds can be flexibly changed by modifying the frequencies and amplitude of the sound grains. The spectrographs include periodical waveforms that are composed of numerous frequencies with flat peak levels. These 10 sound samples were used for synthesis by Max/MSP, which is a visual programming language.

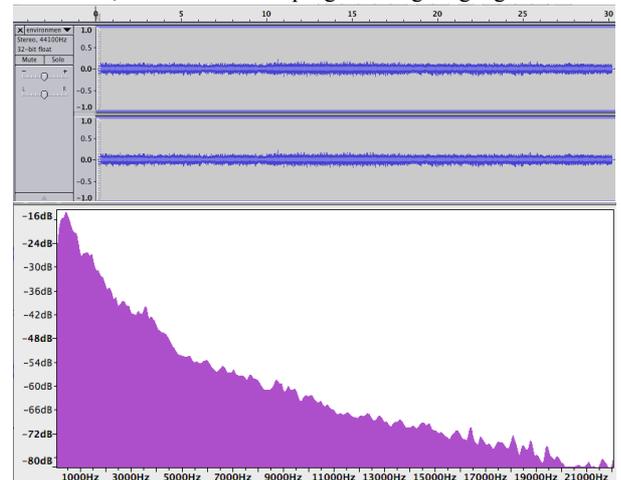


Figure 4 Waveform and spectrograph of environmental sound

2.7. Sound Pictogram

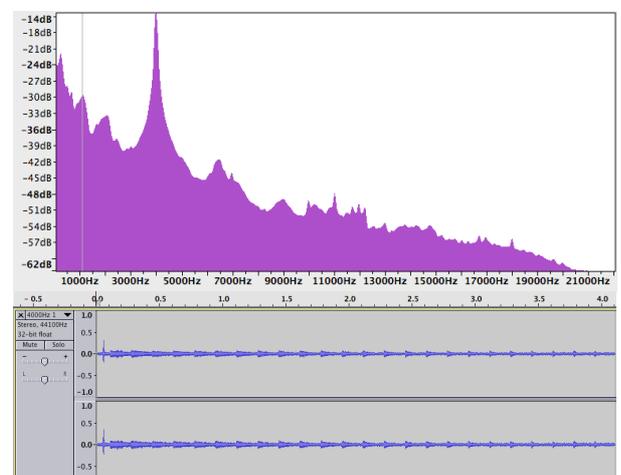


Figure 5 Waveform and spectrograph of a sound pictogram

The timbre, volume, and frequency of sounds are indicated via sine waves or noise in the sound pictograms of artificial sounds in order to indicate changes in state or motion.[6] For example, timbre changes according to temperature. If the timbre includes metallic factor in SD scale collections produced by factor analysis, then the sound gives listeners the impression of cold. High-frequency sound also gives listeners the impression of cold, whereas low-frequency sound gives listeners the impression of low activity.

In our sound installation, SuperCollider was used to create a signal tone generated by a ringing filter in

order to adjust the sound timbre. Figure 5 shows the waveform and spectrum of the signal tone. The signal tone is a periodic sound, and includes only a few frequencies. The spectrum has a frequency peak of 4,000 Hz and also has only a few frequencies. The following items were considered in designing the signal tone:

(1) *To generate an easily heard pure sound to report the state that occurs.*

The signal tone, which is categorized as an abstract sound, is a periodical waveform that contains just a few frequencies. State changes such as detected motion are reported by the signal tone. We next considered how sound should be designed in order to indicate the reported state.

(2) *To report temperature by the signal tone.*

In our demonstration, assignment of sound for a continuously changing parameter such as temperature was considered. The temperature in a room changes continuously over long periods of time but does not change drastically over short periods of time. We also considered how assignment should be made.

The proposed system assigned a sound depending on the temperature as measured by the thermometer, and sounds were generated when moving objects were detected. The table shows the mapping relationship between temperature and pitch of the generated sound. In the case of a low temperature in the detection space, moving objects in the space are indicated by low-pitch sounds, whereas in the case of a high temperature, movements are indicated by high-pitch sounds.

Temperature	Pitch
Low	Low
High	High

Table 1 Mapping relationship between temperature and pitch

3. CONCEPT

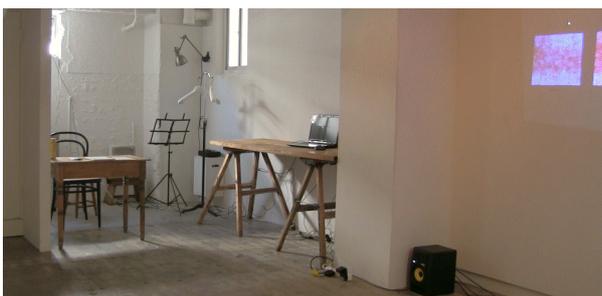


Figure 6 Sound installation

Next, we demonstrated an interactive artwork using sensors and sounds. Figure 6 shows the space used for the interactive artwork. The purpose of the interactive artwork is to represent various data types, such as temperature or motion, using sound and images. In particular, the present study focuses on the simple representation of a large amount of sensor data.

It is expected that individuals will be able to recognize not only states of motion or noise based on auditory and visual pictograms, but also continuous changes in those states, such as multiple individuals

simultaneously walking from an interior to an exterior space. When numerous people walk through the space, the system also detects their trajectories. A signal tone for the people is then reported in the room, and we can recognize the new state from the resulting noise.

Since representation sounds change according to temperature and motion, seasonal changes can be recognized based on pictogram sound in the long-term point of view. Here, the signal tone is reported when the system detects state changes, such as motion, via the camera, and the pitch of the signal tone also changes according to temperature changes. Changes to the signal tone created by temperature changes can be easily recognized if the system is active for long periods of time.

A responsive system for an interactive artwork presents not only the motion or action interactions, but also long-period changes, such as seasonal changes.

4. Implementation

4.1. Visual Pictogram



Figure 7 Visual pictogram in the sound installation

The interactive artwork utilizes simple icons indicating temperature and motion. One purpose of the display is to indicate complex state and temperature information via simple visual and auditory icons. Another purpose is to represent these icons as real, existing objects, such as furniture or shadows. The icon color provides an indication of temperature. High temperatures are shown in red, and low temperatures are shown in blue (Fig. 7).

4.2. System

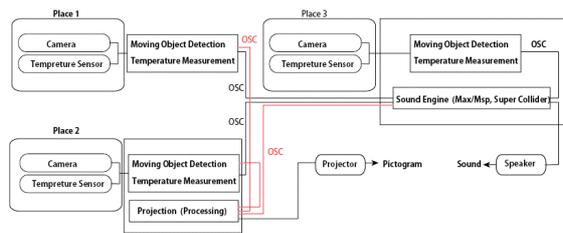


Figure 8 System Architecture

The proposed sound installation system is composed of three cameras and three thermometers connected using network protocols. The OSC protocol sends data to Max/MSP and a number of projectors in real time. Figure 8 shows the system and network used in our installation. After captured image and temperature data are sent to these engines, pictograms are projected as information on the states that exist in these spaces.

4.3. Camera Detection

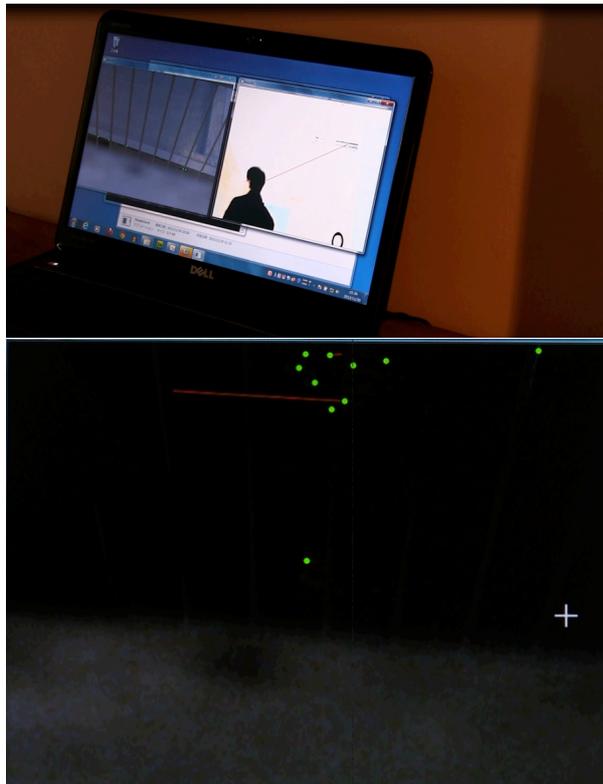


Figure 9 Camera detection system

Three cameras for detecting moving objects were included in our installation (Fig. 9), and the moving objects were analyzed by an optical flow algorithm of open source computer vision (OpenCV). The algorithm used for calculating spatial vectors of luminance estimates the flow related to the vectors in order to detect moving objects.

The system analyzed the states of interior and exterior spaces based on captured images, regardless of whether the images were captured in an active space. For example, for a case in which there are moving objects in

a cold outside space, the system first detects the moving objects and the temperature information. Next, computer-generated images and sounds are created based on the temperature information and the audience can hear temperature sounds based on the activation in the cold space.

4.4. Temperature Detection



Figure 10 Thermometer in the sound installation

A thermometer with an embedded microcontroller is used to generate the sounds and operate the projectors (Fig. 10). Three thermometers were positioned in interior and exterior locations, and the temperature levels were observed in real time. Observed data were then analyzed and sent to visual server via processing and sound engine. The pictogram that results from the processing reflects ambient light based on the temperature in the computer graphics, and thus represents the temperature state.

5. Conclusion

Our proposed system, which can measure temperature in interior and exterior spaces and can detect moving objects by camera, presents the state of these spaces by mapping sounds and pictograms. The pitch of a generated sound changes according to temperature and the activity of the moving object. Based on the detected image, if the temperature of a space is determined to be high, a high-pitch sound is generated. Individuals can then simultaneously recognize activities occurring in both interior and exterior spaces based on generated sounds.

Sound pictograms were designed by combining environmental and abstract sounds, which each have a role in representing the state in the spaces and ambient sounds. If individuals can feel or recognize information about the location states and temperature levels by hearing the mixed sound, they can also understand auditory, visual, and tactile sensations by simultaneously hearing those sounds. Sonification is an augmented reality that is realized by sound design.

The method used by the designer to map the sounds is important, and sensor data must be assigned to valid sounds. These assignments depend on the designer's receptivity, and it is important to remember that the impressions of individuals will also change based on the sound design. For instance, a heartbeat can be accurately mapped to a pulse sound in electrocardiography, which then provides information on the cardiac state by means of an electrocardiogram and auditory information. The designed pulse sound can provide accurate information for predicting imminent cardiac events. When designing the sound to be used, it is important to map the sound to the actual object.

6. ACKNOWLEDGMENT

We would like to thank the members of Cyber Sound Project for their helpful assistance.

7. REFERENCES

[1]"The SonicFinder: an interface that uses auditory icons", William W. Gaver

,Journal Human-Computer Interaction archive Volume 4 Issue 1, March 1989 Pages 67-94

[2] Ref Till Bovermann, Rene Tünnermann, Thomas Hermann (2010). Auditory Augmentation. Int. J. on Ambient Computing and Intelligence (IJACI), 2 (2) , p. 27 – 41

[3] "Sound Jewelry" Takuya Yamauchi and Toru Iwatake Leonardo Music Journal No 18 MIT Press

[4] "the transfinite" Ryoji Ikeda, audiovisual installation 2011

[5] Takuya Yamauchi and Toru Iwatake "Design of a process for interactive product in ubiquitous space " International Journal on Interactive Design and Manufacturing (IJIDeM) Volume 2, Number 2 / 2008,5 Springer Verlag

[6] O.Kitamura, S. Namba and A.Matsumoto "Factor analytical study of tone color",Proc. 6th ICA,A-5-II"

[7] Grey, J.M. "Multidimensional Perceptual Scaling of Musical Timbres", Journal of the Acoustical Society of America, 61(5): 1270-1277

[8]Wessel, D. L., "Timbre Space as a Musical Control Structure", Computer Music Journal, 3(2): 45-52

HYPOTHESIS-DRIVEN SONIFICATION OF PROTEOMIC DATA DISTRIBUTIONS INDICATING NEURODEGRADATION IN AMYOTROPHIC LATERAL SCLEROSIS

William L. Martens

Faculty of Architecture, Design and Planning
University of Sydney, Sydney NSW 2006 Australia
william.martens@sydney.edu.au

Philip Poronnik

School of Medical Sciences
University of Sydney, Sydney NSW 2006 Australia
philip.poronnik@sydney.edu.au

Darren Saunders

School of Medical Sciences
University of New South Wales, Sydney NSW 2052
Australia
d.saunders@unsw.edu.au

ABSTRACT

Three alternative sonifications of proteomic data distributions were compared as a means to indicate the neuropathology associated with Amyotrophic Lateral Sclerosis (ALS) via auditory display (through exploration of the differentiation of induced pluripotent stem cell derived neurons). Pure visual displays of proteomic data often result in "visual overload" such that detailed or subtle data important to describe ALS neurodegradation may be glossed over, and so three competing approaches to the sonification of proteomic data were designed to capitalize upon human auditory capacities that complement the visual capacities engaged by more conventional graphic representations. The auditory displays resulting from hypothesis-driven design of three alternative sonifications were evaluated by naïve listeners, who were instructed to listen for differences between the sonifications produce from proteomic data associated with three different types of cells. One of the sonifications was based upon the hypothesis that auditory sensitivity to regularities and irregularities in spatio-temporal patterns in the data could be heard through spatial distribution of sonification components. The design of a second sonification was based upon the hypothesis that variation in timbral components might create a distinguishable sound for each of three types of cells. A third sonification was based upon the hypothesis that redundant variation in both spatial and timbral components would be even more powerful as a means for identifying spatio-temporal patterns in the dynamic, multidimensional data generated in current proteomic studies of ALS.

1. INTRODUCTION

This study investigated three alternative approaches to the sonification of proteomic data distributions as a means to indicate the neuropathology associated with ALS. A local group of researchers routinely generate large complex proteomic datasets obtained from patient-derived cell lines

and animal models in efforts to understand the changes in the ubiquitin-proteasome system during the progression of ALS. It is common to attempt to interpret these data with the aid of visual displays, using graphics such as that shown in Figure 1. However, these visual displays often provide an unwieldy summary of the structure of complex proteomic datasets, and so it was of great interest to determine if sonifications could provide an additional useful approach to the exploratory analysis required for this data, both as an accompaniment to visual display (as suggested in [1]), and as an independent means by which a stand-alone auditory display might become regarded as potentially useful in its own right (this is not a new idea, such proposals appearing in the early 1980's [2]).

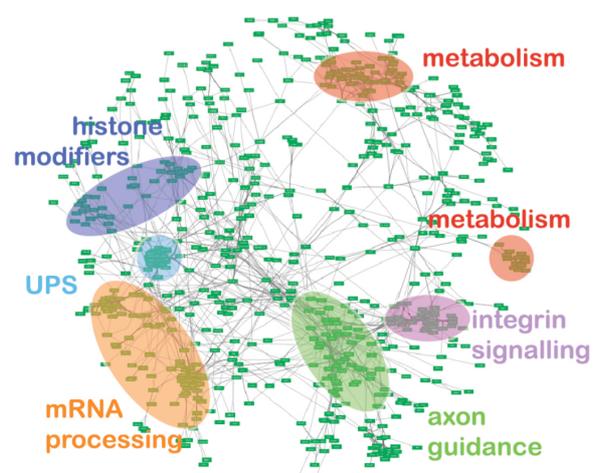


Figure 1: An example of a visual display of proteomic data of the sort that has been utilized to aid in understanding the changes in the ubiquitin-proteasome exhibited in studies of proteomic ALS neurodegradation.

The over-arching assumption here is that auditory enhancement of visually represented data can significantly increase the ability of researchers to detect subtle changes or anomalies in such numerical data sets. Although the motivation for this research was to develop potentially useful auditory display tools for practical applications in the medical sciences in general, this paper reports only preliminary results of a single case study of proteomic data associated with neuropathology in ALS patients. A primary goal of this particular study was to determine which of three sonifications would be judged by domain experts to be most successful in identifying differences in spatio-temporal patterns within the multidimensional proteomic data generated in a single exploratory study. In addition to polling the opinions of domain experts, however, this preliminary study assessed, using in two psychometric tasks, the success of three sonifications in aiding naïve listeners to identify differences between proteomic data distribution of three types of cells. The first task required naïve listeners to make dissimilarity judgments for all pairwise comparison of the set of nine cases defined by the factorial combination of three cell types and three sonification methods. The second task required those same naïve listeners to make ratings of each of the nine stimuli on a number of subjective attributes that might be related to the perceptual dimensions underlying their dissimilarity judgments.

Although these results reveal only the perceptual distinctiveness of the sonifications, they provide a basis for further exploration of the potential value of these sonifications, with the benefit that their perceptual distinctiveness has been established. Clearly, further work will be required to address important issues in sonification system usability for this application, and to determine whether the system will provide a real benefit to domain experts. Indeed, this paper reports preliminary evaluation results as an indication of progress on just one component of the larger medical research project in that it focuses only upon the effectiveness of the display technology used to aid medical scientists in interpreting and understanding their medical data (particularly proteomic data). Nonetheless, it is reasonable to assume that such preliminary examination of the perceptual distinctiveness of sonification system outputs will be a valuable first step in developing and exploring novel sonifications.

2. METHODS

Three alternative proteomic data sonifications were developed and compared in terms of their potential to communicate to the user changes due to neuropathology associated with ALS. Without any prior experience with auditory display of such data, an approach was adopted that is here termed ‘hypothesis-driven’ design of these sonifications, with the expectation that the relative value of the sonifications could be evaluated by the domain experts, who would try to detect differences between the results for ALS cases versus control cases. Of course, ground truth was already available for these particular data, since the ALS cases were selected on the basis of established medical diagnostic procedures. Therefore, relative performance under blind testing conditions has been used to determine which sonifications were best able to communicate to the user regarding neuropathology associated with ALS via changes in sonifications that were directly driven by differences in distribution of the proteomic datasets.

One hypothesis to be tested was that variation in predominantly timbral attributes would be most effective in revealing differences in proteomic data distributions. An alternative hypothesis was that variation in spatial timbral attributes would be more effective in creating audible differences between the sonifications produced for each of the three cell types. Finally, a third hypothesis was that including redundant variation in both timbral and spatial attributes would be more effective than just one or the other of these two individual approaches in isolation. Preliminary results suggested that comparison of these three sonifications, based as they were upon this ‘hypothesis-driven’ design, should allow for the rejection of a hypothesis that had resulted in less effective sonifications, leaving for future consideration only those hypotheses that were not rejected through blind testing. This scientific approach to the initial evaluation of competing sonifications will be examined in more detail in subsequent sections of this paper; however, before describing further this study’s experimental design, the auditory display technology underlying the alternative sonifications will be presented.

2.1. Sound synthesis for the sonifications

In order to generate a sonification for the available proteomic data of interest, a strategy for synthesis that took into account the complexity of the large multivariate dataset was formulated based upon parameter mapping [3]. For nine distinct cases, an assembly of short-duration, temporally overlapping ‘grains’ of sound were created, the parameters of which were selected to approach approximately the minimum perceivable event time for distinct percepts of duration, frequency and amplitude (i.e., approaching auditory resolution of human observers in discriminating between identifiable attributes of loudness, pitch, and those component auditory attributes that are generally regarded as belonging to one of two collections termed timbral or spatial attributes). The ‘hypothesis-driven’ design approach taken here required sound synthesis technology that could offer independent variation of many synthesis parameters to provide identifiable variation in distinct auditory attributes. In the initial stage of this work, synthesis based upon a simple physical model [4] was tested for its versatility in producing a wide range of short sounds exhibiting audibly identifiable timbral variations that all had potential for evoking physical referents in the minds of the listeners. In the next section, the spatial positioning of grains is explained.

The synthesis technology that ultimately was adopted for this project resembles granular synthesis (see [5]), in that a multitude of short sound sources formed an ensemble output (likened to a ‘swarm’) rather than forming clearly separable events that might be heard as distinct in time and space. In all sonifications designed for the current work in this way, there was always a hypothesis to be tested regarding which parameters of the data were ‘mapped’ to particular synthesis parameters. It is beyond the scope of this paper to present the details of the synthesis technology that was developed and refined through experimentation with the available multidimensional proteomic data. Suffice it to say that swarms of percussive ‘grains’ (again, see [5]), synthesized with ‘parameter-mapped’ control over multiple timbral attributes, were distributed in time and space according to the distribution of proteomic data that featured 1815 variables observed over the nine cases to be examined.

2.2. Spatial sound processing for headphone display

Although discrimination of frontward from rearward incidence of sonification components could be well supported if binaural processing were to be coupled with head-tracking technology (see [6]), the experimental stimuli generated in the current investigation were not modified by active sensing of the listener's head turning. Without such tracking of head movements, the sonification designer should not expect the listener to be able to clearly identify whether a source presented at a given lateral angle is being presented with frontward or rearward incidence. Due to the difficulty in supporting reliable front/rear distinctions using uncoupled binaural processing for headphone-based spatial auditory display (again, see [6]), only a simplified model of head acoustics was employed here to move sonification components along the listener's interaural axis. The acoustical cues that were simulated in order to accomplish this manipulation of sound source incidence angle included the interaural time delay (ITD) and the head shadow that generally grows larger at the listener's contralateral ear as the incidence angle of the source is offset laterally from the listener's median plane. This approach offered an advantage over a single-user headphone display in that several listeners could use the system simultaneously without the unexpected variation that would occur if the spatial processing were coupled with head movement of just one of multiple listeners. Of course, using head-coupled updating of headphone-based binaural rendering technology could be added for single-user exploration of the spatial configuration of sonification components (including sensitivity to the listener's translation movements as well as changes in head orientation); however, for the initial studies reported here, only non-head-tracking headphone technology was employed.

2.3. Spatial versus timbral emphasis in sonification

There were nine sonifications created from the factorial combination of three synthesis solutions¹ applied to data from three cell types. So for each of three types of cells that should produce an identifiably different sound, each of three unique parameter-mapping synthesis solutions were applied for presentation. The first of these synthesis solutions was termed the 'Timbral-only' approach, which put emphasis upon timbral differences resulting from spectral variation between grains. The second approach was termed the 'Spatial-only' approach, which held grain spectra constant, and only distributed the grains spatially along the listener's interaural axis. The third approach was termed the 'Spatial-Timbral' approach, and combined redundant variation in the output sound based upon the simultaneous application of both of these parameter-mapping approaches. These sonifications were chosen as candidates for best allowing the differences between cells to be appreciated by any observer, not just those with domain knowledge.

¹ The synthesis solutions employed here were all programmed within the Matlab™ environment. Although the details of the synthesis approaches taken would no doubt be of interest to a subset of readers, those details are considered to be beyond the scope of this paper. The code itself provides the most enabling description of the synthesis approaches. In order to enable interested researchers in replicating the approach taken in the project described in this paper, the employed Matlab code will be provided online (please send an email request to the first author for the URL).

2.4. Experimental Tasks

While discrimination between sonified cases was examined in pilot tests that were run informally during development of competing sonifications, the formal study that allowed more comprehensive analysis of similarities and differences between sonifications comprised two tasks. The first task was a pairwise dissimilarity-rating task, in which the global differences between nine sonification outputs were examined, without respect to particular identifiable attributes. The second task was an attribute-rating task, in which the particular character of each of the nine sonification outputs were examined with respect to identifiable attributes that were exemplified by anchoring stimuli found to be positioned at the extremes of each continuum for those attributes that seemed most distinctly varying within the set of nine stimuli. In fact, for the initial exploration of the characteristics of the nine stimuli, only the sonification developer engaged in the selection of adjectives describing the stimuli through informal discrimination tasks, and so no profiling of the stimuli was done by the five listeners who were naïve regarding the purpose of the experiment. The two formal tasks were completed by these five naïve listeners, but the adjectives used to describe the attribute rating scales were only introduced after the completion of the pairwise dissimilarity-rating task, in order to avoid drawing attention to the experimenter-identified attributes. The instructions for the first task indicated to the listeners that global dissimilarity ratings were required, rather than differences between sonifications based upon particular auditory attributes).

All pairs of nine sonifications were presented to five listeners for their evaluation via Sennheiser HD600 headphones at a comfortable listening level (approximately 75 dBA). Each listener completed one block of 72 trials, which is the number of paired comparisons resulting from the exclusion of the diagonal entries of the 9 x 9 matrix of dissimilarities (i.e., excluding all comparisons between identical stimuli). The sonifications in each pair were presented twice, in two separate trials, with order of presentation reversed for the second presentation, and always separated by a 1-s delay. For each pair of sonifications, listeners recorded their inter-stimulus dissimilarity ratings using a horizontal slider incorporated into an onscreen Graphical User Interface (GUI). On-screen instructions prompted listeners to indicate how similar they thought the sonifications sounded, with the leftmost response indicating that the sonifications sounded most similar, and the rightmost response indicating that the sonifications sounded maximally dissimilar. Each listener had to develop his or her own criterion for the anchoring point of maximal dissimilarity during an initial practice run in which 12 representative pairwise comparison trials were completed. After the initial practice run of 12 trials, each listener completed the formal run of 72 trials. The dissimilarity data matrices produced by each listener in these 72-trial runs could have been averaged to produce a single dissimilarity data matrix for group analysis, however a more powerful analysis using Individual Differences SCALing (INDSCAL) was employed to examine how the five listeners differed from each other, in addition to the summary that is available via examination of the group result.

The combined collected dissimilarity ratings from the group of five listeners were submitted to INDSCAL to obtain two useful outputs: First INDSCAL produced a two-dimensional (2D) spatial configuration of cases (a group 'Stimulus Space'

derived for five listeners taken together) in which each sonification was given coordinates along two dimensions so that the Euclidean distances between the points corresponded to the dissimilarity ratings. The INDSCAL analysis also produced estimates of the differences in weighting that each of the five listeners placed on the resulting dimensions (which weightings are captured by INDSCAL in terms of a ‘Subject Space’). Further details of the analysis are given in the next section of this paper (see the book on *Modern multidimensional scaling* by Borg and Groenen [7] for a more complete explanation of INDSCAL analysis).

The instructions for the second task indicated to the listeners that attribute ratings were required, based upon than differences between sonifications that could be identified with particular auditory attributes. These attribute scales were anchored by adjectives that had been selected by the experimenter to represent the most distinct differences within the set of nine stimuli that seemed likely to be understood by the naïve listeners without much explanation. The selected anchors included the following pairs of adjectives:

- Sparse ↔ Dense
- Tense ↔ Relaxed
- Smooth ↔ Rough
- Compact ↔ Scattered
- Simple ↔ Complex

3. RESULTS

The results of the INDSCAL analysis of the obtained data are shown in Figures 2 and 3. The Stimulus Space shown in Figure 2 uses plotting symbols that indicate the type of cell ('Control', 'ALS', and 'Fibroblast') for which each sonification was generated, as indicated in the legend located in the upper right corner of the graph. Line segments connect the plotting symbols in order to group together the results for the three cell types that were associated with each type of sonification that was employed for the group. The interpretation of this graph may not be obvious at first glance, but it is actually quite straightforward: The three groups of connected symbols will be plotted close to one another if the perceived differences between them is relatively small. For example, the smallest cluster of symbols that are grouped near the origin of the graph (i.e., the [0,0] point) are associated with ‘Spatial-only’ sonifications that were heard to be more similar to each other than those associated with the other two groups of sonifications. The value of this plot is that the relative distance between plotting symbols can be interpreted as providing a uniform indication of both within-group differences and between-group differences (the term ‘uniform’ is used here to indicate that all distances here are based upon a common Euclidean scale). Yet it remains to be asked, what can be concluded from such results. The primary conclusion would be that the sonification type used for the group of cells associated with ‘Spatial-Timbral’ sonifications are showing the greatest inter-stimulus distance of all three groups, and therefore this sonification solution would be preferred according to the criterion that these cells should produce sonifications that are perceptually different as possible.

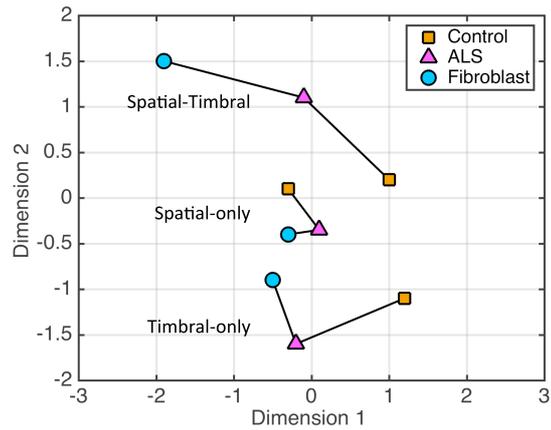


Figure 2: Stimulus Space resulting from the INDSCAL analysis of nine sonifications. Line segments connect the three sets of points associated with each sonification approach, and symbol shapes indicate the cell type being sonified (as identified in the inset legend).

As mentioned in the methods section of this paper, INDSCAL analysis also produces estimates of the differences in weighting that each listener places on the resulting dimensions. This INDSCAL-derived ‘Subject Space’ is shown in Figure 3 for the group of five listeners who participated in this exploratory study. Although the obtained dissimilarity-rating data requires different weights on the two dimensions of the group ‘Stimulus Space’ for each listener, the advantages of INDSCAL is that these differences make it possible to separate such individual differences from the group solution, which shows the common underlying configuration that fits best to all the data. Note that three of the listeners put roughly equal weights upon the two ‘Stimulus Space’ dimensions, indicated by vectors drawn at around 45° from the origin of the graph in Figure 3. One listener put slightly more weight upon Dimension 2, while the remaining listener put more weight upon Dimension 1. Nonetheless, the results are consistent with the hypothesis that for the nine sonifications presented, the five listeners share a common underlying perceptual space that admits of two salient dimensions (although conjecture about the existence of a third underlying dimension might be tempting to consider, comparisons between just nine stimuli do not provide an adequate basis for supported such a conclusion, as explained in the Borg and Groenen [7] book).

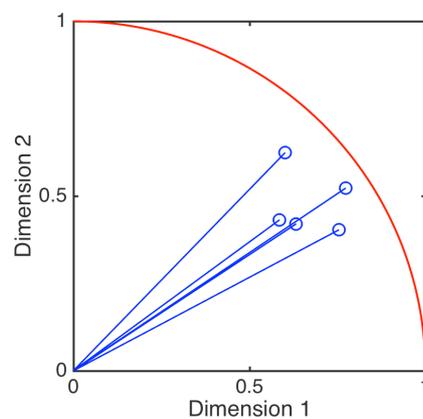


Figure 3: Subject Space resulting from the INDSCAL analysis of nine sonifications for five listeners.

Having concluded that the nine sonifications can be configured within a common perceptual space for the five listeners, it remains to be determined how those two salient dimensions might best be interpreted. One answer to this question would be to find out whether INDSCAL-derived Stimulus Space coordinates could be related to ratings of those stimuli along identifiable auditory attribute scales anchored by the adjective pairs selected by the experimenter. Therefore, the attribute ratings made by those same five listeners, when presented with those nine stimuli individually on separate trials, were submitted as competing predictor variables in a stepwise regression analysis. Although the correlations between sets of ratings could be fairly high, such as that between the smooth-rough and the tense-relaxed rating data, the stepwise regression analysis showed a single set of ratings as the best predictor for INDSCAL Dimension 2, and that was the set associated with the compact-scattered anchoring adjectives (with $R^2=0.66$). After excluding the compact-scattered set from consideration for interpreting INDSCAL Dimension 1, no one of the four remaining predictors showed a particularly high correlation with the coordinates of the nine stimuli on Dimension 1. However, when the smooth-rough and the tense-relaxed rating data were combined to form a new composite predictor, that new predictor accounted for a more of the variance in Dimension 1 coordinates (with $R^2=0.45$).

Taken together, the results of the two tasks serve to show how big the differences were between nine sonifications, and also suggest how one might describe the nature of those differences. The differences between the outputs of the three sonification techniques were best described as varying along a compact-scattered dimension (in the vertical direction of the graph). Clear differences also existed in the configurations derived for the sonifications of the three cell types, which differences were associated primarily with variation in both the smooth-rough and the tense-relaxed ratings.

4. DISCUSSION AND CONCLUSION

The results obtained in this exploratory study only scratch the surface of the problems that must be addressed in developing and evaluating sonifications in this domain of complex, multidimensional proteomic data generated by research studies in medical science. Of the three types of sonifications presented, it seems that sonifications mapping from data to both timbral and spatial parameters provide more distinguishable results than mapping to either timbral or spatial alone, although these results must be regarded as quite preliminary. Nonetheless, the results seem quite promising when compared to the results of typical attempts to visualize such data. One such attempt utilizes multivariate analysis to reduce the complexity of the data to a more easily digestible form. Of course, similar data reduction procedures can be used as a pre-processing step for sonification as well (see [8]). What most such analyses typically attempt to do is to capitalize upon redundancy in the data to find a lower-dimensional perspective on the patterns of underlying variation.

The fact that the 1815 variables are somewhat correlated with each other means that a good deal of the variance in the data is shared, and that shared variance might be represented by a projection of the cases onto a single axis or two through the 1815-dimensional space defined by the proteomic

variables. The most common multivariate analytic technique that seeks out such a projection is Principal Component Analysis (aka PCA). PCA effectively rotates the axes in a multivariate space to find the principal axis along which the variance in the dataset is maximized, taking advantage of the covariance between all the variables. The analysis also finds a second axis, orthogonal to the first, that accounts for the greatest proportion of the remaining variance (see [9]). Figure 4 shows the scores on the principal components resulting subspace projection for the nine cases that were examined in the current study. While the simplicity of the graph in Figure 4 suggests that a simple difference might exist between the three groups of three items here, there is no way of learning from the graph what the meaning of the underlying components might be. Nonetheless, the PCA does provide a potentially more satisfying look at what is going on in the data, even though this involves a somewhat unwieldy graphical analysis of the weights involved in constructing the linear combinations on which the scores shown in Figure 4 are based.

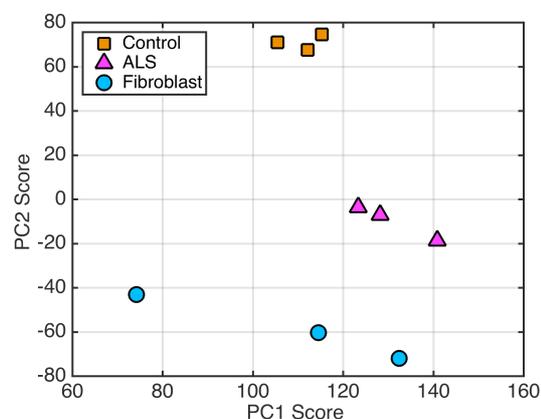


Figure 4: Principal Component (PC) Scores resulting from the multivariate analysis of the proteomic data that featured 1815 variables observed over nine cases.

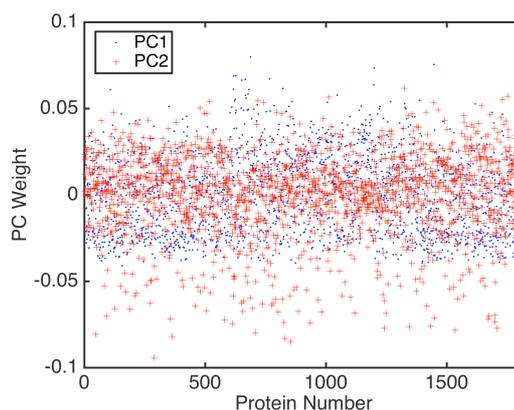


Figure 5: Weights placed upon the 1815 variables that resulted from the Principal Component Analysis (PCA) of the proteomic data.

The weights that were placed upon each of the 1815 variables are illustrated in Figure 5. It is difficult to imagine seeing a pattern here, but it is not so difficult to imagine hearing a change in the underlying pattern of sound ‘grains’ that might be generated through the spatiotemporal

distribution of those ‘grains’ in an appropriately constructed sonification of these data. It is precisely the expectation that such patterns might be appreciated by ear more readily than by eye that motivated the current work.

It is instructive to compare the current results with those of other developers of sonification systems that used similar granular synthesis techniques. An example of early work using granular synthesis, and that also was notable in that it shared a similar psychometric testing perspective, was that of Smith, et al [10]. Their approach was to use a three-alternative forced choice (3AFC) task to track a user’s ability to discriminate between ‘clouds’ of granules varying in frequency with controlled mean and standard deviation within two clouds being compared. Thresholds for hearing a comparison cloud to be ‘higher’ than a reference cloud were measured in the presence of distracting frequency-dependent amplitude modulation resulting from granules closely adjacent in frequency. In presenting their seminal work, Smith, et al [10] identified the types of perceptual discrimination that might be studied as either detection, recognition, or discrimination. The current work extends their psychometric testing approach to include global-dissimilarity-based perceptual scaling with attribute scaling to aid in the interpretation of underlying dimensions. The current results offer the advantage of determining which of a number of competing sonifications created for a small group of listeners the greatest overall perceptual differences between cases for which discrimination is desired (i.e., between cell types). Furthermore, the magnitude of the perceptual differences so observed could be compared to the magnitude of the perceptual differences existing between the multiple outputs of the competing sonification techniques, since these both types of differences were scaled in a common multidimensional space.

In a 1999 review paper, Barrass and Kramer [11] have provided a comprehensive survey of approaches for designing sonifications, and also have outlined ongoing concerns with the existing sonification practice. Of particular relevance is their discussion regarding how knowledge about auditory perception can allow sonification designers to predict how listeners will be able to perceive (if not understand and interpret) variations in novel sonifications. As in the current study, such knowledge can be derived for competing sets of sonifications, however, the point made by Barrass and Kramer [11] is well taken that the theoretical evaluation of new, untried designs requires more than psychoacoustic data. This is because psychoacoustic theories do not involve issues of representation that are central in sonification, since listeners need to hear the underlying data relations in the sounds, rather than just the auditory attributes that are modulated by them.

Thus, the current work in evaluating ALS-related proteomic data sonifications must be regarded as work that is still in the early exploratory stage. As the work enters into a second more confirmatory stage, it will become critically important to establish the means whereby progress and success can be ascertained. Therefore, in a manner that was thoroughly discussed in Bonebright and Flowers’ [12] chapter in the *Sonification Handbook* on ‘Evaluation of Auditory Display’ the initial (and ongoing) evaluation of the current sonifications has been focused upon whether the auditory distinctions displayed in each of the three case studies are in fact as audible and intelligible as the sonification system

developer has designed them to be. The evaluation methods that were used in this regard are those borrowed from perceptual science, and include psychophysical testing for detection, discrimination, and identification of displayed auditory attributes. In addition, the perceptual distinctions displayed in the sonifications presented in the current study, which were intended to distinguish differences that exist in the proteomic data, were assumed to grow larger in perceptual magnitude as differences in the data grew larger. This assumption was directly tested in the experiments reported here, in that known differences in the data (according to the medical science) were used to predict the heard differences reported by observers based strictly upon blind pairwise comparisons between sounds (i.e., the estimated perceptual differences/similarities formed by listeners blindly, on the basis of the auditory display alone). Other simple psychophysical tests involving pairwise discrimination in terms of identifiable attributes can certainly be considered, but have not been executed as yet.

Beyond these more elementary psychological measurement techniques, future development of the sonification systems under test will employ a broad range of evaluation methods, which have been chosen to address the most important issues in sonification system usability. In the final analysis, however, the completed sonification system must meet explicit acceptance criteria before its success is demonstrated. As outlined by Schneiderman and Plaisant [13], these criteria for evaluating system performance might include the following:

- Time for users to learn specific functions
- Speed of task performance
- Rate of errors by users
- User retention of commands over time
- Subjective user satisfaction

In addition to the overall satisfaction with the displayed sonification that may be expressed by system users with domain knowledge, which satisfaction may diminish with time, a more objective evaluation is to be recommended. It is not enough that users think that they can use a system effectively; rather, it is important to determine whether users can reliably make accurate judgments about the information being displayed as part of a typical use-case analysis. Thus, otherwise satisfying sonifications, which are nonetheless finding no support from the results of double blind testing, will eventually be rejected. Ultimately, it is hoped that such an approach will contribute to the formulation of a more general theory of sonification. Empirical results such as these might allow a sonification theory to evolve through a somewhat natural ‘winnowing out’ of unsuccessful approaches, supporting a general approach to sonification with the potential to fill ‘ecological niches’ with truly winning applications.

5. REFERENCES

- [1] C. Scaletti, and A. B. Craig, “Using sound to extract meaning from complex data,” *Electronic Imaging ’91*, San Jose, CA. International Society for Optics and Photonics, 1991.

- [2] J. J., Mezrich, S. Frysinger, and R. Slivjanovski. "Dynamic representation of multivariate time series data," *Journal of the American Statistical Association*, 79(385), pp. 34-40, 1984.
- [3] F. Grond, and J. Berger, "Parameter Mapping Sonification," In Hermann, T., Hunt, A., Neuhoff, J. G., editors, *The Sonification Handbook*, Chapter 15, pp. 363–397. Logos Publishing House, Berlin, Germany, 2011.
- [4] K. Karplus, and A. Strong. "Digital synthesis of plucked-string and drum timbres," *Computer Music Journal*, 7(2), pp. 43-55, 1983
- [5] P. Dutilleux, G. De Poli, A. von dem Knesebeck, U. Zölzer, "Time-segment processing," In *DAFX - Digital Audio Effects* (2nd ed.). John Wiley & Sons Ltd., 2011.
- [6] W. L. Martens, "Perceptual evaluation of filters controlling source direction: Customized and generalized HRTFs for binaural synthesis," *Acoustical Science and Technology*. 24(5), pp. 220-232, 2003.
- [7] Borg, I., and Groenen, P. *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York, NY: Springer, 2005.
- [8] S. Ferguson, W. L. Martens, and D. A. Cabrera, "Statistical Sonification for Exploratory Data Analysis," In Hermann, T., Hunt, A., Neuhoff, J. G., editors, *The Sonification Handbook*, Chapter 8, pp. 175-196. Logos Publishing House, Berlin, Germany, 2011.
- [9] G. H. Dunteman. *Principal components analysis*. SAGE Publications, Inc, Thousand Oaks, CA, USA, 1989.
- [10] S., Smith, H., Levkowitz, R. M., Pickett, and M., Torpey, "System for psychometric testing of auditory representations of scientific data," in *Proceedings of the International Conference on Auditory Display ICAD '94*, Santa Fe, New Mexico. 7–9 Nov, 1994.
- [11] S. Barrass, and G. Kramer, "Using sonification," *Multimedia Systems*, 7(1), 23-31, 1999.
- [12] T. L. Bonebright, and J. H. Flowers, "Evaluation of auditory display," In Hermann, T., Hunt, A., Neuhoff, J. G., editors, *The Sonification Handbook*, Chapter 6, pp. 111–144. Logos Publishing House, Berlin, Germany, 2011.
- [13] B. Schneiderman, and C. Plaisant, *Designing the user interface: Strategies for effective human computer interaction* (5th ed.), Boston, MA: Addison-Wesley, 2010.

SONIFYING THE SOLAR SYSTEM

Michael Quinton, Iain McGregor and David Benyon

Edinburgh Napier University, School of Computing,
Merchiston Campus, 10 Colinton Road, Edinburgh,
EH10 5DT, United Kingdom
{m.quinton, i.mcgregor, d.benyon}@napier.ac.uk

ABSTRACT

Sound is potentially an effective way of analysing data and it is possible to simultaneously interpret layers of sounds and identify changes. Multiple attempts to use sound with scientific data have been made, with varying levels of success. On many occasions this was done without including the end user during the development. In this study a sonified model of the 8 planets of our solar system was built and tested using an end user approach. The sonification was created for the Esplora Planetarium, which is currently being constructed in Malta. The data requirements were gathered from a member of the planetarium staff, and 12 end users, as well as the planetarium representative tested the sonification. The results suggest that listeners were able to discern various planetary characteristics without requiring any additional information. Three out of eight sound design parameters did not represent characteristics successfully. These issues have been identified and further development will be conducted in order to improve the model.

1. INTRODUCTION

Sonification of Scientific Data has gradually become more established since 1985 [01]. The large amounts of data that make up the basis of cosmic research suggest that sonification may be a suitable tool for data analysis, where the technique has been used to an increasing extent. For members of the general public spectacular images of the planets and other space related phenomena could potentially be sonically enhanced to include additional data not easily conveyed through traditional imagery. These sonic signatures might not only increase the entertainment value, but also help educationally. It may also improve accessibility, by providing a richer experience for those who are visually impaired. Sonification acts like a sound effect in a film where it enhances or indicates the nature of a particular parameter. For example if taking a scene from a film where the sound of a steam engine is heard but is not represented visually on screen then the viewer still feels that there is a steam engine in that particular scene and can relate to it.

1.1. Sonification and Planetariums

Scientists, composers and sound artists have explored and implemented sonification in planetariums usually in the form of installations, exhibits or performances based on musical compositions. These sonifications offered an educational and entertainment value to the audiences.

Through sonification, abstract concepts like planetary movements, can be made more tangible and comprehensible to the general public. Since many sonifications have been created as artistic pieces a certain degree of scepticism from scientists has made them wary of using sonification as a scientific tool. Barrass [2] refers to the conflict between the more traditional scientific view with relation to data analysis and the new view, which embraces the advantages of the human auditory system and its cultural significance. There is a fine line between sonification as a means of scientific exploration and merely being perceived as a popular mass media marketing tool.

Out of the 58 examples of sonification that have been mentioned by Dubus and Bresin [3] the majority of these are related to scientific applications. In this list of examples it is interesting to note that only three sonifications related to astronomical sciences are mentioned. When compared to other areas of scientific study sonification ranks as one of the poorer fields of science. Considering the large amounts of data involved in astrophysics the use of sonification in this field can be explored in more detail. The difficulty arises when trying to find the right balance between artistic input and sonifications that are usable in scientific data analysis. It is for this reason that an end user approach was adopted for this study.

1.2. Sonification

The multidimensional and multidirectional nature of hearing, or the *spherical* nature of sound as described by Sterne [4], allows for a number of characteristics of sonified data to be recognised. By listening to data it is possible to perceive patterns and structures that may not be apparent using visual methods. Certain sounds can be relegated to the background and given lower priority allowing the user to carry out other tasks while listening. There is also the ability to filter out certain sounds in order to be able to focus on specific sounds within a dataset [5]. By intentionally not discerning individual sounds, complex sounds can be perceived as a whole allowing the listener to hear multiple audio streams in parallel. Sensitivity to high temporal and high frequency resolution makes us distinctly sensitive to rhythm and pitch allowing listeners to distinguish minute changes in details and enables the use of complex datasets. From slight changes in sound, users can detect a variation in data through which the listener can convey the parameters affected. Apart from being able to handle other tasks while listening to data, there is also the advantage that it is possible to listen to data without having to look at a screen or to even be seated in front of it which is ideal for distance monitoring [6]. Hermann and Hunt [7] mention how humans are capable of identifying sound

sources, spoken words and melodies under noisy conditions. Sound can be a tool for navigation in a fixed image as it portrays movement and offers spatiality. Time dimension is also well supported by sound mapping. Sound can be a potent means of allowing users to create mental image associations and this in turn strengthens memory [8]. It is also advantageous from the point of view of data storage where a single channel of uncompressed 640 x 480 video equals 200 channels of CD quality audio [9].

1.3. Hearing in relation to Sonification

Considering the sophistication of human hearing, sound can potentially be utilised effectively in the process of data analysis. Sound parameters can be attributed to data in order to represent various characteristics. In Table 1 basic elements of hearing as described by Levitin [10], are compared to sonification mappings that were used in a number of sonification projects that have been listed and categorised by Dubus and Bresin, [3]. It is worth noting that the sonification mappings can vary accordingly and are not necessarily arbitrary, and it is often the sound designer who designates the various parameters according to need. Dubus and Bresin have managed to show that the majority of projects used certain mappings related to specific parameters consistently.

The above mentioned sonification mapping parameters have been used in the sonification process of cosmic data, such as the sonification of Kepler space telescope star data [11] where the brightness values of certain stars were observed over long periods of time. Fluctuations in the brightness values indicated that planets were passing between the Kepler telescope and the star that was being observed. In order to sonify the data the software ‘Sonification Sandbox’ was used. ‘Sandbox’ is multi-purpose software used for sonifying data. It allows users to map data to multiple auditory parameters such as timbre, pitch, volume and pan.

Table 1: - The relationship between hearing and sonification mapping possibilities

Parameter	Sonification Mapping
Loudness	Proximity, size, importance, energy
Pitch	Location, size, orientation, velocity, motion, size, distinction
Contour	This would represent the overall sonification
Rhythm	Intensity, density, speed
Tempo	Velocity, event rate
Timbre	Proximity, intensity, importance
Reverberation	Motion, location, proximity, spatialization

Sources: Levitin, [10] and Dubus & Bresin, [3]

One of the most commonly used techniques of sonification in Astronomy is Audification. NASA has published numerous examples of audification, such as those from waves captured by the Cassini spacecraft as it travelled

through our solar system. There are various examples of different solar system phenomena like solar winds, sounds from Jupiter, Saturn, the moons of Titan, Enceladus, the rings of Saturn and the Voyager 1 recordings from outside the *heliosphere* of our solar system (The heliosphere is the sun’s magnetic field inflated to gargantuan proportions by solar winds [12]). These radio waves have been transposed by software platforms such as xSonify and brought into the human range of hearing [13]. Data like these can be immediately transposed in pitch. For example, readings from Jupiter of radio-astronomy data led to the discovery of *whistlers, hiss* and *chorus*.

1.4. Sonification of Astronomical Phenomena

Dubus and Bresin [3] describe how scientists are becoming more accustomed to using sonification as an analysis tool and that the community of researchers using this new tool has grown substantially. Out of the few sonifications made for space physicists none have been designed and tested with the end user [14]. This results in the sonifications being inadequate for the task of data representation and end up not being used. Dubus and Bresin [3] make reference to comments by Scaletti [15] where she states that sound attributed to data can only be called sonification once it has been done with the intent of understanding or communicating something about the original information.

The use of sonification in discerning space data is growing but not a new field. Donald Gurnett has been sonifying data from spacecraft for decades [16]. It was believed that space was a vacuum and therefore sound was unable to travel through it [17] but it has now been discovered that space is not a total vacuum and that stuff does exist between the stars at very low densities and pressures which makes the sound waves inaudible, but sound waves can actually travel through space [18]. This argument is further supported by Professor Carolin Crawford [19] and in a presentation entitled ‘The Sounds of the Universe’ she states that, “sound can be used as a diagnostic of cosmic phenomena, indirectly tracing the behaviour of astronomical objects: - whether the presence of lightning on Jupiter, or the physical structure inside distant stars.”

Crawford [19] argues that sound is an effective means of illustrating certain aspects of astronomy, in particular radio signals. She also referred to the capability of sound to transfer significant amounts of energy across vast volumes of space. During this presentation Professor Crawford plays numerous examples of “sounds from space” but emphasizes the fact that hardly any of the sounds played during her lecture are the actual sounds recorded from space. Some of the sounds never existed. They are conversions of natural radio signals, which are part of the electromagnetic spectrum, into sound. For example, a recording of radio transmissions from the sun have to be speeded up 42,000 times from 0.1 Hz to 4.2 KHz in order for them to become audible.

One aspect of space science that has grown popular recently is that of planet hunting. There are a myriad of exoplanetary systems where thousands of planets are being found, over 5000 found to date and 1800 confirmed as planets [20], [21]. The model of the solar system alone is already a rich and vast playground of possibilities and overwhelming amounts of data. In order to plough through this vast sea of information an efficient and effective means of data analysis has to be utilised. The use of sound could not only facilitate this process and reduce the amount of

time needed to make these analyses, but also allows the listener to create mental images [8] and transport users onto for example, planetary surfaces, nebulae clouds and black holes.

By building an effective sonification of the solar system it can act as a prototype for building sonifications for exosolar planetary systems. The Planetarium model not only works as an educational tool for people who are not familiar with astronomy but it also acts as an indication that if people who are not familiar with the space sciences can discern what the sonification parameters represent then this would mean that a sonification designed for astro scientists using an end user approach may have more positive results.

2. METHOD

2.1. Participants

In order to design the sonification of the planets the user, in this case a planetarium representative, was involved in the data gathering process. An interview was conducted with a trained scientist and teacher.

For the testing of the sonification 11 people from the general public were interviewed, together with the representative from the planetarium. For the experiment 9 males and 3 females participated in the experiment with ages ranging from 24 to 56.

2.2. Materials

The data gathering interview was part multiple choice and part interview, where further questions were asked about the parameters addressed through multiple choice. Audio recordings were made of the interview to be later transcribed.

The testing of the sonified model took place at various locations, as no central meeting place could be found, and to better accommodate the participants' different schedules. The participants were sat down amidst four speakers, two in front and two behind their heads, at close proximity. The choice of four speakers was determined by the fact that the planetarium would be using 5.1 surround system where surround sound would only be working on a flat plain with no up and down movement. Volume levels were kept within a safe range below 85dBA SPL (Peak) and were controlled by the sound designer. The sound designer triggered and manipulated sound live from a DAW using a MIDI controller. The participants were given a questionnaire; the first 4 sections were multiple-choice questions on which the respondents marked their choices, whereas the fifth section contained open-ended interview style questions, which were audio recorded and transcribed.

2.3. Design

Questions for the data-gathering interview were based on information found on a planetary fact sheet by NASA [22]. The resultant sonified model would be a sonic representation of 8 planets and their orbital revolutions around the listener who would be situated in the position of the Sun.

Other questions were also added to the interview. Physical properties such as rock, gas, ice, liquid, metal and fire were included. These elements would be used in order to distinguish between one planet and another. The model

would be represented on a 4 channel surround system working on a flat plane configuration (no up and down movements) and would be working as an audio visual presentation.

Questions that related to the mechanics of the model were also included. The model could be speeded up or slowed down so that planets that revolve around the sun either at very slow or at fast speeds could be regulated. There was also the idea of being able to 'zoom in and zoom out' to a specific planet. This would mean that the planet would be brought closer to the listener by turning up the volume and making the timbre much brighter to replicate the impression that closer objects are louder and clearer than further objects. MaasØ [23] describes this phenomenon in relation to the human voice and its relation to the three acoustic characteristics related to distance i.e. volume, timbre and reverberation. He referred to how listening is more precise on a horizontal level but less accurate at estimating distance. Listeners have difficulties distinguishing whether a sound is coming from 7 or 8 feet away but it can easily distinguish between sounds that are 9 inches or 9 feet away. By adjusting the three acoustic characteristics mentioned it is possible to suggest a sense of distance.

2.4. Procedure

In Section 1 the planetarium representative was asked to grade the importance of each parameter. The grading was based on a five-point scale running from *not important* (1) through to *very important* (5). Parameters that were graded 1 or 2 were left out of the model. These parameters were Orbital eccentricity, planet surface pressure, the global magnetic field of a planet, perihelion and aphelion and rings of a planet. An exception was made in relation to Saturn's rings that are the planets most distinct characteristic. Planetary rings had not qualified as an inclusive parameter. Section 2 clarified which of the parameters graded at score 3 were actually important to the planetarium since the questions delved into more detail. Parameters that scored 4 or 5 were to be included. Some of the parameters that were given importance for the model were not direct parameters that could be sonically represented that easily. For example, diameter could not be directly represented as a sonification parameter. It would have to be represented as the size of the planet in the sound design through pitch. A total of 17 questions out of the 19 attributes mentioned in the NASA planetary fact sheet were included in this section. Five parameters were excluded from the final model and 12 parameters were included.

In the case of the testing of the model participants had to answer a number of multiple choice questions which were designed in order to identify whether users were able to discern what the sonification was representing and to grade accordingly. They were not told which planets they would be listening to. In Section 1 part i participants were asked about the planet Mercury, Section 1 part ii Venus, Section 2 Earth and Mars combined and Section 3 Jupiter, Saturn, Uranus and Neptune combined.

2.5. Development of Model

2.5.1. Parameter Mapping

Pitch was used to reflect the size of the planets. The scale ran between the notes C4 for Mercury (the smallest planet) and C2 Jupiter (the largest planet) the other planets were designated as follows: - Venus B3, Earth G3, Mars A3, Saturn D2, Uranus E2 and Neptune F2.

Sound design elements were used to represent atmospheric conditions, temperatures, air pressure, climate conditions etc. of each planet [24] [25]. Virtual synthesizers were used to create the sounds in the sound design process, which gave more flexibility than samples.

Each planet was also assigned with a low pass filter on each channel of the DAW that could be controlled manually during the experiment by using a MIDI controller allowing parameters such as cut off on the low pass filter to be altered live. This allowed the timbre to be manipulated during testing. Volume control was also controlled manually. The reason for not automating these parameters was so that the sound designer could create the zoom effect during the experiment.

Rhythm was linked to pulses representing the radio waves that are emitted by the planets. Tempo was related to the ability to speed up and slow down the model. This was not given importance by the representative of the planetarium. By speeding or slowing down the tempo one can hear faster planets more clearly and understand their orbits with more appreciation. A case in point is Mercury which only takes 88 days to orbit the Sun. On the other hand, at normal speed the very slow planets such as Neptune only get to revolve once around the surround sound system and therefore had to be speeded up so that the listener could hear the orbit of Neptune a couple of times around the sun. It takes Neptune approximately 165 years to orbit the sun once [24]. A visual model working on the same principle can be found at solarsystemscope.com [26].

A scale was created in order to represent the different speeds and celestial movements of the planets' orbits. When one observes various visual representations of the solar system it becomes evident that these representations are not according to scale. The number of orbits for each planet was worked out over 3 minutes running at a tempo of 5.33 seconds in a measure. Within the 3 minutes all 8 planets of the solar system would have at least looped around the sun once. Neptune the furthest of the planets only makes one revolution within this scale but this represents how slow Neptune actually is. In 2011 the planet completed one orbit since the date of its discovery [27]. Table 2 indicates how much time it takes each planet in order to make one revolution around the sun, and how these data have been sonified.

Table 2: Temporal Scale of planetary orbits [28]

Planet	Sonified time in minutes	Actual orbits in Earth days/ months/ years
Mercury	0.075	88 days
Venus	0.10	224.7 days
Earth	0.15	365 days
Mars	0.22	1 Year 11 months
Jupiter	0.43	11.9 Years
Saturn	1.05	29.7 Years
Uranus	1.53	84.3 Years
Neptune	2.55	164.8 Years

The orbit was replicated by a surround panner which sent the sound through four outputs of a soundcard and was transmitted through a quadraphonic configuration. Every planet was automated so that it would move through the surround in accordance to the planet's speed. The actual sonifications of each planet and of the solar system can be heard [29]: -

3. RESULTS

3.1. Results from the Data Gathering Interview

The parameters density, diameter of the planet, gravity, length of day of a planet, orbital period, mean temperature of the planet's surface and orbital velocity were given most importance by the planetarium representative (PR).

The parameters Mass of a planetary body, Distance from the Sun, Ability to zoom in and out on a planet, Atmospheric Characteristics were given less importance by PR but would still be included in the model.

Orbital eccentricity, planetary surface pressure, ring system, global magnetic field and Perihelion and Aphelion were to be excluded from the model.

There were 20 questions in section 2 and the replies were related to **Timbre**: - Closer planets are clearer than more distant ones (proximity), **Rhythm/ Duration**: - Would represent the radio wave pulse emissions from planets, **Tempo**: - Variable speed of the planetary movements by altering the BPM in the DAW, **Pitch**: - Smaller planets are higher in pitch than larger planets, **Loudness**: - Closer planets are louder than more distant ones (proximity), **Reverberation**: - Would represent distance.

3.2. Results from Testing the Sonification Model

3.2.1. Interpretation of the characteristics of Mercury

The first planet that the participants were asked to discern was information about the planet Mercury. The sonification of this planet was quite successful and participants were able to discern many of the planets characteristics. It is important to note that the listeners had no prior reference or

baseline to which they could associate to or compare and they were not told which planet or planets they would be listening to throughout the experiment. Considering this factor participants were still able to discern the planets size, gravitational influence and atmosphere successfully. The planet was deemed by P2, P4, P5, P6, P7 and P8 as being of an average size and not a large one. P2, P6, P7, P8 and P9 discerned the gravitational pull as being of average strength. Almost all the participants except for P2 were able to discern Mercury’s Magneto Sphere as one the planets main attributes and participants P1, P5, P6, P9, P10 and P11 were able to determine a lack of atmosphere or atmospheric conditions which is a precise discernment considering that Mercury does not have an atmosphere but instead has something called an *exosphere* made up of atoms which are blasted off its surface by solar radiation [27].

3.2.2. Interpretation of the characteristics of Venus

Venus was poorly represented sonically as a planet and listeners were not able to discern that many characteristics successfully. The only parameters that the participants were able to discern correctly was the strength of Venus’s gravitational pull. P1, P2, P3, P4, P9 and P11 could also successfully discern that Venus was a larger planet than Mercury and that Venus is close to the sun (P3, P4, P5, P6, P9, P10, and P11). Participants scored poorly with regards to the planets type were Venus was designated as an Icy planet of cold temperatures and of mild atmospheric conditions (P1, P3, P4, P5, P6, P7, P8, P11 and PR). The sound design for a fiery planet like Venus was suggested by PR and was described as a ‘chime-like popping’ sound. It was probably the metallic properties of the sound that gave the impression of Venus being a cold icy planet. PR had also felt that the planet was of an average temperature like many of the other participants. This gives a further indication that the sound design for this planet is ineffective and has to be revised.

3.2.3. Interpretation of the characteristics of Earth & Mars

The overall discernment between the two planets of Earth and Mars was fairly successful although there were slight problems with the sound design of Mars that had a negative effect on the results with regards to parameters concerning the planets’ size. Participants P1, P3, P5, P7, P8, P10 and PR could successfully determine that both planets were close to each other and that there were slight differences in size between the two planets (P4, P6, P7 and P9). The only problem was that participants P1, P2, P3, P6, P7, P8, P9 and P11 mistakenly deemed Mars as being bigger than the Earth. With regards to characteristics P1, P4, P5, P6, P7, P9 and PR were able to determine that Earth had rocky and liquid characteristics but were unable to successfully discern Mars as being a Rocky planet except for PR who was the only candidate to designate Mars as being a rocky planet. The sound design of Mars gave the impression of being a larger and colder planet with ice qualities due to the horn like qualities that were used for the sound design. The metallic quality of the sound gave the impression of coldness and the depth of the sound gave the wrong impression of largeness. The sound design for the planet Mars would have to be revised.

3.2.4. Interpretation of the characteristics of Jupiter, Saturn, Uranus & Neptune

In this section participants were asked to listen to the four planets of Jupiter, Saturn Uranus and Neptune simultaneously. The planets were introduced to the listeners one by one and then left to play at the same time for approximately three minutes. The results of this section were quite successful. This part of the listening experiment clearly indicated that P1, P2, P3, P5, P7, P8, P10, P11 and PR were able to hear different characteristics simultaneously and to be able to discern differences between the planets and recognize various characteristics from each planet. P1, P2, P3, P4, P5, P6, P10 and P11 found the experience of listening to four planets at the same time to be immersive. P2, P3, P5, P7, P8, P10, P11 and PR were able to distinguish the orbits of each planet clearly. Table 3 indicates the answers that participants gave for questions 4, 5 and 6 where for example in question 4 participants were asked to indicate how many planets they thoughts were either rock, ice, liquid, gas or fiery by writing down a number which would range from 1 to 4. In Question 5 the listeners had to distinguish how many planets’ were small, medium or large in size and in question 6 the participants had to work out the orbit speeds of the four planets by stating how many planets were orbiting at a fast, average or slow rate.

Table 3: Question 4 -6 Reported instances of parameters when comparing four planets

Parameter	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	PR	
4	Rock	1	2	1	1	1	2	1	1	1	1		
	Ice	1	1		1		1			1	1	1	
	Liquid	1		1	2	1	1	1			1		
	Gas	1	1	2		1		1	1		1		
	Fire					1		1	1		1	1	2
5	Small		1	1	2	1		2	1		1	2	
	Medium	2	1	2	2	1	1	2	1	1	2	1	2
	Large	2	2	1		2	3		2	1	1	1	2
6	Fast	1	1	1	1	1	1		1	1	1	2	2
	Average	2	2	2	2	2	2	2	1	2	2	1	
	Slow	1	1	1	1	1	1	2	2		1	1	2

3.2.5. General interpretation of all eight planets playing simultaneously

In this last part of the listening experiment the participants were asked to listen to all eight planets at the same time. The sound of each planet was triggered at the same time and was left to play for approximately three and a half minutes. In this section participants were asked more general questions related to main aesthetics. P2, P3, P4, P5, P8, P10, P11 and PR found that the soundscape was immersive and P1, P6 and P9 found that the sonification had a musical quality to it. P1, P3, P4, P5, P6, P9, P11 and PR found it to be harmonious and P2 and P8 found it familiar. There was only one participant, P7 that found the soundscape to be confusing. The listeners were then asked to determine whether or not they could follow differences

in orbit speeds, planet size, proximity and climate. Table 4 indicates the scores for these parameters. From the table the parameter climate is the least one that listeners were able to distinguish due to all the different sounds that were happening at once. Proximity was the characteristic mostly discerned by the listeners where they were able to perceive planets that were closer and others that were further away. Finally the participants were asked to classify the quality of the sound design ranging from poor to good. Most of the participants found the sound design to be good. P7 and P10 found it to be fair while P9 graded it as not bad.

Table 4: Parameters that listeners were able to distinguish clearly

Parameter	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	PR
Orbit
Size
Speed
Proximity
Climate

3.2.6. The Interview

In question 1 participants were asked the reason for the rating they had given the model in section 4. The participants generally commented that they found the model to be immersive, it gave a good idea of spatiality, that the sounds of the planets were distinct but at the same time there was a balance between them, and that the sounds evoked images of planets and planetary landscapes.

In the second question the participants were asked to elaborate upon what they liked about the model. In general participants liked the distinction between the planets, and the amount and quality of the detail that were portrayed by the sounds. PR was impressed by the way that proximity was presented and how through immersion one felt that something was coming closer or moving away.

In question 3 participants were asked what they disliked about the model. There were certain resonances from one particular planet that did disturb some of the participants, P6 and P7 were particularly bothered by this. PR said that there were times when it got confusing. Rather than being an element of dislike it was more a matter of losing focus and not being able to discern the detail anymore.

Question four asked the participants how they would create their own sound design of the solar system. Many of the participants felt that they were unable to answer this question due to a lack of knowledge regarding the subject of sound design. Other participants such as P3, P4, P6, P7, P10 and P11 picked up on points relating to using the size and representing the dimensionality by using pitch or using smaller sounds for smaller planets and larger sounds for larger planets. PR replied the following:

"... I think you got it quite right according to my ideas and tastes..."

Question 5 asked participants whether they thought the sound design was an effective tool for representing data and

question 6 asked whether sound design could enhance visual presentation of planets and whether the participants would use sound to represent the planets. All the participants including the PR agreed that sonification was an effective tool for representing data in one way or another and all felt that that sound would enhance the visual experience.

Only P10 disagreed that sound could be used as a scientific tool due to its subjective nature. P10 did agree that sound enhances the visual and could be used effectively to that extent. P8 and P11 felt that sonification helps to make scientific data accessible to the layman. This was a valid point especially when considered in the case of a planetarium where the users will not necessarily be knowledgeable about the planets. P3 made an interesting comment by stating that sound is a language that can be taught. Once sound has semantic value then discernment of data can be comprehended more easily by the listener and can be relayed to other listeners more effectively. P5, P9, P10 and PR found that sound acts as an effective memory tag and would also enhance visual memory much more. The general feeling was that sonification creates immersion that allows the listener to be drawn into the data and to share a more intuitive response to it.

4. DISCUSSION

4.1. The Subjectivity of Sound

In reference to Hegarty [30] where an isolated sound is important to guide the listener. If the listener is unable to create the suitable mental image through the sonification then it is ineffective.

Should sonification follow strict parameters where certain sounds always represent certain characteristics? This reminds us of Hermann's [31] comments that there were no specific guidelines determining how sonification is made. This remains to be a question of 'for and against' in the auditory display community. Let us take for example the high pitched and timbre sounds that seemed to evoke feelings of coldness in the participants of this experiment. There is an apparent trend in relation to sounds of this nature and the images they evoke in the listener. If further testing does indicate that the majority of participant's related high frequency sounds to a feeling of coldness, then this could become a standardised representation of this sensation. This makes the sonification reproducible, that the system can be used with different data and in repetition with the same data Hermann [31].

Although sound is subjective and everyone experiences it in their own way, there are common elements that work collectively [32]. There are various examples from the results that show common elements that the participants were able to discern. Idhe [33] emphasizes that the listener should be aware that one's beliefs will determine ones perspective of the sound and that the listener should listen to the sound itself. The beginning of the test might have induced a form of listening that searched for association since the null existence of one automatically caused the listeners to create an internal marking scale that they could relate to in order to be able to perceive the first planets that they listened to. As the test proceeded participants became more confident in their listening abilities. P10 had stated that the subjectivity of sound would not allow it to be an efficient scientific tool, but sound is measurable and it is

mathematical making it an efficient tool for representing numerical data.

4.2. Human Hearing vs. Parameter Mapping Sonification

The parameters of hearing given by Levitin [10] and the sonification parameters mentioned by Dubus and Bresin [3] in relation to hearing are loudness, pitch, contour, rhythm, tempo, timbre and reverberation and the sonification mappings in relation to these parameters are compared.

Participants could discern particular parameters because these sounds were mapped in accordance with the way in which human hearing works. In the case of proximity participants scored high. When the planets were played individually they were brought closer or taken further away from the listener by reducing the timbre and amplitude.

In the case of rhythm, listeners were able to relate to the orbits of the planets and to distinguish between the different speeds that the different orbits were moving at in comparison to each other rather than in comparison to the sun. Elements of reverberation were not emphasized in the testing and participants were not asked about it in the questionnaire.

Contour would be related to the participant's ability to hear the model as a whole. Eight participants, including the planetarium representative, heard the model as harmonious and two others heard it to be familiar. These participants were able to hear the overall impression created by the sound of all eight planets playing and to take it in as one whole composition.

Pitch was calculated by participants in relation to each other. This meant that one planet always had to act as the baseline for the others to be compared to. When planets were played individually it was difficult to guess the size, especially in the case of the very first planet where there was nothing to compare it to.

Regarding the human hearing vs. parameter mapping of sonification when considering the positive results from the test and especially since the participants had no visual aid, guidance or even a comparative baseline then it can be concluded that the parameter mapping was successful.

4.3. The End User Approach

In this research and in research carried out by Diaz Merced *et al.* [14] no sonifications have been found that have been tested with the end user in the field of Space sciences. Diaz Merced *et al.* [14] have been designing the sonification software xSonify over the years and the final improvements made to the program have been made by involving the end user in the process.

There is one common aim in both studies that is clearly defined in the Diaz Merced *et al.* [14] report. The sonification must act as a common platform where blind and sighted people can be aware of the same events through the sonification and will be able to share the similar knowledge of what they have achieved through the sonification process.

One difference noted between the studies is that in the Merced *et al.* [14] report there is no mention of any spatial representation of the sonified data. Only volume, pitch and timbre have been mentioned as the parameters within which the sonification is made. The planetarium model gave

importance to the spatial element. The spatial element helps to create immersion as was seen in the experiment carried out by Turner *et al.* [34] where a sense of a place was simulated in another place successfully. This is the level of immersion that can be achieved by using a surround configuration and also allows parameters to be mapped out more easily. For space data exploration this could be essential as multiple layers of data can be distributed in ways that make the data easier to listen to. Sterne [4] states that hearing immerses its subjects, it also places the listener inside an event and vision gives perspective.

5. CONCLUSION

It is interesting to note that this listening experiment was conducted on an audience of non-scientific people whose knowledge of astronomy is limited and yet they were able to discern a lot of detail from the model. The people were not involved in the sound design process itself. It is also not known how much knowledge the participants actually had with regards to the solar system. The planetarium representative was able to determine characteristics more precisely because he was more knowledgeable about the subject and because he was involved in the sound design process too. If the lay person was able to determine so many details of the sound design without any prior guidance or knowledge of what they were listening to then this reflects that sonification is an effective means of representing data. With a couple of adjustments that would have to be made in order to address the problems with certain aspects of the sound design with regards to Venus and Mars then the Solar System model could then act as a comparative model for exosolar planetary systems. This is the same approach that is usually employed by scientists by comparing exosolar systems to our solar system in order to determine how these planetary systems work. If the sonification of a planetary system could be conducted with Astro scientists as the end user and where the sonification is specifically designed and mapped out according to their needs then they may be able to determine much more from such a sonification model and to be able to use it efficiently as a scientific tool. If the sonification is used with a visual component then the effectiveness of any solar planetary model will be enhanced.

As future work there are improvements that need to be made with regards to the sound design of certain planets. Once the sound design has been arranged testing can start again and a fresh batch of participants can be chosen in order to widen the sample to see what works and what does not and a more consistent sonified model of the solar system can be built and can also find use outside the planetarium market. The model of the solar system can act as a guideline or basis so that further sonifications for exo-solar planetary systems can be built and can also be used as a comparative model against exo-solar systems. The work on exo-solar planetary systems will be aimed at Astro scientists that work in the field of exo-solar planetary science. The sonification can be used as a scientific tool which the scientists can use in order to analyse large portions of data and to find similar patterns or differences between the different systems.

6. REFERENCES

- [1] Hermann, T., & Hunt, A. (2005). Guest editors' introduction: An introduction to interactive sonification. *IEEE multimedia*, (2), 20-24.
- [2] Barrass, S. (2012). The aesthetic turn in sonification towards a social and cultural medium. *AI & society*, 27(2), 177-181.
- [3] Dubus, G., & Bresin, R. (2013). A systematic review of mapping strategies for the sonification of physical quantities. *PLoS one*, 8(12), e82491.
- [4] Sterne, J., (2012). *The sound studies reader*. New York :Routledge,
- [5] Lunn, P., & Hunt, A. (2011). Listening to the invisible: Sonification as a tool for astronomical discovery.
- [6] Vogt, K., de Campo, A., & Eckel, G. (nd) An Introduction to Sonification and its Application to Theoretical Physics.
- [7] Hermann, T., & Hunt, A. (2011). *The sonification handbook*. Berlin: Logos Verlag.
- [8] Minghim, R., & Forrest, A. R. (1995, October). An illustrated analysis of sonification for scientific visualisation. In *Proceedings of the 6th conference on Visualization'95* (p. 110). IEEE Computer Society.
- [9] McGee, R. (2009), *Auditory Displays and Sonification: Introduction and Overview*.
- [10] Levitin, D. J. (2011). *This is your brain on music: Understanding a human obsession*. Atlantic Books Ltd.
- [11] Winton, R. J., Gable, T. M., Schuett, J., & Walker, B. N. (2012). A sonification of Kepler space telescope star data.
- [12] NASA/gov, 2014, accessed June 2015, http://science.nasa.gov/science-news/science-at-nasa/2013/01nov_ismsounds/
- [13] NASA/ JPL, 2006, accessed June 2015, <http://saturn.jpl.nasa.gov/news/cassini/features/feature20060424/>
- [14] Diaz-merced, W.L., Brewster, S., Candey, R.M., Schneps, M. (2013) "A study of the use of a sonification prototype by astrophysicists". CHI'13, Paris, France
- [15] Scaletti C (1994) *Auditory display: sonification, audification and auditory interfaces*, Addison Wesley Publishing Company, chapter 8: Sound synthesis algorithms for auditory data representations. pp. 223-251.
- [16] Feder, T., (2012) 'Shhhh. Listen to the data', Print edition, 65, 20-22, DOI: <http://dx.doi.org/10.1063/PT.3.1550>
- [17] Blackstock, D. T. (2000). *Fundamentals of physical acoustics*. John Wiley & Sons.
- [18] O'Brien, T (2014) 'Sounds of Space', <http://proftimobrien.com/2014/03/sounds-of-space/>, accessed April 2015
- [19] Crawford, C. (2011) 'Sounds of the universe', Gresham College, accessed May 2015, <http://www.gresham.ac.uk/lectures-and-events/the-sounds-of-the-universe>
- [20] PlanetQuest (2015). 'Exo planets 2020', NASA Jet Propulsion Lab, accessed August 2015, <http://planetquest.jpl.nasa.gov/news/208>
- [21] Exoplanets.org (nd), accessed August 2015, <http://exoplanets.org/>
- [22] NASA, 2004, 'Planetary fact sheet', accessed July 2015, http://nssdc.gsfc.nasa.gov/planetary/factsheet/planetfact_not es.html
- [23] Maasø, A. (2008). The proxemics of the mediated voice. *Lowering the boom: critical studies in film sound*, 36-50.
- [24] Space.com, 2015, accessed July 2015, <http://www.space.com/>
- [25] NASA Solar System Exploration, 2015, accessed June 2015, <https://solarsystem.nasa.gov/planets/index.cfm>
- [26] solarsystemscope.com, nd, accessed July 2015, <http://www.solarsystemscope.com/>
- [27] Space.com, 2014, accessed July 2015, <http://www.space.com/36-mercury-the-suns-closest-planetary-neighbor.html>
- [28] Universe Today, 2009, Last accessed July 2015, <http://www.universetoday.com/37507/years-of-the-planets/>
- [29] Quinton, M (2016), Soundcloud, last accessed April 2016, https://soundcloud.com/michael-quinton_napier
- [30] Hegarty, P., (2012) 'A Chronicle Condition: Noise & Time', Chapter 1, Goddard, M., Halligan, B., & Hegarty, P. (Eds.). (2012). *Reverberations: The philosophy, aesthetics and politics of noise*. Bloomsbury Publishing USA.
- [31] Hermann, T. (2008). Taxonomy and definitions for sonification and auditory display.
- [32] Chion, M. (1994). The three listening modes. *Audio-Vision: Sound on Screen*, 26-32. COMOS EU, (2013) 'From X rays to music', accessed June 2015, http://www.fp7-space.eu/Newsletter_Archive/COSMOS+_Newsletter12.pdf
- [33] Ihde, Don. 1974. The Auditory Dimension. In *Listening and Voice: A Phenomenology of Sound*. Athens: Ohio University Press. Pp. 49-55, Sterne, J. (Ed.). (2012). *The sound studies reader*. Routledge.
- [34] Turner, Phil. McGregor, Iain. Turner, Susan. Carroll, Fiona., 2003. "EVALUATING SOUNDSCAPES AS A MEANS OF CREATING A SENSE OF PLACE" (July): 6-9.

3D TIME-BASED AURAL DATA REPRESENTATION USING D⁴ LIBRARY'S LAYER BASED AMPLITUDE PANNING ALGORITHM

Ivica Ico Bukvic

Virginia Tech
SOPA, DISIS, ICAT
Blacksburg, VA, USA
ico@vt.edu

ABSTRACT

The following paper introduces a new Layer Based Amplitude Panning algorithm and supporting D⁴ library of rapid prototyping tools for the 3D time-based data representation using sound. The algorithm is designed to scale and support a broad array of configurations, with particular focus on High Density Loudspeaker Arrays (HDLAs). The supporting rapid prototyping tools are designed to leverage oculocentric strategies to importing, editing, and rendering data, offering an array of innovative approaches to spatial data editing and representation through the use of sound in HDLA scenarios. The ensuing D⁴ ecosystem aims to address the shortcomings of existing approaches to spatial aural representation of data, offers unique opportunities for furthering research in the spatial data audification and sonification, as well as transportable and scalable spatial media creation and production.

1. INTRODUCTION

In today's rich data driven society strategies for optimal data experience and comprehension are more important than ever. Humans are biologically predisposed to experiencing rich environmental data multimodally [1], warranting research into individual modalities' potential in promoting data comprehension and interpretation delivered through technology. Such research serves as the foundation for their combined utilization to broaden cognitive bandwidth and clarity [2]. In this respect, visual data exploration or visualization has arguably seen greatest progress. This may be in part because of human predisposition to visual stimuli, as well as because visualizations have had a rich history [3] that both predates and inspires today's technology-centric approaches.

Audification and sonification [4] are relatively new but nonetheless thriving research areas. In particular, they offer a diverse array of complementing and competing approaches to spatial aural representation of data. With auditory spatial awareness covering practically all directions [5], it is a dimension that exceeds the perceivable spatial range of the visual domain. Apart from the simple amplitude panning [6], audio spatialization approaches include Ambisonics [7], Binaural [8], Depth Based Amplitude Panning (DBAP) [9], Vector Based Amplitude Panning (VBAP) [10], and Wave Field Synthesis (WFS) [11]. The following paper fo-

cuses primarily on spatialization strategies that are reproducible in physical environments and offer physical affordances with minimal amount of idiosyncrasies, such as the vantage point, without requiring additional technological support, e.g. a motion tracking system. For this reason, due to its specific context that does not meet the aforesaid criteria the paper excludes the Binaural approach from the discussion below.

2. CATALYST

This project was inspired primarily by the newfound space whose hybrid HDLA implementation exposed new audio spatialization research opportunities and challenges. Virginia Tech Institute for Creativity, Arts, and Technology's (ICAT) Cube is an innovative space with a hybrid audio infrastructure capable of supporting all of the aforesaid approaches to spatializing sound, with particular focus on WFS, Ambisonics, and VBAP (Fig.1). It is a 50x40x32-foot (WxLxH) blackbox space with catwalks and mesh ceiling whose audio infrastructure is centered around the idea of discovery and experimentation, including audification and sonification. ICAT's Cube offers a unique hybrid 148-channel audio system designed in collaboration with ARUP Acoustics inc. In order to accommodate the various spatialization algorithms, it consists of a 124.4 homogeneous loudspeaker array offering several horizontal layers of varying density: a high density ear-level equidistant 64-channel array and additional three loudspeaker layers with 20 channels each, including a 20-channel ceiling raster. The 124-channel system is complemented by 4 symmetrically positioned subs centered on each side of the first level catwalk. The system also offers an additional 17-inch sub focusing primarily on sub-50Hz frequencies. It can be further complemented by 10 mobile floor-level loudspeakers. Cube also offers nine ceiling-mounted ultrasonic audio spotlights, including four mounted onto a motorized, remotely-controlled arm.

In a space designed for transdisciplinary research that needs to be capable of near seamlessly transitioning from one spatialization technique to another and/or concurrently employing multiple approaches, such an implementation is not without a compromise. Cube's WFS relies on a proprietary Sonic Emotion Wave 1 system [12] that enables its implementation using sparser loudspeaker configuration. Ambisonics require careful calibration due to cuboid shape of the loudspeaker configuration [10]. Finally, VBAP due to algorithm's inability to handle irregular densities, particularly the ear-level layer, utilizes only select ear-level and ceiling loudspeakers, therefore relying more on the virtual sound positioning than what a localized amplitude panning system may



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

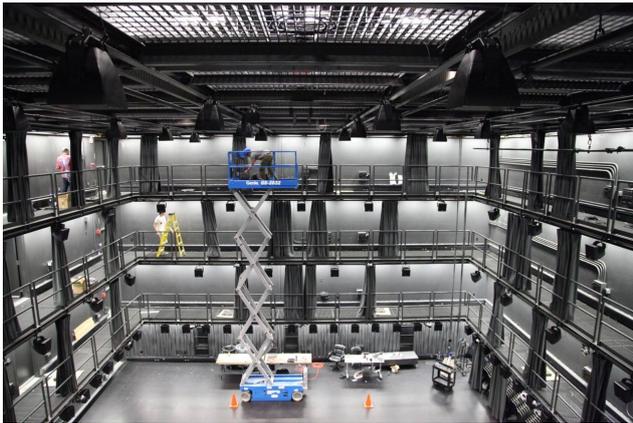


Figure 1: Virginia Tech ICAT Cube.

ostensibly require (Fig.3). Furthermore, each of the aforesaid configurations provides limited transportability among various spaces, including 900 square foot ICAT Perform Studio's 2-layer 24.4 Genelec system, and the the Digital Interactive Sound & Inter-media Studio (DISIS) classroom offering 8.2 single-layer Genelec system. Apart from the WFS plane wave [11] and VBAP's source spread (a.k.a. MDAP) [13] that manifests itself in a form of a regular circle-like shape around the source's center, none of the spatialization approaches offer an easy and controlled way of projecting sounds through multiple physical sources, particularly when it comes to irregular shapes. Likewise, none of the currently available technologies provide the aforesaid features in a way that can easily scale among varying loudspeaker configurations while utilizing all of the available physical sources and their superior localization over that of virtual ones [14].

Based on the observations attained through the newfound Virginia Tech signature audio research space, several inconsistencies have emerged that limit broader applicability of the preexisting approaches to audio spatialization with particular focus on audification and sonification scenarios (listed in no particular order):

1. Support for irregular High Density Loudspeaker Arrays (HDLAs);
2. Focus on the ground truth with minimal amount of idiosyncrasies;
3. Leverage vantage point to promote data comprehension;
4. Optimized, lean, scalable, and accessible, and
5. Ease of use through supporting rapid-prototyping time-based tools.

Recent Computer Music Journal solicitation [15] has defined HDLAs as "systems addressing 24 or more independent loudspeaker". In this paper HDLAs are further defined as loudspeaker configurations capable of rendering 3D sound without having to rely solely on virtual sources or post-processing techniques.

2.1. Support for flexible loudspeaker layouts

While most of the aforesaid spatialization algorithms are HDLA and therefore 3D capable, most implementations favor certain loudspeaker configurations, e.g. tightly spaced loudspeakers in

WFS or triangular loudspeaker placement in VBAP and High-Order Ambisonics [16]. DBAP is configuration-agnostic, but also requires additional features, such as *spatial blur*, designed to minimize problems of spatial coloration and spread typical of both VBAP and DBAP [9]. Recent research further suggests for some of these approaches there may be ways to utilize less common configurations (e.g. Blue Ripple Sound's Rapture3D for irregular loudspeaker arrangements using HOA [17] or the proprietary Sonic Emotion systems that allow for sparse WFS arrays [12]). Due to their proprietary nature, currently the limits of these solutions is not known, nor how well and/or how reliably they may be able to scale and/or accommodate systems whose irregularity significantly deviates from the prescribed configuration (e.g. sparse vs. irregular loudspeaker distribution in WFS), particularly in HDLA environments. VBAP solution, like the one implemented in the Virginia Tech's Cube utilizes only some of the loudspeakers in order to attain the desired triangular organization among the loudspeakers, leaving a number of physical sources unused (Fig.3). Given the physical sources' superior audio spatialization potential over the virtual ones, such a solution was found incapable of harnessing the full potential of CUBE's audio system.

2.2. Ground Truth with Minimal Amount of Idiosyncrasies

Each of the aforesaid spatialization approaches is encumbered by unique idiosyncrasies that limit the ease of their applicability in a broad range of scenarios. These idiosyncrasies can be seen as a detriment towards developing generalizable sonification strategies in part because they can also cloud the prospect of identifying the ground truth. The most obvious one is the aforesaid sensitivity of various approaches to loudspeaker configurations. Similarly, positioning a virtual (e.g. Ambisonics [7] and specialized cases in WFS [18]), and physical sound sources (e.g. 4DSound [19]) inside the listening area offers great promise. Yet, their respective idiosyncrasies, such as the sweet spot (e.g. WFS aliasing, and lower order Ambisonics), custom and ostensibly intrusive hardware (4DSound), and the computational complexity (e.g. Ambisonics) that currently lacks out-of-box solutions, particularly when associated with non-standard loudspeaker layouts, limits their universal applicability. Similarly, WFS's ability to place sounds outside the listening space may allow for more uniform perception of the sound source, yet doing so will also limit the power of the vantage point that may help in clarifying source's location and its relationship to other adjacent sources depending on listener's location. Lastly, DBAP introduces spatial blur to compensate for potential spatial coloration and spread inconsistencies.

2.3. Leveraging Vantage Point

In this paper the author posits for a system to provide optimal listening environment, it needs to mimic affordances of our everyday lives as long as its implementation does not exacerbate one of the observed limitations. Vantage point is one such affordance that enables listeners to perceive both the rendered aural data within the context of their immediate environment, as well as perceive rendered data communally in a location-specific fashion. Unlike virtual sources within the listening area that also introduce limiting idiosyncrasies, vantage point is essentially intrinsic to simpler amplitude-based algorithms. This allows for a closer study of a particular angle, or even positioning oneself closer to the loudspeaker perimeter to elevate perceived amplitude of a source or

texture of interest, something that may prove particularly useful in data audification and sonification.

The vantage point limitation brings out another important consideration in pursuing a more universal and transportable approach to spatializing sound—loudspeaker perimeter based spatialization. Instead of relying on idiosyncratic sound processing that enables rendering of virtual sources within and outside the loudspeaker perimeter, the spatialization should ideally focus on the perimeter-centric rendition, an approach that offers relatively straightforward mapping of multidimensional data onto the loudspeaker perimeter, reinforces the vantage point, and makes it considerably easier to reproduce in varied and flexible HDLA scenarios.

2.4. Optimized, Lean, Scalable, and Accessible

Ideally, a system should be lean—it should rely on the preexisting tool frameworks where possible, ensuring that at its very core it is simple and maintainable with minimal redundancies. This is certainly the case with some of the implementations that are typically embedded in digital signal processing languages, including Max [20], Pure-Data [21], and Supercollider [22], or provided as plugins (e.g. VST, LV2 plugins, or Audio Units). Such implementations can leverage the vast resources of those toolkits to further enhance their functionality and flexibility.

In terms of rendering spatial data using sound, one of the additional considerations is system’s responsiveness and how that responsiveness scales from conventional stereophonic to HDLA scenarios. Ideally, such a system should be capable of rendering a scene in real-time and under low-latency conditions. While low-latency operation is not necessarily critical in controlled tests, its absence may limit system’s applicability and broader appeal, both of which are essential for wider adoption and potential standardization across multiple sites and contexts.

Although all of the aforesaid systems offer real-time and low-latency performance, some (e.g. WFS and HOA) require careful space- and loudspeaker-layout-specific calibration that may not be easily accessible out-of-box. In particular, when considering systems with cutting-edge features (e.g. Wave 1), their proprietary nature may render them as prohibitively expensive black box implementations with more complex HDLA configurations requiring special design and licensing. This can also be seen as a potential factor in limiting the access to such solutions and consequently their transportability.

2.5. Rapid-Prototyping Tools

If implemented well, rapid prototyping tools have a unique ability to go well beyond representing loudspeaker positions and their respective amplitudes. By interfacing with multidimensional data sources, such tools have the potential to lead to cross-pollination of generalizable standards across various modalities and by doing so serve as a scaffolding in domains whose standards are yet to be solidified. For instance, being able to interact with visual representation of audio spatialization may lead towards leveraging standards and techniques associated with visual drawing and painting and using those to guide the development of corresponding methodologies in the spatial aural domain.

Sound is a time-based modality and for this reason, rapid prototyping tools should go beyond providing the ability to position a sound source. They could also include a way of altering their location over time, as well as visualizing the outcomes of such

a change. With the exception of Sonic Emotion’s Wave 1 [12], 4DSound [19], D-Mitri system [23], VBAP-based Zirkonium [24], and recently introduced Sound Particles [25], HDLA spatialization systems are devoid of any time-based data that can be easily synced with other time-based content (e.g. video or an abstract data feed), typically requiring users to create their own middleware to drive such systems in real-time and/or render their audio feeds into a multichannel audio file. While offering ability to visualize loudspeaker configuration, it is currently unclear if DBAP offers any rapid prototyping tools. Similarly, it remains unclear whether Sound Particles is capable of rendering real-time low-latency audio nor what is its CPU overhead in doing so.

Within the context of audification and sonification, none of the existing off-the-shelf systems offer easy interfacing with multidimensional data sets and their translation into a spatialized sound.

2.6. Other Considerations

Based on the observed limitations, the author of this paper posits that the ideal platform for pursuing a generalizable approach to spatial data representation using sound should mimic as closely as possible real-world environmental conditions our multisensory mechanisms are accustomed to experiencing, leveraging, and interpreting. More so, it should do so with minimal technological complexity and idiosyncratic limitations. Such a system is more likely to integrate and cross-pollinate with other modalities and in return leverage their preexisting body of research to identify optimal mapping strategies. Furthermore, the author argues that such cross-pollination in particular between visual and aural may offer a useful scaffolding to sonification theory based on the existing body of research in the visual domain. In a pursuit of such a solution the technology presented in this paper focuses primarily on data sets with up to four dimensions.

3. INTRODUCING D⁴

D⁴ is a new Max [20] spatialization library that aims to address the aforesaid limitations by:

1. Introducing a new lean, transportable, and scalable audio spatialization algorithm capable of scaling from monophonic to HDLA environments, with particular focus on advanced spatial manipulations of sound in audification and sonification scenarios, and
2. Providing a collection of supporting rapid prototyping time-based tools that leverage the newfound audio spatialization algorithm and enable users to efficiently design and deploy complex spatial audio images.

Below we will focus primarily on the spatialization algorithm that in part builds on author’s prior research [26] and its newfound affordances that have a potential to serve as a foundation for the further exploration of the auditory display paradigm.

3.1. D⁴’s Algorithm

At the very core, D⁴ is driven by the newly proposed Layer Based Amplitude Panning (LBAP) algorithm. LBAP is rooted in a straightforward sinusoidal amplitude panning algorithm which amounts to:

$$L_{amp} = \cos(L_{distance} * \pi/2), \quad (1)$$

$$R_{amp} = \sin(L_{distance} * \pi/2), \quad (2)$$

L and R variables stand for left and right channels spatially oriented from listener's perspective in clockwise fashion, respectively. $L_{distance}$ is a normalized value between 0 and 1 and where the ensuing amplitude value between 0 and 1 is used to modulate the outgoing audio signal for both L and R channels.

In 2D arrays of varying densities, e.g. horizontal ear-level arrays, the math for manipulation between loudspeakers remains essentially the same, with the only addition being the awareness of the loudspeaker and source positions in horizontal space expressed as an angle (0-360 degrees). By a simple calculation, one can either identify perfect physical source (a loudspeaker) or two adjacent loudspeakers and using the aforesaid function calculate the amplitude ratios between the two. What makes this approach particularly convenient is its ability to utilize irregular densities across the perimeter with the only caveat being decreased angle perception resolution in areas that may be sparser in terms of loudspeaker spacing and therefore more reliant on virtual sources (Fig.2).

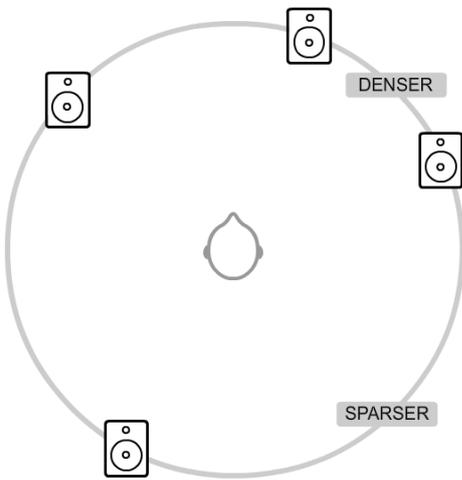


Figure 2: Irregular 2D perimeter loudspeaker array's localized reliance on virtual sources and its inversely proportional relationship to the array's immediate density.

When applying the same algorithm in a 3D environment where there are multiple horizontal layers of loudspeakers positioned around the perimeter, the aforesaid algorithm is typically superseded by VBAP [27] and more recently DBAP [9]. Where VBAP begins to fall apart is when horizontal loudspeaker layers are populated with varying densities and consequently irregular distances among loudspeakers. This is certainly the case with the ICAT Cube where the upper levels host only 20 loudspeakers as opposed to 64 loudspeakers at ear level and where such a configuration make sense given human decreased spatial perception accuracy of elevated sound sources. This, however, is not the only scenario. Similar limitations can theoretically also occur in spaces whose architectural design precludes equal loudspeaker distribution, something that DBAP aims to address albeit with added complexity and ensuing idiosyncrasies. For instance, there may be acoustic considerations, structural beams, pillars, walls, and other physical structures that prevent loudspeaker placement. When employing

VBAP, such setups fail to provide usable adjacent triangles, as is the case with ICAT Cube (Fig.3), and while one can skip physical sources in order to retain triangular configuration, such a solution precludes the use of all physical sources, resulting in a less than ideal scenario, particularly when considering preferred higher loudspeaker density at ear level where human perception, depending on head orientation, offers greatest angular resolution. Another option is using a hybrid system, so that the secondary spatialization approach utilizes the higher density layer. This, however, further limits system's transportability and introduces an entirely new array of idiosyncrasies.

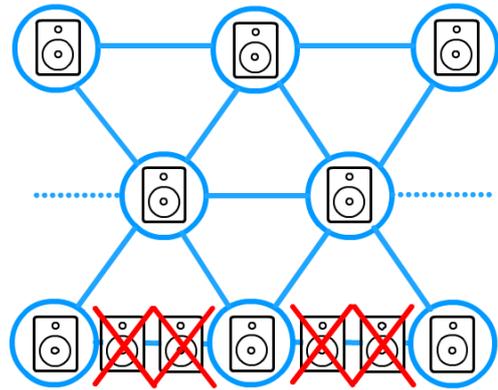


Figure 3: VBAP's selective use of loudspeakers in irregular layered loudspeaker configurations.

LBAP aims to address this problem by introducing an amplitude panning variant that relies on the core notion that the entire perimeter-based audio system is separated into a series of layers with each layer being assigned shared elevation and each loudspeaker further identified by its azimuth (Fig.4). In this respect vertical surfaces with loudspeaker rasters above, as is the case with the ICAT Cube's ceiling, and below are treated as a series of concentric circles, which is a feature that loosely resembles D-mitri and MIAP's grouping. In cases where there are no loudspeakers below or above, the lowest and highest layers assume any sound that moves below or above their elevation respectively should be cross-faded across the layer itself (Fig.5). While this is less than ideal, it can be easily remedied by adding an additional layer below (should the architecture allow for doing so), while leveraging existing infrastructure to the best of its ability.

Once the layers are identified, LBAP uses one vertical cross-section as the elevation reference. Doing so will enable for the sound to easily traverse individual layers horizontally (as it should) without having to compensate for vantage point deviations in elevation (e.g. loudspeaker in a far corner will effectively have lower elevation than one immediately next to the listener that belongs to the same layer (Fig.6). In cases where sound does not neatly fall onto one of the physical sources or a single horizontal layer, LBAP based on source's elevation first identifies its closest two layers, the one below and one above where the virtual source is located. Once the two layers are identified LBAP calculates their respective amplitude ratios as follows:

$$Above_{amp} = \cos(Below_{distance} * \pi/2), \quad (3)$$

its integration into the ICAT Cube which utilizes four subs centered on each of the four sides of the 1st level catwalk as the first independent layer and with an additional 17-inch subwoofer that provides rumbling lows for the entire space from a single source as the second independent layer. Consequently, the algorithm inherently allows for use of a single loudspeaker per layer, resulting in 100% of the original generated amplitude emanating from that speaker regardless of the source's position and/or radius.

3.2. Advanced Sonification Features

Apart from WFS' plane wave [11] or VBAP's source spread (a.k.a. MDAP) [13] that manifests itself in a form of a circle-like shape around the source's center projected onto the 3D loudspeaker perimeter, none of the spatialization approaches offer an easy and controlled way of projecting sounds through multiple physical sources, particularly when it comes to irregular shapes. What arguably sets D⁴ apart from other spatialization algorithms is its re-imagined approach to growing point sources using Radius, and the Spatial Mask, as well as a suite of supporting spatialization tools that leverage these newfound affordances.

3.2.1. Radius

Each point source's default radius is assumed to be 1°. As it grows, based on proximity calculated as a linear distance between source's location and radius and physical loudspeaker's position, it spills over adjacent loudspeakers with its amplitude decreasing in all directions using the sinusoidal amplitude panning curve. As a result sounds with a diameter of 180°, cover entire sphere with the opposite edge being essentially inaudible. At 360° diameter, the overlap between the outer diameters when coupled together (and further limited not to exceed maximum allowable amplitude) amount to 100% of the original amplitude (Fig.7).

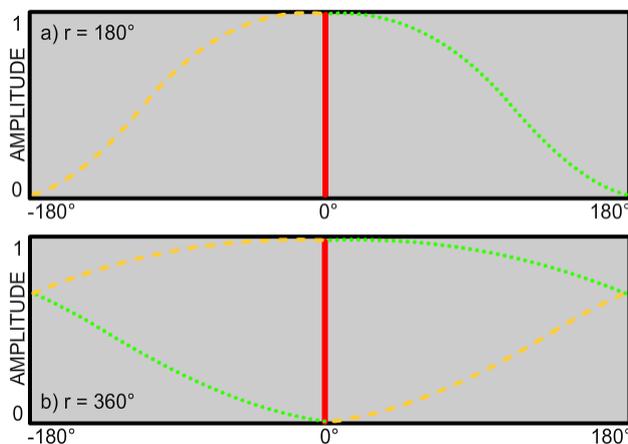


Figure 7: Sinusoidal amplitude curve applied to source radius in a single dimension. Thick red line denotes relative source's angle within one dimension. Yellow striped and green dotted lines denote two radius vectors across the said dimension. Example a) shows 180° diameter or 90° radius with no overlap, and b) 360° diameter or 180° radius with overlap.

3.2.2. Spatialization Mask

D⁴'s Spatial Mask (SM), akin to that of its visual counterpart considers the entire spherical space to have the default mask of 1. This means wherever the point source and whatever its radius, it will populate all the loudspeakers based on the computed amplitude. The spatial mask, however, can be changed with its default resolution down to 0.5° horizontally and 1° vertically, giving each loudspeaker a unique maximum possible amplitude as a float point value between 0 and 1. As a result, a moving source's amplitude will be limited by its corresponding mask value as it traverses the ensuing spherical perimeter. This also allows a situation where a point source with 180° radius or 360° diameter that emanates throughout all the loudspeakers can now be dynamically modified to map to any possible mask. When coupled with time-based visual editing tools, this equates to essentially aural painting [26] in both 2D and 3D. We will further explore SM and its features as part of the D⁴ Rapid Prototyping Tools section below.

4. SIDE-STEPPING LIMITATIONS

D⁴'s implementation of the LBAP algorithm is a lean implementation in that it relies on Max's framework. Consequently, when coupled with Max's battery of digital signal processing objects, it allows for greater extensibility. For instance, through the use of a collection of included abstractions, D⁴ library offers access to an otherwise complex form of movable sound sources, including angled circular motion, and the ability to control attack and trail envelope, effectively resulting in the aural equivalent of the motion blur (MB). D⁴ also offers easy way of interfacing with Max's Jitter library [28] that offers vector optimization when calculating multidimensional matrices, something that has proven particularly useful when working with the Spatialization Mask.

Informal LBAP an D⁴ tests have shown it is capable of providing critical low-latency real-time rendering of spatialized audio sources even in >100 HDLA and high-audio-stream-count scenarios. This makes it particularly useful in interactive environments. Its inaugural implementation as part of a Tornado simulation that premiered in the fall 2014 features 1,011 internal 24-bit 48KHz audio streams or channels stemming from two dozen concurrent point sources that are mixed down and outputted through the 124.4 CUBE loudspeaker system in real-time with audio latency of 11ms (512-byte buffer) between the time an action is initiated and the sound leaving the computer. D⁴'s implementation of the LBAP algorithm is designed to scale from monaural to as many loudspeakers as the system (CPU and audio hardware) can support. The current version offers a growing array of optimizations, including omission of unnecessary audio streams and bypassing redundant requests.

D⁴ offers both single- and multi-threaded implementations. The multithreaded version, however, has offered only marginal improvement over its single-threaded counterpart. This is likely due to the fact that the built-in algorithm's implementation maximizes reliance on the built-in Max objects and as such in and of itself does not bear significant CPU footprint. More so, whatever the savings in terms of CPU utilization due to distribution across multiple CPU cores are replaced by the newfound overhead required to synchronize concurrent audio streams through a high number of interrupts required by the low-latency setting. Further testing is warranted to attain a better understanding of the CPU overhead in single- and multi-threaded scenarios.

One of the greatest challenges of the HDLA audio content is its transportability. Fixed media tends to be distributed as pre-rendered multichannel sound files that are often accompanied by a simple Max patch or an equivalent tool capable of interfacing with often unconventional HDLA configurations. The target venue, however, may not have the same number of loudspeakers, requiring either sound to be re-rendered (assuming the existing system lends itself to easy reconfiguration), or calling for compromises in determining which channels need to be omitted or doubled. As an alternative, a live version may be used where sound sources are coupled by a system that renders entire piece in real-time, requiring engine that is adaptable and reconfigurable. D⁴ aims to address transportability by providing a simple one-step reconfiguration consisting of loudspeaker channels and their respective azimuths and elevations provided in an ordered (bottom-up, clockwise) layered approach that instantly updates all instances within the Max ecosystem and adapts the spatialization algorithm for a newfound loudspeaker arrangement. With its real-time low-latency scalable engine D⁴ can also leverage the aforesaid implementation within the live and interactive aural spatialization of data, as well as artistic contexts.

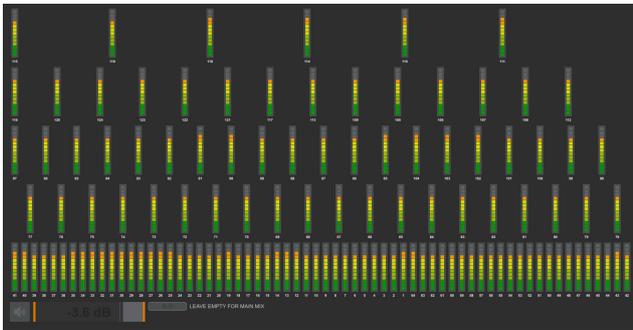


Figure 8: An instance of D⁴ library's signal monitor.

With its integration into Max, D⁴ immediately benefits from the built-in debugging and signal monitoring tools. With the help of the Jitter library, its spatialization capabilities can be easily translated into visual domain. The same has enabled D⁴ access to external control surfaces. For instance, the aforesaid Tornado simulation offers iPad interface for controlling the simulation from the Cube's floor through the use of the Max's Mira library. In addition, it offers a visual level monitor built out of a collection of abstractions that enable users to easily customize and design new space-specific level monitors. Given the exponential complexity of signal flow in HDLA scenarios, the entire D⁴ ecosystem is virtual audio bus aware, and offers a collection of visual tools, a.k.a. monitors specially tailored to harness this feature (Fig.8). By assigning a bus name to a particular monitor, it will automatically switch to monitoring all outputs from that bus, while leaving the bus name blank will revert to monitoring main outs. Similarly, the library provides a global main out whose adjustments affect all its instances. By default, D⁴ comes with monitors for three Virginia Tech spaces, including DISIS, and ICAT's Cube and Perform studio, and offers easy way of creating new site-specific level monitors using a collection of abstractions.

5. D⁴'S RAPID PROTOTYPING TOOLS

D⁴ library also offers a series of rapid prototyping tools. Below we'll provide a brief overview of its 2D and 3D editors and means of importing data sets.

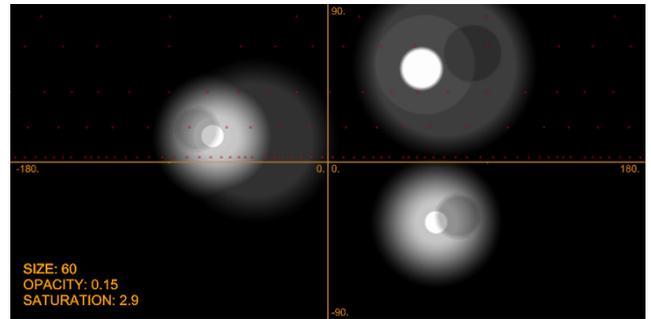


Figure 9: D⁴ library's 2D mask editor.

2D mask editor (Fig.9) is a two-dimensional representation of the loudspeaker perimeter unfolded onto a plane with the x axis covering the full circle and y axis covering 90 degrees above and below. Apart from the usual representation of angles and a cursor, the visualization also auto-populates various layers, or as is the case with the ICAT Cube, the 124-speaker array (red dots) and its complementing 4-sub array (green dots). Jitter is used to allocate area around each loudspeaker up to the half-point between it and the adjacent loudspeakers. This area is used to calculate loudspeaker's overall amplitude based on its average grayscale color, with the black color denoting silence and white color 100% of the original amplitude.

To edit sound's mask, user is provided a customizable cursor that, akin to that of a digital drawing software, can be resized and its brush altered by varying transparency and saturation. Furthermore, the user can translate the SM both in conjunction with sound's rotation or independently of it. The editor also provides brush mirroring around the texture's x axis edges to simplify cross-fading across the visual seam generated by unfolding the mask onto a finite 2D plane. The ensuing mask can be fed either in real-time or on demand to the desired sound object. It can be also stored for time-based use and/or storage we will briefly discuss as part of the time-based editing features below.

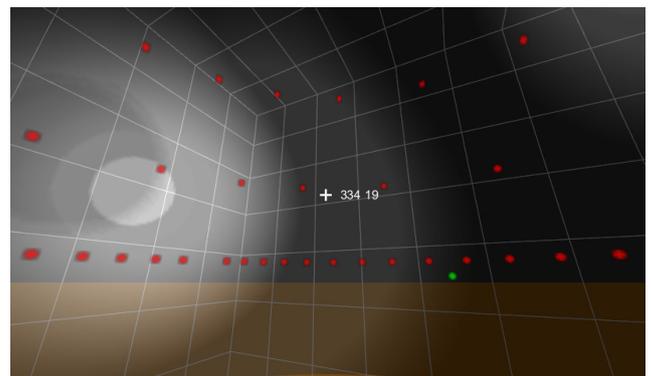


Figure 10: D⁴ library's 3D mask editor.

3D Mask Editor (Fig.10) is a three-dimensional counterpart

to the aforesaid 2D version. It allows for the exact same viewing and editing of the mask albeit from a 3D vantage point within the space, allowing user to pan the view around. This allows users to place 2D drawings in the context of the actual 3D space. While the default visualization provides a space-agnostic cuboid space, inspired by ICAT Cube's setup, given Max's flexibility, its layout can be easily altered, including importing actual 3D meshes of the target space. The ensuing drawing is stored in the identical way as the 2D drawing and the two are mutually interchangeable.

5.1. Time-Based Editing

While Mask Editors provide an easy way of populating sound sources in a cloud-like configuration throughout the HDLA space, their full potential is realized by leveraging accompanying time-based editor. Each of the mask snapshots can be stored as a 4-dimensional matrix (x, y, color, sequential keyframes). The matrix is accompanied by a coll object data structure which further contains timing of each keyframe and information on whether the transition between the current and next keyframe requires interpolation or not. In its initial release there is only one linear form of interpolation available between frames with other forms to be introduced in a later version based on user feedback. In addition, SM can be translated across x and y axes (corresponding to azimuth and elevation). Such interpolation is processed in parallel to interpolation (cross-fading) between SM keyframes. This can effectively serve as a secondary means of simulating cloud (as opposed to point) source's location.

Given the mask editor is fairly CPU intensive, for more complex real-time rendering saved editor renditions packaged as matrix-coll data containers can be retrieved and replayed using a considerably leaner Spatial Mask Player. Doing so enables playing multiple concurrent instances with minimal CPU overhead.

What makes D⁴'s approach to spatialization potentially useful as a platform for auditory displays is its ability to interface with the vast collections of spatial data and their translation into the 3D aural domain. Transferring visual data can be achieved by exporting it into an array of grayscale images, using format such as MJPEG, and importing it as a matrix into the editor. By relying on this feature alone, one could separate RGB channels into separate layers, effectively creating an audification engine of a movie footage. D⁴ editor also allows for synchronization with external clocks using SMPTE, and can adjust internal pace in respect to the sync.

6. ADVANTAGES

Based on the observed features, LBAP and the D⁴ library offer a number of advantages over the existing approaches that may be relevant to the audio display research, as well as the live and production scenarios, including support for irregular HDLAs, transportability, focus on the ground truth with minimal idiosyncrasies, vantage-point aware, optimized, lean, scalable, and accessible, and with the help of a growing number of rapid prototyping tools, the ease of use with particular focus on mapping multidimensional data onto spatial audio.

D⁴'s design focuses on rapid prototyping and implementation, leveraging existing battery of Max objects wherever possible, and consequently the pursuit of maximum flexibility. Such hybrid, mostly open source (MOSS) approach to software distribution is envisioned to isolate aspects that are easily modifiable by community and thereby encourage iterative improvement through

community participation, while retaining control over the core algorithm and its still evolving APIs. As a result, the library is also implemented as a potential drop-in replacement for the existing approaches to spatialization that predominantly rely on the azimuth/elevation value pairs. Although amplitude overages are unlikely, as a safety precaution, LBAP further implements hard limiting per physical output channel, preventing amplitudes that exceed 1 or 100% of the incoming sound.

7. LIMITATIONS AND FUTURE WORK

While LBAP's simplicity essentially makes it capable of addressing just about any 3D loudspeaker layout that can be reasonably described as a collection of horizontal layers, D⁴ library's rapid prototyping tools do not account for corners in cuboid scenarios as potentially special cases, something that would affect both the amplitude and the vantage point elevation. In informal listening tests the dichotomy between the assumed spherical azimuth/elevation loudspeaker location assignment and the actual cuboid layout of the ICAT Cube has not revealed observable deviations mainly because the azimuth and elevation hold true in both cases, with the ostensible amplitude variation due to differing distances between the listener and individual loudspeakers being below the observable threshold.

The same layered approach may make LBAP not applicable to certain scenarios. While some such scenarios are delineated for instance in DBAP paper [9], it is currently unclear how necessary or useful such a feature may be, particularly within the context of spatial audification and sonification. To address this, LBAP's layers could be ostensibly applied in a way where such layers are not treated as parallel, albeit at a potentially significant increase in algorithm's complexity.

Unlike D-Mitri and MIAP, D⁴ is currently not capable of grouping loudspeakers. While similar results can be achieved through the use of the Sound Mask and/or independent layers, there is clearly a need for potential use of groups in the system's future iterations. Another limitation is the lack of multiple independent multilayered contexts. Currently, the system supports one multilayered context and virtually unlimited number of additional independent layers. It is unclear whether it makes sense to have multiple concurrent multilayered contexts exist within the same space.

It is worth noting that Jitter operations D⁴'s SM relies on are not designed to take place per audio sample and as such its visual tools have more limited resolution than the audio itself. While the system provides built-in audio interpolation this remains one of the potential limitations, particularly when it comes to exploring innovative approaches to spatial amplitude modulation that goes beyond the fifty keyframes per second. Outside such extreme cases, the number of possible keyframes has proven more than adequate.

8. OBTAINING D⁴

D⁴ is currently under development with the anticipated commercial release in the summer 2016. For licensing enquiries contact the author at ico@vt.edu.

9. REFERENCES

- [1] M. H. Giard and F. Peronnet, "Auditory-visual integration during multimodal object recognition in humans: a

- behavioral and electrophysiological study,” *Journal of cognitive neuroscience*, vol. 11, no. 5, pp. 473–490, 1999. [Online]. Available: <http://www.mitpressjournals.org/doi/abs/10.1162/089892999563544>
- [2] S. Y. Mousavi, R. Low, and J. Sweller, “Reducing cognitive load by mixing auditory and visual presentation modes,” *Journal of Educational Psychology*, vol. 87, no. 2, pp. 319–334, 1995.
- [3] M. Friendly, “A Brief History of Data Visualization,” in *Handbook of Data Visualization*, ser. Springer Handbooks Comp.Statistics. Springer Berlin Heidelberg, 2008, pp. 15–56, doi: 10.1007/978-3-540-33037-0_2. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-33037-0_2
- [4] G. Kramer, *Auditory display: Sonification, audification, and auditory interfaces*. Perseus Publishing, 1993. [Online]. Available: <http://dl.acm.org/citation.cfm?id=529229>
- [5] T. R. Letowski and S. T. Letowski, “Auditory spatial perception: Auditory localization,” DTIC Document, Tech. Rep., 2012. [Online]. Available: <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA563540>
- [6] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning,” *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=7853>
- [7] J. Daniel and S. Moreau, “Further study of sound field coding with higher order ambisonics,” in *Audio Engineering Society Convention 116*. Audio Engineering Society, 2004. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=12789>
- [8] B. Carty and V. Lazzarini, “Binaural HRTF based spatialisation: New approaches and implementation,” in *DAFx 09 proceedings of the 12th International Conference on Digital Audio Effects, Politecnico di Milano, Como Campus, Sept. 1-4, Como, Italy*. Dept. of Electronic Engineering, Queen Mary Univ. of London., 2009, pp. 1–6. [Online]. Available: <http://eprints.maynoothuniversity.ie/2334>
- [9] T. Lossius, P. Baltazar, and T. de la Hogue, “DBAPdistance-based amplitude panning.” Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2009. [Online]. Available: <http://www.trondlossius.no/system/fileattachments/30/original/icmc2009-dbap-rev1.pdf>
- [10] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning,” *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=7853>
- [11] K. Brandenburg, S. Brix, and T. Sporer, “Wave field synthesis,” in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2009*. IEEE, 2009, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5069680
- [12] E. Corteel, A. Damien, and C. Ihssen, “Spatial sound reinforcement using Wave Field Synthesis. A case study at the Institut du Monde Arabe,” in *27th TonmeisterTagung-VDT International Convention, 2012*. [Online]. Available: <http://www.wfs-sound.com/wp-content/uploads/2015/03/TMT2012.CorteelEtAl.IMA.121124.pdf>
- [13] V. Pulkki, “Uniform spreading of amplitude panned virtual sources,” in *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*. IEEE, 1999, pp. 187–190. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=810881
- [14] G. Parsehian, C. Jouffrais, and B. F. Katz, “Reaching nearby sources: comparison between real and virtual sound and visual targets,” *Frontiers in neuroscience*, vol. 8, 2014. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4151089/>
- [15] “CMJ Call for Works.” [Online]. Available: <http://www.computermusicjournal.org/HDLA-call.html>
- [16] F. Hollerweger, “An Introduction to Higher Order Ambisonic,” *April 2005, 2013*. [Online]. Available: <http://flo.mur.at/writings/HOA-intro.pdf>
- [17] “Custom Layouts in Rapture3d Advanced | Blue Ripple Sound.” [Online]. Available: <http://www.blueripplesound.com/custom-layouts>
- [18] E. Corteel, C. Kuhn-Rahloff, and R. Pellegrini, “Wave field synthesis rendering with increased aliasing frequency,” in *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008. [Online]. Available: <http://www.aes.org/e-lib/online/browse.cfm?elib=14492>
- [19] “4dsound,” <http://4dsound.net/>. [Online]. Available: <http://4dsound.net/>
- [20] M. Puckette, “Max at seventeen,” *Computer Music Journal*, vol. 26, no. 4, pp. 31–43, 2002. [Online]. Available: <http://www.mitpressjournals.org/doi/pdf/10.1162/014892602320991356>
- [21] —, “Pure Data: another integrated computer music environment,” in *PROCEEDINGS, INTERNATIONAL COMPUTER MUSIC CONFERENCE*, pp. 37–41, 1996. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.3903>
- [22] S. Wilson, D. Cottle, and N. Collins, *The SuperCollider Book*. The MIT Press, 2011. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2016692>
- [23] “D-Mitri : Digital Audio Platform | Meyer Sound,” <http://www.meyersound.com/product/d-mitri/spacemap.htm>. [Online]. Available: <http://www.meyersound.com/product/d-mitri/spacemap.htm>
- [24] C. Ramakrishnan, “Zirkonium: Non-invasive software for sound spatialisation,” *Organised Sound*, vol. 14, no. 03, pp. 268–276, Dec. 2009. [Online]. Available: <http://journals.cambridge.org/article.S1355771809990082>
- [25] “Sound Particles - Home,” <http://www.sound-particles.com/>. [Online]. Available: <http://www.sound-particles.com/>
- [26] I. Bukvic, D. Gracanin, and F. Quek, “Investigating artistic potential of the DREAM interface: The Aural Painting,” in *Proceedings of the International Computer Music Conference, 2008*.
- [27] J. Cofino, A. Barreto, and M. Adjouadi, “Comparing Two Methods of Sound Spatialization: Vector-Based Amplitude Panning (VBAP) Versus Linear Panning (LP),” in *Innovations and Advances in Computer, Information, Systems Sciences, and Engineering*, ser. Lecture Notes

in *Electrical Engineering*, K. Elleithy and T. Sobh, Eds. Springer New York, 2013, no. 152, pp. 359–370, doi: 10.1007/978-1-4614-3535-8_31. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4614-3535-8_31

- [28] M. Song, S. A. Mokhov, and S. P. Mudur, “Course: Rapid advanced multimodal multi-device interactive application prototyping with Max/Jitter, processing, and OpenGL,” in *Games Entertainment Media Conference (GEM), 2015 IEEE*. IEEE, 2015, pp. 1–2. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7377246

DESIGNING INFORMATIVE SOUND TO ENHANCE A SIMPLE DECISION TASK

Keith Nesbitt, Paul Williams, Patrick Ng, Karen Blackmore, Ami Eidels

University of Newcastle
University Drive, Callaghan, NSW, Australia
Keith.Nesbitt@newcastle.edu.au

ABSTRACT

In this paper we examined the role of informative sound in a simple decision-making game task. A within-subject experiment with 48 participants measured the response time, success rate and number of timeouts of the players in a number of eight-second decision tasks. As time proceeds, the task becomes easier at the risk of players timing out and reducing the overall opportunities they will have to attempt the task. We designed a simple informative sound display that uses a tone that increases in amplitude over the duration of the task. We test player performance in three conditions, no sound (visual-only), constant (non-informative) sound and increasing (informative) sound. We found that the increasing sound display significantly reduced timeouts when compared with the visual only and constant sound versions of the task. This reduction in timeouts did not impair the players' performance in terms of their success rate nor response time.

1. INTRODUCTION

The aim of this study is to understand the informative use of sound in a simple decision-making task. We are motivated to better understand interaction in computer games where a player's fast decision-making is often critical to performance. The amount of time taken to make such decisions can be affected by the amount of information the player possesses in their current situation. Assuming that extra information allows players to perform better at such game tasks, this paper investigates the way sound can be used to provide multi-modal support for decision-making.

This 'informative' use of sound is designed to provide additional feedback related to the players' own actions, as well as key events and states of the game world. That is, the player can gather information about the game environment by relying upon auditory as well as visual cues. This might be advantageous in situations where visual cues are unhelpful because the eyes are already engaged in processing other signals. Alternatively, auditory displays may provide a more optimal modality for information when temporal cues are required.

This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

In terms of computer games, the role of sound in has certainly evolved since the classic laser sound effects and monotone background music used in nostalgic games such as Space Invaders [1]. Indeed, sound has become an integral part of the experience provided by modern computer games.

Sounds for computer games have traditionally been designed much like sound for motion pictures, as an adjunct to the visual experience. In films, music is used to establish the mood of the scene as well as to evoke tension and emotional responses from the audience [2]. Sound effects in film tend to enhance the realism of the scene with the intention of creating greater levels of immersion for the viewer.

Rightly or wrongly, computer game designers tend to focus on visual perceptual cues when designing the game levels [3]. Like the sound in films, the auditory effects are mainly added to enhance the visual experience [4]. In accordance with this approach, much research on sound display within video games has focused on how sound enhances players' experience and immersion [4-8].

Of course, one consequence of using sound solely for visual enhancement is that the design of more informative sound can be overlooked. For some user groups, such as the visually impaired, this can even exclude them from being able to play the game [9]. A more subtle consequence is that the full potential of using sound to convey useful messages is not always exploited in games. This is despite many studies within the field of auditory display [4,10-17] that provide evidence for the value of auditory feedback. It is clear that in many situations, well-designed sounds can provide important, additional feedback for computer users [7, 17-22].

While sound is usually designed as an adjunct to visual experiences in computer games, some games do exploit sound in more informative ways. These include Papa Sangre [29], a horror themed audio game for the iOS mobile platform that uses sound effects to guide the player in the dark environment. Likewise, the recent Thief series [30] integrates sound into the gameplay and uses it as the primary feedback for navigation. Informative sound is also present in some online multiplayer games such as World of Warcraft [31] where the sounds provide a more general informative function that supports player orientation and the identification of key situations and states [21].

Many approaches exist for supporting the design of sound displays. These include a case-based, metaphorical approach for aligning the informative function of sounds to listening encounters from the real world [23] and a structured multi-sensory taxonomy with guidelines that considers all

modalities for the display [24]. Another approach relies on the use of Auditory Design Patterns [12,25,26]. This approach has been considered for both general auditory display [12] and specifically for use in video games [25,26].

Probably, the two most well-known techniques described for displaying information through sound are Auditory Icons [10] and Earcons [13]. The technique known as Auditory Icons was first investigated as a means of extending the use of visual interface icons to the auditory dimension [10]. Based on ‘everyday listening’ skills, this approach maps information to recognisable sounds from the real world. By using a recognisable sound, the user can intuitively understand the current action or event suggested by the sound. For example, hitting a tin with a stick is an event that generates a sound. The sound itself conveys information about the material and size of the tin and if it is full or hollow. The sound also conveys information about the materials involved, the frequency of hitting and the force of the hitting. This is information we naturally learn to interpret from our everyday experiences.

There are many instances of natural-like sounds used to augment computer interfaces. For example, the SonicFinder integrated running and pouring sounds in the Macintosh interface to represent file manipulations on the desktop [11]. In the SharedARK application, a virtual physics laboratory for distance education included sounds such as hums that mapped to the state of the physics being simulated [27]. The ARKola bottling system, mapped sounds to equipment in a soft-drink factory, introducing audio cues for monitoring the bottling process [27].

The second common method of designing informational sound is the use of Earcons [13]. Earcons are abstract, synthetic tones that are structured to create auditory messages. This approach relies on ‘musical listening’ skills as it conveys information using musical properties of sound such as rhythm, timbre, and the pitch of notes. This can be contrasted with Auditory Icons that use everyday listening skills rather than acquired musical expertise.

Studies on the effectiveness of Earcons for conveying information have been conducted since the 1990s. Brewster et al [14] experimentally tested the effectiveness of Earcons in providing navigational cues within a structured menu hierarchy. The results found that 81.5% of participants successfully identified their position in the hierarchy, indicating that Earcons can be a powerful method of conveying structured information. Polotti et al [28] evaluated the use of rhetorical Earcons to map common operating-system functions in a graphical interface and found that subjects benefited from additional sound feedback when performing key tasks such as cutting and pasting.

Compared with Auditory Icons, Earcons have the advantage of being able to convey complex information about events to the user without any natural associations with a sound source. On the downside, Earcons require prior understanding of the mapping between the sound and the event before the information can be recognised. By contrast, Auditory Icons are considered to be more intuitive as they leverage the existing listening skills of users.

Currently, only a limited amount of work relating Auditory Icons and Earcons to computer games has been reported. Jørgensen [21] noted that both these approaches can play a role in terms of enhancing control functions for the player by extending the player’s current range of vision. There are also some taxonomies of sound usage described in the context of games [32-34]. The use of Earcons and Auditory Icons and their relationship to player performance

in Defence of the Ancients 2 (Dota 2) [35], a popular multiplayer online battle arena game, have also been reported [36].

However, the informative use of sound has a much longer history of study in domains outside of games [18]. With applications have been reported in diverse domains ranging from file management [11] to hospital operating rooms and vehicle safety systems [39-41]. Leveraging these informative approaches to using sound can potentially allow more critical information to be integrated into game interfaces. This intent would be to improve traditional usability criteria such as effectiveness, utility and efficiency [10,11,26,37,38] without impacting on the immersive experience that games strive for.

Interestingly, the effects of additional sound information, on top of existing visual information, are not always beneficial. The auditory Stroop effect [42] demonstrates how performance can deteriorate when the visual and auditory information are in conflict. Moreover, given the ultimately limited capacity of the brain to process information [43,44], additional sources of information, though relevant to the task at hand, may overload the system and impair performance. Thus, the potential benefit from adding auditory information to visual displays is not trivial and requires careful empirical scrutiny. This is precisely the aim of the current study, to evaluate user performance in a multimodal decision-making task.

2. A SIMPLE DECISION MAKING TASK

We developed a simple, custom-built decision-making game, called Buckets. It was designed to allow for the controlled collection of performance data, something that can be difficult in commercial games due to the complexity of interactions. Our Buckets game consists of a repeated task that was initially designed as a visual-only perceptual challenge for measuring how players employ strategy, balancing the risks and rewards associated with game mechanics [45].

Balancing risk and reward is an important consideration in the design of computer games and has even been likened to the thrill of gambling [46]. Of course, if players gamble on a strategy, they assume some odds, some amount of risk. In gameplay it is reasonable to expect that greater risks will be compensated by greater rewards. Adams not only states that each “risk must always be accompanied by a reward” [46] but also describes this as a fundamental rule for designing computer games.

In the Buckets game, players must solve a perceptual challenge, deciding which of four rectangles (buckets) is filling up with dark blue dots (rain) the fastest. The game is comprised of repeated trials. On each trial a new display of four buckets appears and the player has one attempt to determine by a key press which of the four items is the target. A new trial with a fresh display appears after a response, irrespective of whether it was correct or not. The player’s overall goal in the game is to identify as many target buckets as possible within a fixed time period. The longer a player waits on each individual attempt (trial), the more likely it is that the attempt will result in a positive outcome.

This is because as each attempt progresses, more pixels accumulate in the target bucket, making it easier to discern the one correct bucket from the three incorrect ones. A typical strategy might be to attempt faster responses as this rewards the player with more time for additional attempts. An alternative strategy is to reduce the risk of each attempt

by waiting longer to improve chances of correctly identifying the target bucket.

At the start of each trial four buckets are displayed (see Fig. 1). Each bucket consists of 5,000 pixels (50 wide x 100 high). At the start of the game each of the four buckets are 50% filled with 2,500 dark blue pixels and 2,500 white pixels. The actual position of the 50% of dark pixels is chosen randomly at every update of the display. For each trial one of the four buckets is randomly chosen to serve as the target. Each decision may last only up to 8 seconds within the trial, where this target bucket will gradually increase its number of blue pixels until it is 52.5% filled. This number was chosen by trialing the game and using the empirical data to set a difficulty level that gave players a 40-60% chance of success. If no response was given within 8 sec, a new trial begins and no score will be added to the player.

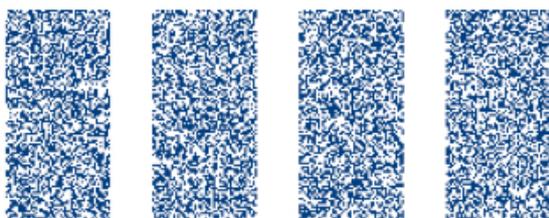


Figure 1: Buckets Perceptual Challenge

The frame rate of the game is configured to ensure that the display is updated at 10 frames per second. This means that about 14 pixels of extra filling are added to the target bucket at each frame. The actual frame rate of the game is monitored to ensure it meets this required number of frames per second.

Note that a player has only 8 seconds to respond, if they wait too long they timeout, losing the opportunity to make a selection. After collecting initial empirical data for the game we found an unexpected consequence of the design was that many players would experience these timeouts. This would negate the benefit that was meant to accrue by waiting longer to make a decision. To address this timeout problem, a simple sound alarm was designed. A beep was used to warn the player that they had 2 seconds left to respond. Unfortunately when trialing this solution the alarm proved distracting to some players, diverting them from their primary task and forcing them to make an immediate decision rather than allowing them to maintain focus on their primary perceptual challenge.

While such auditory alarms are commonly used as warnings, they are intended to divert user attention away from their current task. However, peripheral sounds have also been found useful for background monitoring of system states. We therefore implemented a simple background sound that increased in amplitude over the 8 seconds of the task decision. This sound can be described as an auditory icon as the increasing sound acts much an alarm of an approaching car. The sound becomes louder (and more dangerous) towards a critical moment in time. There are some contraindications for using amplitude in this way [14], so we also considered increasing the frequency of the sound, however this required us to resolve the complex relationship between pitch and amplitude [17], something that was difficult to resolve in the software platform we used. Initial trials with this increasing amplitude sound, anecdotally at least, created a suitable alarming signal and thus we adopted this approach for further empirical testing.

3. METHOD

The sound-augmented Buckets game was tested by comparing player performance in three conditions: *no sound*, *constant sound* and *increasing sound*. The first, ‘no sound’ condition used the original, visual only version of the Buckets game. The second condition used a ‘constant sound’ generated using a sine wave of fixed frequency (440 Hz). Musically, this corresponds to A in the fourth octave. This sound played at constant amplitude throughout the 8 seconds of each attempt. This condition was intended as a further control for comparing performance with the ‘increasing sound’ condition. The increasing sound was generated as a sine wave where the frequency of the signal was held constant at 261 Hz. Musically this corresponds to middle C. The amplitude of the signal gradually increased, in a linear fashion, over the 8 seconds of each trial. The increasing sound was pre-recorded as an 8 second wav file that was triggered for play at the start of each trial.

We tested the game using a repeated measure design with 48 psychology and computing students, and academics, from the University of Newcastle. Psychology students were awarded course credit for their participation. The study was approved by the University of Newcastle’s Research Ethics Committee. Participants were predominantly male (71%) and ranged in age between 18-54 years, with an average age of 21. All participants had normal, or corrected normal, vision and hearing.

The experiment was conducted within a computer laboratory. On arrival, each participant was assigned a workstation that displayed the Buckets game. All data was collected using an Apple Mac Pro running OS X 10.8 Mountain Lion. The game was played online using the Mozilla Firefox (Version 22) web browser and the Flash Player (version 11.4). Each participant wore a full sized headphone (AKG K44) during the whole experiment, even in the no sound condition. During the experiment the volume level on the operating system was set at the lowest possible volume (1 out of 16 bars).

Each participant played in each of the three conditions: no sound, constant sound and increasing sound. The order of the three conditions was counter-balanced across participants to control for effects learning and fatigue effects. Participants were randomly allocated to an order of conditions. Regardless, each participant received one minute of practice time in each condition before playing that condition competitively for 15 minutes. These 15 minutes were further divided into three blocks of, five minutes each.

As each condition began, the participant was presented with the game rules. These rules emphasized the importance of both accuracy and speed in the task. At the end of each individual trial the player received feedback for 500 milliseconds regarding their choice of buckets. A green tick was displayed below the target square if the decision was correct. This was accompanied by a cash-register sound. If the player choose incorrectly a red cross was displayed below the target and a sigh-of-disappointment sound played. Where the player timed out an alarm clock was displayed, and a typical alarm clock sound was played.

At the completion of each of the five minute blocks, participants received summary information about their performance, namely their number of correct and incorrect responses. At the end of each sound condition the subject was allowed a two minute break before commencing the next assigned condition. Overall subjects completed the experiment in about 60 minutes.

The player's response time for each attempt, and their timeout status for that attempt, were recorded for later analysis. Data regarding the correct target bucket (1-4) and the player's actual selected bucket (1-4) for each attempt were also saved.

4. RESULTS

Overall, the players completed 20,173 trials, with 8,347 of these trials resulting in correct responses (41.38%). There were 11,369 (56.36%) incorrect responses and a further 457 (2.27%) timeouts recorded. The task was designed to allow for a success rate between 40-60% (to allow a sufficient number of both correct and incorrect trials for analysis, see [45].) The results show this preliminary goal was achieved, so we move on to two further types of analysis. We first performed (section 4.1) further analysis on the pooled data to gain an overall appreciation of the data. This pooling process results in unequal number of trials for each condition, so within-group analysis cannot be used. A more traditional within-group analysis of the data is performed in section 4.2

4.1 Pooled Trial Results

The average response time of participants was 3.97 seconds (SD=1.98). A paired-samples t-test was conducted to compare the response time in winning trials and losing trials (excluding timeouts). There was a significant difference in the response time for winning responses (M=4.40, SD=1.81) and losing response (M=3.50, SD=1.88); $t(19718)=1.96$, $p < 0.05$. Again this was expected, as the task was designed so that responding more slowly would improve the player's chance of success.

Next we considered all trials in relation to the three experimental conditions. Overall, the 48 players completed 6,818 trials in the no sound condition, 6,661 trials in the constant sound condition and 6,694 trials in the increasing sound condition. In the no sound condition there were 2,794 (40.98%) correct responses, 3,830 (56.17%) incorrect responses and 194 (2.85%) timeouts. In the constant sound condition there were 2,717 (40.79%) correct responses, 3,773 (56.64%) incorrect responses and 171 (2.57%) timeouts. In the Increasing sound conditions there were 2,836 (42.37%) correct responses, 3,766 (56.26%) incorrect responses and 92 (1.37%) timeouts.

We designed the increasing sound as a temporal cue to reduce timeouts; it seemed to be effective with 2.85% of timeouts in the no sound condition, 2.57% in the constant sound and 1.37% in the increasing sound condition. A chi-square test of goodness-of-fit was performed to determine whether timeouts occurred equally across all sound conditions. Timeouts were not equally distributed in the experiments, $\chi^2(2, N=457) = 49.09$, $p < 0.05$. Unlike the timeouts, there were no significant differences in the number of correct responses $\chi^2(2, N=8,347) = 2.37$, $p=0.30$ or incorrect responses $\chi^2(2, N=11,369) = 0.15$, $p=0.93$ across the three conditions. This suggests that apart from the reduction in timeouts, there were no changes in players' hit rate (accuracy) when sound was included in the display.

However, using a one-way ANOVA we found a significant effect of sound on mean response time for all trials at the $p < 0.05$ level for the three conditions [$F(2, 19713) = 15.26$, $p = 0.00$]. Post hoc comparisons using the Tukey HSD test indicated that the mean response time for the no sound condition (M = 3.77, SD = 1.93) was significantly faster than both the constant sound condition (M = 3.93, SD

= 1.89) and the increasing sound condition (M = 3.94, SD = 1.88). The constant sound condition did not significantly differ from the increasing sound condition.

We then considered mean response time from winning trials separately from the response time for losing trials. Again there was a significant effect of sound on mean response time for both the winning trial data at the $p < 0.05$ level for the three conditions [$F(2, 8344) = 12.31$, $p = 0.00$] and the losing trial data [$F(2, 11366) = 5.03$, $p = 0.01$].

In terms of wins, post hoc comparisons indicated that the mean response time for the no sound condition (M=4.26, SD=1.85) was significantly faster than the constant sound condition (M=4.48, SD=1.81) and the increasing sound condition (M = 4.47, SD = 1.76). The constant sound condition did not significantly differ from the increasing sound condition. This overall pattern was consistent with the loss data where post hoc comparisons indicated that the mean time for the no sound condition (M=3.42, SD=1.91) was significantly faster than the constant sound condition (M=3.53, SD=1.85) and the increasing sound condition (M=3.54, SD=1.87). Again, the constant sound condition did not significantly differ from the increasing sound condition.

These results were pleasing from our design goals, as they provided further indication that (i) players avoided timeouts in the increasing sound condition, and at the same time (ii) were able to wait longer to respond than in the original no sound condition. What was most surprising about these results is that players also seemed to wait longer to respond in the constant sound condition, although this produced no significant reduction in timeouts. This constant sound condition was included as a control condition and was not expected to produce any variation in the way players performed the task.

4.2 Player by Player Results

After examining effects from pooled data, we also considered the player-by-player results. That is, the mean result for each player in each condition was calculated before analyzing these results in a one-way repeated-measures design. On average players completed 420.27 (SD=81.43) trials, 142.04 (SD=33.40) in the no sound condition, 138.77 (SD=27.69) in the constant sound condition and 139.46 (SD=31.54) in the increasing sound condition. The minimum number of trials completed by a player was 318. The maximum number of trials by a single player was 697.

Given the variation in number of trials that players completed we were concerned that our overall results could be biased, or over weighted, by individual performance. We therefore repeated our pooled-data analysis by using the averaged results for the 48 players. This entailed averaging all trials for each of the 48 individual players to find their averages and then finding the average of these 48 results.

First we considered the average number of winning trials for each player in the three conditions, no sound (M=58.21, SD=17.82), constant sound (M=56.60, SD=18.86) and increasing sound (M=59.08, SD=17.01). A repeated measures (within subjects) one-way ANOVA showed no significant difference between the number of wins in the three sound conditions, $F(2,47) = 0.70$, $p = .497$.

Next we considered the number of losses per player in the three conditions, no sound (M=79.79, SD=39.37), constant sound (M=78.60, SD=7.62) and increasing sound (M=78.46, SD=39.18). Again a repeated measures (within subjects) one-way ANOVA showed no significant difference, $F(2,47) = 0.06$, $p = .946$.

We then analyzed the number of timeouts in the three conditions, no sound ($M=4.33$, $SD=4.87$), constant sound ($M=3.54$, $SD=3.79$) and increasing sound ($M=1.65$, $SD=2.55$). In this case a significant difference was found between the number of timeouts in the three sound conditions, $F(2,47) = 13.36$, $p < .05$ (0.000). Post hoc comparisons with Bonferroni correction confirmed that increasing sound resulted in a significantly lower number of timeouts compared to the no sound condition ($p = .01$). There were no significant differences between either the no sound and constant sound or the constant sound and increasing sound conditions.

We then considered the average response time for all trials, per player ($n=48$), in the three conditions, no sound ($M=4.08$, $SD=1.21$), constant sound ($M=4.18$, $SD=1.14$) and increasing sound ($M=4.22$, $SD=1.21$). A repeated measures (within subjects) one-way ANOVA showed no significant difference between the response time in the three sound conditions, $F(2,47) = 0.51$, $p > .599$.

Next, we compared response times for all winning trials per player ($n=48$) in the three conditions, no sound ($M=4.22$, $SD=1.23$), constant sound ($M=4.40$, $SD=1.11$) and increasing sound ($M=4.40$, $SD=1.20$). No significant difference was found for the players average winning response time, $F(2,47) = 0.98$, $p = .379$.

Finally we considered just the response time for losing trials per player in the three conditions, no sound ($M=3.88$, $SD=1.16$), constant sound ($M=3.94$, $SD=1.08$) and increasing sound ($M=4.04$, $SD=1.19$). Again no significant difference was found for the players average losing response time, $F(2,47) = 0.72$, $p = .492$.

5. DISCUSSION

The key design goals of our multimodal display were validated in the experiment. First the decision-making task had a general success rate of about 41%, within the desired range of 40-60% (necessary to allow a sufficient number of both correct and incorrect trials [45]). However, there was considerable variation between players with four of the 48 players averaging below a 25% success rate and another four achieving higher than 70% of correct responses when all their trials were considered. Fig. 2 shows the distribution of win-percentage across players (the line marks the mean correct rate of 41%).

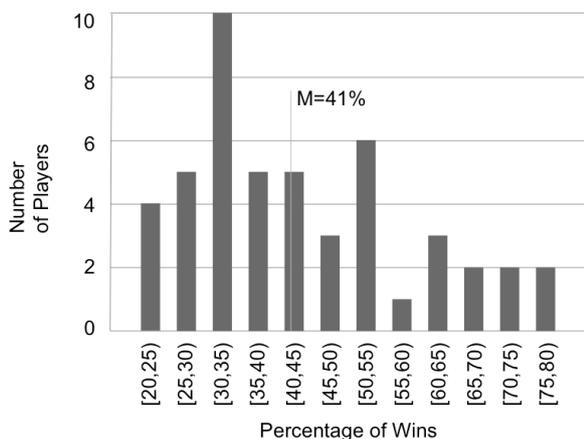


Figure 2: Percentage of wins for players

Second, the application of increasing sound cue to prevent timeouts was also successful. When we analysed all trials together, and then player-by-player, we found a significant reduction in timeouts in the increasing sound conditions compared to the no sound condition. This reduction in timeouts did not seem to come at the expense of players responding more quickly in the task. Indeed when we examined all trials together we found a significant increase in reaction time, so players actually slowed their response time in the increasing sound condition compared to the no sound condition. Overall, they also recorded more wins and fewer losses in the increasing sound condition than the no sound condition. This was consistent with the game design, as we expect the task to become easier as players wait longer.

These results are mitigated by the fact that when we compared the data by averaging player by player outcomes the difference in timeouts was still significant, however, the increase in response time and wins in the increasing sound condition was no longer evident. This suggests there was some bias introduced into the overall results by the performance of individuals in the experiment. Regardless, we can be confident that the player's performance in the multimodal task did not reduce to offset the reduction in timeouts.

The most surprising result was that that when we compared the overall trials we also found a significant difference between the response time and number of wins in the constant sound condition, compared to the no sound condition. Again this is mitigated by the fact that this significance was not evident when we compared the average results over the 48 players. Regardless this is a surprising result and worth further discussion.

The increasing sound signal was specifically designed to allow players to wait longer before deciding. The constant sound was introduced as a control, yet somewhat surprisingly players seemed to wait longer before responding in this condition as well. This could imply they also receive some timing information from this constant sound signal. One explanation for these results is that players have an internal mechanism for measuring time that is activated by a constant sound signal. Indeed such a model has been proposed that describes an internal pacemaker sending a regular series of pulses to some kind of counter mechanism [47].

In this model the internal clock mechanism can also be calibrated by external events. In the buckets game this calibration might occur using the visual updates or the game time outs. The model has also been used to consider how such a clock could impact of cognitive function such as the decision-making task [48]. Because of the need to track time in the constant condition this model predicts that we would see slightly longer response times, yet lower success rates. While this is the case in this experiment, these differences were not found to be significantly different in the constant sound condition compared to the increasing sound condition. Regardless, this interesting result is probably worth further study.

In terms of decision-making tasks in game designs, we have demonstrated the usefulness of gathering empirical data to test player performance on game-like tasks. We have also demonstrated how simple informative sound displays can provide useful information in a perceptual visual challenge.

6. CONCLUSION

Fast decision-making is often a critical task that underpins performance in computer games (and real life). Players in

competitive games are often faced with numerous, rapid decisions in which the final outcome is decided based on the player's current awareness of the situation. This requirement is also true in many business decisions.

In this experiment we compared the performance of players in a simple visual decision-making task and a multimodal version of the same eight-second task. Players must choose between one of four possible outcomes within the eight seconds before timing out. The longer players wait the easier the challenge becomes.

We augmented the visual only task by adding auditory feedback in the form of a sound slowly increasing in amplitude over the eight-second period. When compared with the visual only task, we found that there is a significant reduction in the number of timeouts experienced by players in the increasing sound display. This reduction does not seem to come at the expense of performance, as players seem to wait longer and make more correct responses in the increasing sound condition.

An interesting result that needs further validation is that players also seem to wait longer when a non-informative constant sound was added to the display. This result was difficult to validate as considerable player-to-player variation occurs in the task with average success rates ranging from 21% to 78%. Some players (n=15) performed at least 10% better in the increasing sound display while others (n=14) performed 10% worse with this display. For some (n=19) performance seems relatively unchanged between display modes.

Such variation in performance has previously been reported with multimodal displays [15, 38] and categorised as conflicting, complementary and redundant [49,50]. Where individuals perform worse with multimodal information, the display can be categorised as conflicting; where they perform better it can be described as complementary; and where there is no change in performance the display can be described as redundant. This variability in performance is also worth further study to see if it is consistent among individuals across other multimodal tasks, indicating a particular individual preference or also if it might be mitigated by training.

7. REFERENCES

- [1] Space Invaders (1978). Developer: Taito Corporation, Publisher: Taito Corporation.
- [2] Ekman I (2005) Meaningful noise: Understanding sound effects in computer games. In: Diddle A (ed) Proceedings of the 6th Digital Arts and Culture 2005 Conference : Digital experience: design, aesthetics, practice. IT University of Copenhagen, Denmark.
- [3] El-Nasr MS, Yan S (2006) Visual Attention in 3D games. In: Proceedings of the ACM SIGCHI international conference on advances in computer entertainment technology, ACM, New York, USA.
- [4] Gårdenfors D (2003) Designing sound-based computer games. *Digital Creativity* 14(2):111-114
- [5] Grimshaw M, Lindley CA, Nacke L (2008) Sound and Immersion in the First-Person Shooter: Mixed Measurement of the Player's Sonic Experience. In: Proceedings of 3rd Audio Mostly Conference, Interactive Institute, Piteå, Sweden.
- [6] Parker JR, Heerema J (2008) Audio Interaction in Computer Mediated Games. *International Journal of Computer Games Technology*, vol 2008, Article ID Article ID 178923, 8 pages, doi: 10.1155/2008/178923
- [7] Röber N, Masuch M (2005), Leaving the Screen, New Perspectives in Audio-only Gaming. In: Proceedings of the 11th ICAD: Limerick, Ireland, pp 92-98
- [8] Wolfson S, Case G (2000) The effects of sound and colour on responses to a computer game, *Interact Comp* 13(2):183-192
- [9] Valente L, Souza CSD, Feijó B (2008) An exploratory study on non-visual mobile phone interfaces for games. In: Proceedings of the VIII Brazilian Symposium on Human Factors in Computing Systems, Sociedade Brasileira de Computação, Brazil pp 31-39
- [10] Gaver W (1986) Auditory Icons: using sound in computer Interfaces. *Human-Computer Interaction* 2:167-177
- [11] Gaver W (1989) The SonicFinder: An interface that uses auditory icons, *Human-Computer Interaction*, 4(1):67-94
- [12] Barrass S (2003) Sonification Design Patterns. In: Proceedings of the 2003 International Conference on Auditory Display, Boston, MA, USA, pp 170-175
- [13] Blattner MM, Sumikawa DA, Greenberg RM (1989) Earcons and icons: their structure and common design principles. *Human Computer Interactive*, 4(1):11-14. doi: 10.1207/s15327051hci0401_1
- [14] Brewster SA, Wright PC, Edwards ADN (1993) An evaluation of earcons for use in auditory human-computer interfaces. In: Ashlund, Mullet, Henderson, Hollnagel & White (Eds.), Proceedings of INTERCHI'93, Amsterdam: ACM Press, Addison-Wesley. pp 222-227
- [15] Nesbitt KV, Barrass S (2004) Finding trading patterns in stock market data. *IEEE Comput Graph Appl* 24(5):45-55
- [16] Tan S, Baxa J, Spackman MP (2010) Effects of Built-in Audio versus Unrelated Background Music on Performance in an Adventure Role-Playing Game. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, 2(3):1-23. Doi: 10.4018/jgcms.2010070101
- [17] Walker BN, Kramer G (2006) Auditory Displays, Alarms, and Auditory Interfaces. In: W. Karwowski (Ed.), *International Encyclopedia of Ergonomics and Human Factors* (2nd ed.) New York: CRC Press. pp 1021-1025
- [18] Kramer G (1994) Auditory Display: Sonification, Audification, and Auditory Interfaces. Perseus Publishing.
- [19] Brewster SA, Walker VA (2000) Non-Visual Interfaces for Wearable Computers. In: Proceedings of IEE Workshop on Wearable Computing, London.
- [20] Adcock M, Barrass S (2004) Cultivating design patterns for auditory displays. In: ICAD Proceedings of the 10th International Conference on Auditory Display, Sydney, Australia.
- [21] Jørgensen K (2006) On the Functional Aspects of Computer Game Audio, In: Proceedings of the 2nd Audio Mostly Conference, Interactive Institute, Piteå, Sweden, pp. 48-52
- [22] Ramos D, Folmer E (2011) Supplemental Sonification of a Bingo Game. In: Proceedings of the 6th International Conference on Foundations of Digital Games, ACM, New York. pp 168-173
- [23] Barrass S (1996) EarBenders: Using Stories About Listening to Design Auditory Interfaces. In: Proceedings of the First Asia-Pacific Conference on Human Computer Interaction APCHI'96, Information Technology Institute, Singapore.

- [24] Nesbitt KV (2004) Comparing and Reusing Visualisation and Sonification Designs using the MS-Taxonomy, In Proceedings of International Community for Auditory Display, Sydney, Australia.
- [25] Alves V, Roque L (2010) A pattern language for sound design in games. In: Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound, Interactive Institute, Piteå, Sweden.
- [26] Ng P, Nesbitt K, (2013) Informative Sound Design in Video Games. In: Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death, Melbourne, Australia.
- [27] Gaver W, Smith R, O'Shea T (1991) Effective sounds in complex systems: the ARKola simulation. In: Proceedings of CHI '91. ACM, New York, pp 85-90
- [28] Polotti, P., & Lemaitre, G. (2013). Rhetorical Strategies for Sound Design and Auditory Display: A Case Study, *Int J Des* 7(2):67-82
- [29] Something Else (2013) Papa Sangre (iOS Software), London.
- [30] Square Enix (2014) Thief (PC Software), Shinjuku, Tokyo, Japan.
- [31] Blizzard Entertainment (2012) World of Warcraft: Mists of Pandaria (PC/Mac Software). Irvine, California, USA. Activision Blizzard Inc
- [32] Grimshaw M, Schott G (2008) A Conceptual Framework for the Analysis of First-Person Shooter Audio and its Potential Use for Game Engines. *International Journal of Computer Games Technology* Volume 2008 (2008), Article ID 720280, 7 pages
<http://dx.doi.org/10.1155/2008/720280>
- [33] Friberg J (2004) Audio games: New perspectives on game audio. In: Proceedings of the 2004 ACM SIGCHI Advances in Computer Entertainment Technology, pp148-154
- [34] Stockburger A (2003) The Game environment from an auditive perspective. In: Copier M, Raessens J (eds) *Level Up: Digital Games Research Conference*. Digital Games Research Association, Utrecht University Press.
- [35] Valve Corporation (2013) Dota 2 (PC/Mac Software). Bellevue, Washington, USA: Valve, LLC.
- [36] Ng P, Nesbitt K, Blackmore K (2015) Sound improves player performance in a multiplayer online battle arena game. In: *Artificial Life and Computational Intelligence: Lecture Notes in Computer Science*. Springer International Publishing, pp166-174
- [37] Smith DR, Walker BN (2005). Effects of auditory context cues and training on performance of a point estimation sonification task. *Appl Cogn Psychol* 19(8):1065-1087
- [38] Nesbitt KV, Hoskens I (2008) Multi-sensory game interface improves player satisfaction but not performance. In: Proceedings of the Ninth Conference on Australasian User Interface, Volume 76, Australian Computer Society, Australia, pp13-18
- [39] Stanton NA, Edworthy J (1999) Auditory warnings and displays: An overview. In N. A. Stanton & J. Edworthy (eds), *Human factors in auditory warnings*,. Aldershot, UK: Ashgate, pp 3–30
- [40] Graham R (1999) Use of auditory icons as emergency warnings: evaluation within a vehicle collision avoidance application, *Ergonomics*, 42(9): 1233-1248
- [41] Patterson RD (1982) Guidelines for auditory warning systems on civil aircraft. CAA Paper 82017, Civil Aviation Authority, London.
- [42] Morgan AL, Brandt JF (1989) An auditory Stroop effect for pitch, loudness, and time. *Brain Lang* 36(4):592-603
- [43] Kahneman D (1973) *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.
- [44] Townsend JT, Eidels A (2011) Workload capacity spaces: A unified methodology for response time measures of efficiency as workload is varied. *Psychon Bull Rev* 18(4):659-681
- [45] Williams PG, Nesbitt KV, Eidels A, Elliott DJ, (2011) Balancing risk and reward to develop an optimal hot-hand game. *Game Studies*, 11 online (2011)
- [46] Adams E (2010) *Fundamentals of Game Design*. Pearson Education.
- [47] Zakay D, Block RA (1995) An attentional-gate model of prospective time estimation. In Richelle M, Keyser VD, d'Ydewalle G, Vandierendonck A (eds), *Time and the dynamic control of behavior*, Liège, Belgium: Universite de Liege, pp 167-178
- [48] Block RA (2003) Psychological timing without a timer: The roles of attention and memory. In Helfrich H (ed), *Time and mind II: Information processing perspectives*. Göttingen, Germany: Hogrefe & Huber, pp 41-59
- [49] McGee, MR, Gray PD and Brewster SA (2000) The Effective Combination of Haptic and Auditory Textural Information. In: Brewster SA, Murray-Smith R(eds) *Proceedings of the First International Workshop on Haptic Human-Computer Interaction*. Springer-Verlag, London, UK, pp 118-126
- [50] Pao LY, Lawrence DA (1998) Synergistic Visual/Haptic Computer Interfaces. In: Proceedings of Japan/USE/Vietnam Workshop on Research and Education in Systems, Computation and Control Engineering, pp 155-162

ORAL PAPERS

*Aesthetics, Philosophy, and
Culture of Auditory Displays*

THE AESTHETICS OF CAUSALITY: A descriptive account into Ecological Performativity

Teresa Marie Connors

The University of Waikato Conservatorium of Music
Hamilton, New Zealand,
tmconnor@waikaro.ac.nz

ABSTRACT

In this paper, I offer a perspective into a creative research practice I have come to term as *Ecological Performativity*. This practice has evolved from a number of non-linear audiovisual installations that are intrinsically linked to geographical and everyday phenomena. The project is situated in ecological discourse that seeks to explore conditions and methods of co-creative processes derived from an intensive data-gathering procedure and immersion within the respective environments. Through research the techniques explored include computer vision, data sonification, live convolution and improvisation as a means to engage the agency of material and thus construct non-linear audiovisual installations. To contextualize this research, I have recently reoriented my practice within recent critical, theoretical, and philosophical discourses emerging in the humanities, sciences and social sciences generally referred to as ‘the nonhuman turn’. These trends currently provide a reassessment of the assumptions that have defined our understanding of the geo-conjunctures that make up life on earth and, as such, challenge the long-standing narrative of human exceptionalism. It is out of this reorientation that the practice of *Ecological Performativity* has evolved.

1. INTRODUCTION

“As techno-science increasingly reaches into every aspect of life, formerly fast held distinctions between the inert and the active, the human and non-human and life and matter are cracking.” [1]

In April 2011 researchers from the natural and social sciences, the humanities, and a variety of creative practitioners gathered in Rotterdam, The Netherlands. Under the title *The Vibrancy Effect: An Anti-Disciplinary Meeting*, the focus was to discuss and “explore the aesthetic-political-technical-ethical effects of vibrant matter [1].” The term *Vibrancy*, here, is in direct reference to Jane Bennett’s concept of vibrant materiality or “thing-power” that, as Bennett claims, attempts to give voice to the energetic vitality intrinsic to matter and the active, earthy, and complex entanglements of the human and nonhuman [2]. At this meeting, participants presented their unique thoughts, approaches, and concerns for considering vibrant materiality, or, what sociologist of science Andrew Pickering calls material agency—“the material that comes at us from outside the human realm [3].”

 This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

Jump ahead to May 2012 and a gathering of scholars at the University of Wisconsin, Milwaukee under the rubric of *The Nonhuman Turn*, and similar to *The Vibrancy Effect*, the discourses to emerge explored the agency of human and nonhuman bodies. Bennett describes these discourses as an attempt: “to find new techniques, in speech and art and mood, to disclose the participation of nonhumans in “our” world [4].” Erin Manning adds that the “art of participation does not find its conduit solely in the human. [...] Art also does its work without human intervention, activating fields of relation that are environmental or ecological in scales of intermixings that may include the human but don’t depend on it. How to categorize as human or nonhuman the exuberance of an effect of light, the way the air moves through a space, or the way one artwork catches another in its movement of thought [5].”

Broadly speaking, the nonhuman can refer to objects such as “climate change, drought, and famine; to biotechnology, intellectual property, and privacy; to genocide, terrorism, and war [6].” Such wide-ranging perspectives on what constitutes a nonhuman are, as Salter claims above, a cracking of distinctions [1]. But given the many concerns arising in the twenty-first century, in the time of ecological emergency, this turn towards the nonhuman has particular relevance, as Timothy Morton suggests, “to exit modernity [7].”

2. ISSUES OF AGENCY

“Thinking issues of agency through the experiential encounter with the ‘stuff of the world’ encourage a radically different vision of the world—dynamic, temporally emergent, contingent, and performative [8].”

Thinking in terms of agency and performativity is nothing inordinately new, and in Western thought has evolved from a variety of philosophical, scientific and artistic research that took place over the last century [6, 9, 10]. Of late however, the reinvestigation into these notions is, as Salter suggests... “encouraging a radically different vision of the world [8].” From Karen Barad’s “intra-action” [11] and Pickering’s “dance of agency” [3] to Bennett’s “thing-power” [2], Morton’s notion of the “hyperobject” [12] and Tim Ingold’s “meshwork” [13], a reconceptualization is taking place which challenges the fundamental understanding of the interdependence and interconnectedness of all life and matter and the notion of human exceptionalism.

As these thinkers grapple with the notion of agency in human and nonhuman bodies, a host of ecological, social, cultural, and political observations and concerns are being raised and challenged; the urgency of which is energized by



Figure 1: The work *Motion Parallax* (1998) was the first large-scale audiovisual collaboration between Andrew Denton and I and was created using field recordings captured on a cross-Canada trip from Tofino, British Columbia to Cape Spear, Newfoundland. This photo was taken on the last day of shooting on top of Signal Hill historical site located in St. John's NFLD.

what has now been embraced as the Anthropocene¹—the epoch in which the effects of fossil-fuel-burning humans have fundamentally altered the earth's geological composition.

For my own part, I was drawn to these discussions as a means to answer my own question: What does making art from the lived and experiential encounters with the 'stuff of the world' do? In other words, what is the purpose of an art form set in the context of time and place? Since the 1990s my creative practice has revolved around the exploration of the day-to-day situated encounters in the real world (Figure 1). These works were deeply embedded in time and place and explored the impact that human activities have had on the landscape. Since then, this practice has evolved from a fixed-media format to one that explores non-linear systems. This was motivated by a growing curiosity to explore a more dynamic aesthetic that could include agential properties available in the respective environments. Field recordings have taken place throughout North America, New Zealand, and Australia resulting in a catalogue of audiovisual works that are intrinsically linked to geographical factors and everyday phenomena.

But what is it about these experiential encounters that have held my curiosity? What significance does it have on my mode of artist practice, and how does this practice motivate the conditions in which creative possibilities are activated, assembled, and processed? More specifically, what would motivate my collaborators and I to venture on field recordings that would place us in Death Valley in 53°C heat, the polluted wastelands of the Salton Sea in Southern California, the crowded sidewalks of Los Angeles, and the tourist-filled paved pedestrian trails in Sequoia National Park? The answer to these questions, I believe, resides in practice.

By reorienting my creative practice with these different thinkers and writers, the process that I refer to as *Ecological Performativity* has evolved. Central to this idea is the fundamental questions: What tendencies emerge in the

*making-doing-thinking*² of creative practice when human and nonhuman agency is located as a co-creative device? What capacities do these tendencies have on the creative process and how do they affect the resulting artefacts? Can this encourage an attunement to the reality of the coexistence of all things on Earth? And if so, as a creative practitioner, what, then, is my response and response-ability?

3. ECOLOGICAL PERFORMATIVITY

“The world is an open process of mattering through which mattering itself acquires meaning and form through the realization of different agential possibilities [11].”

Open processes and different agential possibilities are central to the creative practice of *Ecological Performativity*. *Ecological* is located within the philosophical provocations of Brian Massumi and Erin Manning as being that of a relational experience: “Organisms-that-person agitate in the mix, but always in a witness of environment: a becoming ecology of practices [16].” Thus, this practice considers emergence and material agency as co-creative apparatuses. Accordingly, *Performativity* draws specifically upon Andrew Pickering's notion of the “dance of agency [3].” Here, agency and performative are entwined in what Pickering posits as the performative idiom [18]. This is Pickering's attempt to move away from the idea that agency is specific only to humans, or to what he refers to as “human exceptionalism [18].” He suggests that the world, in all its heterogeneous multiplicity, is full of agency and processes of emergence. By exploring these processes and performative relationships between things, including those beyond the human realm, Pickering suggests that we invite the “possibility of a non-modern stance of revealing rather than enframing which, in turn, invites open-ended extensions [19].”

Similar to other ecologically-grounded creative practices,³ *Ecological Performativity* explores the relationships of environment, material, and process, and are derived from an intensive data-gathering procedure and immersion within the respective environments. Each work begins in a matter-of-fact manner (making sure all batteries are charged etc.). However, the effect and affect these environments have on my collaborators and I become an operative agent. Bennett's discourse on “thing-power” surmises that: “Earthy bodies, of various but always finite durations, affect and are affected by one another. And they form noisy systems or temporary working assemblages that are, as much as any individuated thing, loci of effectivity and allure [4].” There is a causal dimension that, as Morton argues, is “wholly an *aesthetic* phenomenon [7].” Of this my long time collaborator Andrew Denton writes: “Once time is taken to absorb [the location], I attempt to record material that communicates my sensations and experiences of being there [25].” He reflects that by “letting go of a need to understand, comprehend, and categorize [...] the intensity of the making-feeling-thinking [could] take over in the moment of capture, leave[ing] the reflection

¹ Ecologist Eugene Stoermer and atmospheric chemist Paul Crutzen coined the term Anthropocene as a means to name the geological transformations that have occurred in the oceans (acidification) and coral reefs (bleaching) [14].

² Donna Haraway describes this figure of practice as a back and forth passing of patterns, similar to string figure games [15].

³ Terms used to denote other ecologically-grounded creative practices include *ecomposition*, *sonic ecologies*, *EcoSon*, *ecosystems*, and *audible ecosystems* [20, 21, 22, 23, 24].

and reinterpretation for a later distanced encounter with the material during post-production [25].”

The post-production exploration of materials is done in part through the development of specifically designed computational systems. These systems vary in construction and are intrinsically linked to the collected location data of audio field recordings, moving images and photos, as well as weather, meteorological, and environmental data gleaned from these situated encounters. Through research the techniques include computer vision processes, data sonification, live convolution, and improvisation as a mean to engage the agency of material and thus construct the non-linear audiovisual installations. Data sonification has been enlisted in a variety of ways as a co-creative apparatus. This includes transcoding environmental data (numbers) into triggering agents on audio volume controls, audio delay units, audio effects units, and as a means to construct algorithms to run the overall architecture of the non-linear installations. In addition, the compositional technique of sonification has been employed on still and moving images for the manipulation of audio field-recordings and the generation of sonic material through motiongrams [26]. What emerges does so in an iterative, non-deterministic manner, which affords an open-ended interaction within an “ecology of practice [27].”

Subsequently, this ecology of practice has come to involve the recording of live musical improvisations in response to the developed system. This has become an important component of *Ecological Performativity*,—which is within the iterative developments of these systems out of the material gathered; an acoustic musician is then invited into the process to respond improvisationally to the material. Recordings have taken place in live multimedia concert improvisations, studio settings, and the respective environments. What this provides is a cumulative database that in turn folds back into the final system. Motivated by the desire to explore non-linear systems, the installation platform provides a space where the constraints of beginnings, middles, and ends are eliminated. The artwork can then exist as a transformative apparatus.

Operating in this discursive register, from the core of creative-research, provides a platform for experimentation-as-process that contributes to new ways of thinking by insisting that every practice is a knowledge that can speak and act through the differences and emerging possibilities.

4. Dark Ecology, the Sonic Potentials of Data and the Salton Sea

“With dark ecology, we can explore all kinds of art forms as ecological: not just ones that are about lions and mountains [...]. The ecological thought includes negativity and irony, ugliness and horror [28].”

Anybody who has ventured into the writings of Timothy Morton will be familiar with the complexity of ideas spun on every page. From his book *Ecology Without Nature* and ideas of the “hyperobject” to his dark ecological thoughts, Morton’s philosophical ponderings purpose a way of thinking and being (of which he considers thinking, in and of itself, an ecological event) that embraces ambiguity, uncertainty and the uncanniness of the entangled mesh. Morton is a strong advocate for art, philosophy, and music stating that: “...art forms have something to tell us about the environment, because they can make us question reality [28].”

“Thus the art in the time of hyperobjects explores the uncanniness of beings, the uniqueness of beings, the irony and interrelationships between beings, and the ironic secondariness of the intermeshing between beings [29].”

When considering Morton’s idea of the hyperobject, that of, “agents or objects so massively distributed in time and space as to transcend localization, such as the biosphere, global warming, or the sum of all the whirring machinery of capitalism [12],” the creative practice of making works from field recordings and data becomes multifaceted. When one reflects on the interwoven interactions that occur in any given encounter; between what is seen and unseen, heard and inaudible to our human ears, the complexity of the mesh is immense. For Morton, “the mesh” substitutes words such as interdependence and interconnectedness [28]. For Tim Ingold, the mesh is a metaphor for the relational interwoven lines of lived experience [13]. In my creative research, thinking in terms of the mesh underpins the practice of *Ecological Performativity*. By engaging in a non-deterministic way with what is present in any given environment, “the poetic potential of locational data has the capacity to draw you to the multiplicity and complexity of the content [25].”

This practice was put into play when collaborator Andrew Denton and I recently embarked on an audiovisual collection process throughout the Southwestern drought regions of the United States. This three-week field recording session involved many extreme locations including Bombay Beach on the Salton Sea. Parking our vehicle and venturing into this environment, the odour itself stopped us in our tracks. The shoreline was littered with dead fish and birds and human objects in varying stages of decline, all of which were covered with a dusty white mixture of salt and dried thermal mud. This environment is the result of early 20th century weather systems and ensuing human activities.

In 1905, when the Colorado River swelled and breached its banks, the water ran into the Salton Sink, a geographical region 220 feet below sea level. After two years of continuous flow, a 15-by-35 mile lake formed that became known as the Salton Sea [30]. Taking advantage of California’s newest and now largest lake, the Salton Sea became a favorite getaway spot for nearby Los Angeles and San Diego residents. During the 1950s and ‘60s, Bombay Beach, which is located on the lake’s eastern side, became a prosperous resort town filled with sunbathers, water-skiers, and yacht club parties. During the 1970s, however, it became apparent that the ecosystem of the Salton Sea was quickly deteriorating. With no drainage outlet and little to no annual rainfall, the inflow of industrial pollutants and untreated sewage began to increase the lake’s salient level and caused the water to deoxygenate. What had become an angler’s well-stocked paradise quickly transformed into a rotten layer of dead fish and birds [30, 31].

The indexical signs of the human and nonhuman now litter Bombay Beach, which has been described as “the most depressing place in California [32].” Once Denton and I had adjusted to the initial shock of this environment, we proceeded to record these indexical signs. Denton finds himself visually drawn to the monotonous awe of water reflections, birds in flight and the seemingly endless convoy of cattle trains that shimmer in the desert heat, on their journey from Mexico (Figure 2). For my part, I became transfixed with the numerous objects scattered throughout this landscape: rusty metal objects sticking out of the ground, wooden refuse from



Figure 2: The Salton Sea. Photo Andrew Denton

dilapidated buildings, sections of concrete slab, plastic bags entangled and flapping in dead bushes, and a lone broken piano (Figure 3). Using contact microphones, I recorded the sonic textures and tones by tapping, plucking and playing these objects. Equally striking was the sound resounding at the waters edge. Primarily comprised of crushed fish and bird bones, the sonic quality activated by wave and human footsteps has a sharp percussive high-pitched resonance. I captured this using a hydrophone.

It was during this field recording I posed the question to my collaborator: what does making art from these lived and experiential encounters in the world do? Beyond technical considerations, this research attends to the frailty, vulnerability and the performative substance of time and place. Morton surmises “to be located “in” space or “in” time is already to have been caught in a web of relations [7].” From a sonic arts practice Kim-Cohen suggests that: “Every work of art is a response to the conditions within which it is produced and received [. . .], the assumptions and problems inherent to its time and place [33].” Or, perhaps, by choosing to engage with the negativity, irony and ugliness of these environments—Morton’s dark ecology, the capacity to recalibrate the world through our practice is opened by drawing out the evocative and emotional that, in turn, provides the opportunity to see, hear and be in the world differently [25]. The artwork becomes an apparatus of change.



Figure 3: Piano on Bombay Beach.

5. ACKNOWLEDGMENT

This research was made possible in part by the University of Waikato International PhD Scholarship.

6. REFERENCES

- [1] C. Salter, “The Vibrancy Effect: An Anti-Disciplinary Expert Meeting” in *The Vibrancy Effect*, eds. C. Salter, H. Smoak and M. V. Dartel, Rotterdam: NAI Publishing, 2011, pp.16–23.
- [2] J. Bennett, *Vibrant Matter: A Political Ecology of Things*, Durham & London: Duke University Press, 2010.
- [3] A. Pickering, *The Mangle of Practice: Time, Agency, and Science*, Chicago: University of Chicago Press, 1995.
- [4] J. Bennett, “Systems and Things. On Vital Materialism and Object-Oriented Philosophy,” in *The Nonhuman Turn*, ed. R. Grusin, Minneapolis: University of Minnesota Press, 2015, pp. 223–239.
- [5] E. Manning, “Artfulness.” In *The Nonhuman Turn*, ed. R. Grusin, Minneapolis: University of Minnesota Press, 2015, pp. 45–79.
- [6] Richard Grusin ed., *The Nonhuman Turn*, Minneapolis: University of Minnesota Press, 2015.
- [7] T. Morton, *Realist Magic: Objects, Ontology, Causality*, Ann Arbor: Open Humanities Press, 2013.
- [8] C. Salter, *Alien Agency*, Cambridge, MA: The MIT Press, 2015.
- [9] C. Salter, *Entangled: Technology and the Transformation of Performance*, Cambridge, MA: The MIT Press, 2010.
- [10] F. Capra and P. L. Luisi, *The Systems View of Life: A Unifying Vision*, Cambridge, MA: Cambridge University Press, 2014.
- [11] K. Barad, *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*, Durham & London: Duke University Press, 2007.
- [12] T. Morton, *Hyperobjects: Philosophy and Ecology after the End of the World*, Minneapolis, MN: The University of Minnesota Press, 2013.
- [13] T. Ingold, *Essays on Movement, Knowledge and Description*. London: Routledge, 2011.
- [14] P. Crutzen and E. Stoermer, The “Anthropocene”, *Global Change Newsletter* 41, May 2000, pp. 17–18.
- [15] D. Haraway, “Staying with the Trouble: Sympoiesis, String Figures, Multispecies Muddles,” Keynote at University of Alberta: 2014.
- [16] E. Manning and B. Massumi, *Thought in the Act: Passages in the Ecology of Experience*, Minneapolis, MN: University of Minnesota Press, 2014.
- [17] A. Pickering, “Being in an Environment: A Performative Perspective.” *Natures Sciences Sociétés* 21, 2013, pp. 77–83.
- [18] A. Pickering, “Art and Agency,” in *The Vibrancy Effect*, eds. C. Salter, H. Smoak, and M. V. Dartel (Rotterdam: NAI Publishing, 2012), pp. 28–32.

- [19] A Pickering, *The Cybernetic Brain*, Chicago: University of Chicago Press, 2010.
- [20] D. Keller and A. Capasso, "New Concepts and Techniques in Eco-composition," *Organised Sound*, Vol. 11, No. 01, 2006, pp. 55–62.
- [21] L. Barclay, "Sonic Ecologies: Environmental Electroacoustic Music Composition in Cultural Immersion," PhD, Griffith University, 2013.
- [22] M. Burtner, "Ecosono: Adventures in Interactive Ecoacoustics in the World," *Organised Sound*, Vol. 16, No. 03, 2011, pp. 234–44.
- [23] A. Di Scipio, "Listening to Yourself through the Otherself: On *Background Noise Study* and other works," *Organised Sound*, Vol. 16, No. 02, 2011, pp. 97–108.
- [24] S. Waters, "Performance Ecosystems: Ecological Approaches to Musical Interaction." In *Electroacoustic Music Studies Network*. De Montfort, Leicester, 2007.
- [25] A. Denton, *Affective Moving Image and the Anthropocene*, Monash University: PhD dissertation, 2016.
- [26] A. R. Jensenius, "Motion-Sound Interaction Using Sonification Based on Motiongrams." *ACHI 2012: The Fifth International Conference on Advances in Computer-Human Interactions*, 2012, 170–175.
- [27] I. Stengers, "An Ecology of Practices." *Cultural Studies Review* 11, no. 1, 2005.
- [28] T. Morton, *The Ecological Thought*, Cambridge and London: Harvard University Press, 2010.
- [29] T. Morton, "Dawn of the Hyperobjects", from <http://www.youtube.com/watch?v=zxpPJ16D1cY>
- [30] The Salton Sea. "A Brief Description of its Current Conditions, and Potential Remediation Projects", from <http://www.sci.sdsu.edu/salton/Salton%20Sea%20Description.html>
- [31] T. Paiva, "Lost America: The Salton Sea", from <http://lostamerica.com/photo-items/the-salton-sea/>
- [32] R. Riggs, "Strange Geographies: Bombay Beach", from <http://mentalfloss.com/article/24260/strange-geographies-bombay-beach>
- [33] S. Kim-Cohen, *Against Ambience*. New York, NY: Bloomsbury Publishing. Kindle Electronic Edition, 2013.

SONIFYING FOR PUBLIC ENGAGEMENT: A CONTEXT-BASED MODEL FOR SONIFYING AIR POLLUTION DATA

Marc St Pierre

Simon Fraser University
mkstpier@sfu.ca

Milena Droumeva

Assistant Professor
Simon Fraser University
mvdroume@sfu.ca

ABSTRACT

In this paper we report on a unique and contextually-sensitive approach to sonification of a subset of climate data: urban air pollution for four Canadian cities. Similarly to other data-driven models for sonification and auditory display, this model details an approach to data parameter mappings, however we specifically consider the context of a public engagement initiative and a reception by an ‘everyday’ listener, which informs our design. Further, we present an innovative model for FM index-driven sonification that rests on the notion of ‘harmonic identities’ for each air pollution data parameter sonified, allowing us to sonify more datasets in a perceptually ‘economic’ way. Finally, we briefly discuss usability and design implications and outline future work.

1. INTRODUCTION

Sonification has, over the last two decades, established itself as a growing modality for conveying information and an increasingly legitimized tool, useful in many different circumstances. It can integrate successfully into workflows for control room monitoring, scientific data exploration, and even physiotherapeutic treatment [1]. Arguably, in any circumstance where there is (ongoing or continual) information that requires perception and/or action, sonification can play a part in its communication, either alone or as a complement to visual displays.

Within the broader collection of ICAD literature, there have been numerous advances in scientific sonification and accompanying issues of auditory stream perception, aesthetics and usability. However, there has been relatively little attention as yet given to circumstances in which scientific data is communicated in the public sphere. That is, when sonifications have been designed to engage a mass audience through the audible representation of scientific data, with the goal of raising awareness and allowing an everyday listener access to challenging and often ‘cold’ scientific information. As with other contexts, this form comes with its own design ideals, aesthetic affordances, and functional constraints to consider.

Recent sonification projects, which include the discovery of

the Higgs Boson [2], Rosetta Comet [3], and gravitational waves [4], have received widespread public appeal. They are among the more salient examples of aesthetically driven mappings that are meant to engage the public in a particular way, with, of course, the openly political motive of trying to foster a public interest in scientific discovery. And also, perhaps, with the underlying political motive of trying to justify funding for costly programs. It is important to acknowledge this, because in circumstances where sonification is used as a ‘public relations’ tool, there is always a message underneath which serves as the guiding principle for its design: a note made most famously in Alexandra Supper’s discussion of the legitimacy of sonification [5].

With sonification’s increasing presence in the public domain, one of the most pressing implications is the need to rethink perceptual mappings and training as part of ongoing design developments in sonification to involve an increasingly non-specialized audience that is more used to listening to music than data. The visual equivalent of this would be the rise and popularity of conveying popular scientific/quantitative data through accessible and visually appealing infographics and data visualizations. Here is where sonification, akin to visualization, provides a unique entrypoint into understanding complex and often ‘invisible’ scientific, but also potentially social, cultural, and political processes. For this reason, it is useful to explore the unique ways in which sonification operates in the context of public relations. This paper will chart a practice-based research approach to a mapping strategy meant to engage the public with issues surrounding air quality in cities across Canada. Sonification, in this case, serves as a tool to communicate a broader message of mitigating emissions to the public, for citizen health, as well as for climate change activism.

2. DATA DOMAIN AND DESIGN PRINCIPLES

Climate and weather data have been a popular resource for sonification research in the past, possessing time-varying and dynamic characteristics, ripe with temporal patterns for the ear to perceive. Previous explorations of climate data have often cited these as reasons for developing new sonification methods [6]. In more recent cases, such as Goudarzi’s user-centered approach [7], the stated end-goal of her design methodology is to develop mapping strategies that serve the research needs of climate scientists. The data used is complex and multidimensional, requiring many perceptually diverse mappings sounding in parallel. Scientists are also required to learn the mappings over an extended period of time.



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

If we consider public presentation as the context of reception, the requirements are different. Given time constraints, sonifications often need to be simplified to drive home one or two highly important and impactful points. Fewer mapping strategies are used, and they may be more aesthetically driven. The message behind the sonification needs to be self-evident, and would not require specialized ear training to perceive.

In this project air quality was chosen among the broad spectrum of climate data based on its availability, and its recurring presence in the media's continuous coverage of issues surrounding urban environments. Because of this coverage, a public is primed to understand the connections between air pollution, their health, and the contributions of these emissions to the climate change. A sonification of air quality data, then, serves as an access point for the public to further understand the urgency of our changing environmental condition. It is this underlying outcome that guides our design.

Sonifications for four Canadian cities, Vancouver, Edmonton, Toronto, and Sarnia during the year of 2014 were analyzed. Data is retrieved from provincial websites for British Columbia, Alberta, and Ontario respectively [8], [9], [10]. Within each dataset, five metrics of air pollution are sonified. These metrics are Ozone (O₃), Particulate Matter (PM_{2.5}), Nitrogen Dioxide (NO₂), Carbon Monoxide (CO), and Sulphur Dioxide (SO₂). Each is sampled hourly.

3. A REVIEW OF PREVIOUS METHODS: THE AUDITORY GRAPH

Of the three conventional methods of sonification, audification, model-based, and PMSon (parameter mapping sonification), the latter is most widely used for small to large datasets with multiple data properties or dimensions. deCampo's Design Space Map [11] proves a useful starting point when designing sonifications, however, there appears to be a canonical propensity, in cases where the design space map is applied, to view a one-to-one data-to-pitch mapping as the archetype for an effective sonification, i.e. the auditory graph [12]. It is true that the ear's ability to resolve frequency deviations - the JND - is apt compared to other mapping strategies within certain frequency ranges, however this often results in sonification designs which can be aesthetically challenging to listen to for extended periods of time [13], as well as semantically limited in terms of what pitch change denotes in terms of meaning. Furthermore, if more than one data dimension is sonified, resulting pitch streams must be kept sufficiently separate in frequency so they don't overlap and risk confusion as to which data parameter is changing. These remarks are ultimately situated within one of two typical 'requirements' of sonification listening: 'exact value perception' or a high degree of correct value identification, versus the general perception of significant shifts in a continuous dataset. In the case of raising awareness, rather than scientific identification of data, this challenges perceptual considerations when designing sonifications. Conventional auditory graphs may seem appropriate for air quality data given an exploratory or purely scientific context. However, under the aesthetic and functional requirements for public relations outlined in the previous section, an alternative interpretation of the auditory graph is used, which we outline in the following sections.

4. TIME SCALING AND PARAMETER MAPPING

Before the details of the design are revealed, it is encouraged that you listen to the sonifications for both Vancouver and Sarnia first. They are publicly available at <https://soundcloud.com/marcstpierre> [14]. We contend that this is meant to demonstrate an implementation which does not require detailed understanding, or prolonged training, in order to holistically perceive which city is more polluted. But it is ultimately in the hands of the listener to support this argument or critique it.

To begin, there are a few stated goals of a sonification such as this one. The first is to holistically map more pollutants in the atmosphere to noise, which takes advantage of its negative connotations. This is an aesthetic decision to help communicate the detrimental effects of emissions to the environment. The second is to be able to differentiate and compare the relative levels of pollutants in each city. The third is to do this with temporal effectiveness. In short, the mapping requires enough aesthetic and functional flexibility to yield five differentiable streams that do not interfere with each other or cause auditory fatigue.

The first and most crucial design consideration is time-scaling. For this sonification, each data value representing an hour of real time, is reduced to 0.2 seconds of sonification time. A 12-hour day is therefore represented in 2.4 seconds, or a year in roughly half an hour. Daily emission patterns in the data are easily perceivable at this scale and rest comfortably within the echoic memory range [11]. With regard to streaming, several fundamental ideas from psychoacoustics and sound synthesis provide a framework for the final design. The first is Bregman's seminal stream segregation grouping cue, which states that "when two concurrent sounds have different fundamental frequencies, the brain can use the fact that the harmonics that comprise each sound will be a whole number multiple of the fundamental." [15]

Considering the design criteria for perceivability, the nature of our dataset, and these perceptual attributes, we chose frequency modulation (FM PMSon) as the most suitable mapping strategy based on its ability to efficiently generate rich harmonic spectra in relation to a fundamental. Frequency modulation is a waveshaping synthesis technique, which uses one waveform to modulate the frequency of another waveform. One wave is called the carrier, the other the modulator. When modulation occurs at frequencies above the audio rate, "sinusoidal sidebands are created at frequencies equal to the carrier frequency plus and minus integer multiples of the modulator frequency". The index of modulation is a ratio that indicates the amount of deviation from the carrier signal, and this value determines the number of sidebands on either side of the carrier, resulting a subjective experience of 'noise' [16].

4.1. Creating "Harmonic Identities" Using Stream Segregation

The five metrics are sonified in parallel using a simple mapping strategy with a positive polarity (larger data values equal to larger acoustic values). 4 of the pollutants, CO, O₃, SO₂, and NO₂, were scaled and mapped to the modulation

index of an FM synth in SuperCollider. Each pollutant was given its own fundamental frequency and carrier-to-modulator ratio in different regions of the auditory spectrum. This mapping possesses multiple affordances: the first, given different c/m ratios, each pollutant occupies a fixed fundamental frequency in the auditory spectrum which remains unchanged throughout the sonification. This means that once the pollutant positions are known, it becomes very easy to identify which one is changing at any given time. Furthermore, because of the different ratios, each pollutant also possesses a distinct array of sidebands that are harmonic multiples of the modulator, creating unique timbral structures, or *harmonic identities*, for each pollutant. Importantly, this is what allows for the superimposition of streams on top of each other without perceptual and cognitive occlusion. As the modulation index goes up for each of the pollutants to the point of overlap between the sidebands, the streams remain differentiable, based on the consistent harmonic relationship to an unchanging and unique fundamental. In the same way that you are perceptually able to ‘parse’ out the sounds of individual instruments playing together in an orchestra, pollutants which are sonified using FM can be segregated (to a degree) based on their harmonic identities. As pollutant levels go up, the sidebands increase, becoming, fuller, brighter, and ultimately noisy. The effects of this mapping are evidenced by comparing a relatively less polluted city like Vancouver to an industrial town like Sarnia, where one sounds much more distorted and aesthetically ‘harsh’ than the other.

4.2. Particulate Matter (PM_{2.5})

The fifth pollutant, PM_{2.5}, is measured differently than the other chemicals and therefore receives a different mapping. Particulate matter is commonly cited as the most dangerous air pollutant among those measured [17]. Instead of a single chemical, PM_{2.5} is composed of multiple substances, some of which are quite toxic, that penetrate deep into the lungs causing cancer and other related diseases. Because of this PM_{2.5} is mapped to a granular synth whose click rate increases as the amount of particles increase. This is meant to evoke the sonic archetype of a Geiger counter, where the increasing click rate signifies increased urgency / proximity, and in this case danger to listeners who experience it.

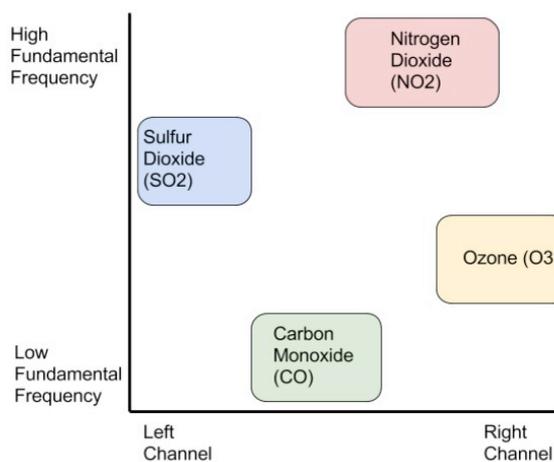


Figure 1: An approximate representation of the model’s frequency domain and stereo-space mappings (PM_{2.5} not represented here)

```
(
SynthDef.new(\granular, {
  arg amp, out, clickRate, pan;
  var sig;
  sig = WhiteNoise.ar(1);
  sig = GrainIn.ar(2, Impulse.kr(clickRate), 0.003, sig, pan);
  sig = sig * amp;
  Out.ar(out, sig);
}).add;

SynthDef.new(\fm, {
  arg freq, carPartial, modPartial, index, amp, pan;
  var mod, car;
  mod = SinOsc.ar(freq * modPartial, 0, freq * index);
  car = SinOsc.ar((freq * carPartial) + mod, 0, amp);
  Out.ar(0, Pan2.ar(car, pan));
}).add;

Pdef(
  \sonification,
  Ppar([
    Pmono(\fm,
      \freq, 50,
      \carPartial, 1,
      \modPartial, 2,
      \index, Pseq(~coindex, 1),
      \amp, Pseq(~coamp, 1),
      \dur, 0.2,
      \pan, 0,
    ),
    Pmono(\granular,
      \clickRate, Pseq(~pm25rate, 1),
      \amp, Pseq(~pm25amp, 1),
      \dur, 0.2,
      \pan, 0,
    ),
  ]),
).play;
)
```

Figure 2: Example synth definition and parameter mapping for CO and PM_{2.5} implemented in SuperCollider.

4.3. The Importance of Redundancy in Communicating a Message

Stream segregation between the four pollutants (CO, O₃, NO₂, and SO₂) is further reinforced by the redundant encodings of the same data in different mapping strategies. Data is encoded onto amplitude, so that pollutants become louder and more salient as they increase, as well as the stereo field, so that each pollutant occupies a fixed position in space, making it accessible to a wide array of commercially available (non-specialized) speaker arrangements and listening conditions. Previous research supports the idea that redundant integral mapping strategies improve performance in auditory graph comprehension [18].

In total, each of the four pollutant chemicals possesses four dimensional attributes. Two of them: spatial position and fundamental frequency, are meant to facilitate fast and easy identification of the pollutant. The other two, loudness and number of sidebands, afford the perception of change. What is unique and promising in this design – and particular dataset used – is that a redundant one-to-many parameter mapping actually becomes a perceptual strength instead of weakness, owing to FM’s maintenance of consistency of *harmonic identities*, promoting coherence of the overall listening experience.

4.4. A Brief Note on Spatialization for Public Sonification

A practical consideration, chronically under-addressed by sonification work is that the context in which sonifications are presented, at conferences, in auditoriums, classrooms etc. offer widely differing conditions for both audio quality and spatialization. Designing for public presentation means designing for a variety of conditions in mind. In many cases, good quality speakers are not readily available and they may not be spaced at a wide enough distance to encompass the entire audience within an immersive ‘sweet spot’. Under

these constraints, sonifications that rely solely on spatial mappings struggle to produce meaning for the audience, unless they are already encoded redundantly to other parameters. That is to say, redundant mappings are important not only because their integration with other auditory dimensions emphasize perceptibility, but also because in contexts where one mapping fails, the data can remain comprehensible. Designers in visual modes of representation, for instance, routinely account for this: it is important to choose colours that, when printed in black and white, still offer enough contrast to delineate the image. Auditory representations in practical contexts can and should operate under the same principle.

5. PRELIMINARY FINDINGS

While we have not yet had the opportunity to formally ‘user test’ this sonification with an actual public of everyday listeners, we have informally presented and tested it in several working groups, including two sets of audiences of 20+ each, and a sonification working group of 5 participants. Here we report first impressions from these presentations to a portion of a ‘general public’. We discuss the design in terms of perceivability, but also in terms of affective engagement with the issues at hand and its utility in generating fruitful dialogue about climate issues and city infrastructures as a result of accessing complex data in this way.

Ultimately, a successful sonification design solution reveals auditory gestalts, perceptible artifacts, or patterns in the data that hopefully inspire further research directions in the respective data domain. Based on the listening experiences during the 3 working group presentations of this sonification, the dataset anecdotally generated these results:

- Most everyone was able to readily identify the most polluted city (Sarnia) and the least polluted city (Vancouver)
- Most everyone was able to comfortably identify if not interpret the four harmonic identities of the chemical pollutants, and their spatial position in the recording
- The ‘Geiger counter’ mapping of particle matter pollution seemed intuitive
- With some re-listening, most people were able to identify a rhythmic pattern in the datasets in terms of ozone ebbs and flows related to solar activity

What is more interesting to us in terms of auditory displays as a form of public engagement is that listening collectively (rather than individually) proved a rich way of raising and discussing a number of questions related to specific patterns of air pollution in these geographic areas. Sharing the sonification experience – and having it as a reference to come back to – allowed us to communicate important additional information to the publics we interacted with, who otherwise would not have volunteered to learn more about air pollution or its health and climate repercussions. Further, while parameter mapping might seem a ‘technical’ detail, folks raised a number of interesting perspectives on the use of ‘logical’ and counter-logical approaches to representing

degrees and type of pollution. One person suggested that while noisy sidebands convey the idea intuitively, a more impactful and artistically driven approach could be using silence somehow – to signify the loss of healthy air and environments. Comments such as these opened an interesting discussion of not only semantic mappings of data, but potential connotations of mapping choices as culturally specific, and as part of a larger ecology of accessible information visualization and public knowledge translation.

One specific instance of having a fruitful and informative discussion based directly on the shared listening experience comes from listening to air pollution in Edmonton. This dataset exhibits a unique temporal emissions pattern that is not present in any of the other cities. Short bursts of emissions from different sources can be heard at extremely regular and predictable intervals throughout longer sections of the sonification for Edmonton. Our collective discussion specifically included brainstorming about possible causes of this pattern implying a spark of interest for further research into the infrastructural and industrial, as well as seasonal and environmental character of the city. Rush hour, although initially thought to be the cause, is not a possible explanation for the bursts; other cities would yield the same pattern if this was the case. What is curious is that, even though the short bursts sound at regular intervals, the source of the burst changes between the pollutants. At times there are prominent bursts in SO₂ and CO, but they will then switch to O₃ and NO₂ and back again. One current collectively-generated hypothesis is that the bursts are a result of a shared industrial practice across multiple sources of these pollutants, e.g. an active industrial complex or factory perhaps interacting with the weather. The point is that while we are not familiar enough with the data domain to make assumptions about causes, the listening experience of the sonified data provided not only enough recognition of relevant shifts and patterns, but also generates unique questions for further scientific inquiry and more importantly – public engagement.

6. CONCLUSIONS

By charting a practice-based design methodology for sonification, we demonstrate an array of aesthetic, functional, and practical design choices, which coalesce to produce sonic information. In the case of public engagement, our design work begins at a high level with a message and aesthetic we wish to convey, then works down to lower level decisions that reference synthesis methods and psychoacoustics. An enquiry into the use of sonification in the context of public relations reveals certain generalizable principles: at the most basic level, it illustrates how semiotic decisions, such as pollution-to-noise mappings, can be an integral and effective part of collectively interacting with information. Making a comparison to visualization practices for climate data, there are arrays of similar cultural/semiotic decisions that are not typically discussed as part of public presentation: e.g. the colour red is often chosen to demonstrate the most detrimental environmental effects of pollution. What we are getting at is that these decisions exist already in the contexts of public knowledge translation, and that they must be acknowledged as relying on already established archetypes and ‘perceptual mappings’. Listening to sonifications, as a novel form of public engagement opens the door to having these sorts of conversations collectively and bringing attention to the culturally-specific and semiotically-driven

mappings of data-to-modality. The alternative is that these decisions are ignored at the cost of truly understanding what constitutes an effective information design strategy, visual, aural, or otherwise. Our assertion is that opening up a sonification design conversation in this way can uniquely and meaningfully inform the diversity of design choices for sonification, taking also in account the context of reception and variety of listening outcomes at hand. In that sonification sheds light on the ways we choose to communicate scientific data for all potential listeners, including the broader public. Living in an ocularcentric world, it is oftentimes easy to forget that all the graphs, Venn diagrams, box plots, infographics, and visualizations we use are full of calculated design choices meant to engage a viewer in particular ways – from scaling, to color theory, to graph and chart shapes. Sonification is no different. By exploring sensory modes alternative to visual designs, we can begin to rediscover the latent practices that govern how we communicate knowledge. Sonification opens the doors to critique them, and offers solutions to improve them.

Attention”, in *Proceedings of the International Conference on Auditory Display*, Limerick, 2005

7. REFERENCES

- [1] T. Hermann, A. Hunt, and J. Neuhoff, *The Sonification Handbook*. Logos Publishing House, Berlin, 2011.
- [2] http://geant3.archive.geant.net/Media_Centre/Pages/Higs-like-Boson-Sonification.aspx
- [3] <http://blogs.esa.int/rosetta/2014/11/11/the-singing-comet/>
- [4] <http://createdigitalmusic.com/2016/02/a-major-breakthrough-in-physics-is-heard-not-seen/>
- [5] A. Supper. “Sublime frequencies: The construction of sublime listening experiences in the sonification of scientific data,” *Social Studies of Science*, February 2014, 44(1), pp. 34-58.
- [6] J. H. Schuett, R. J. Winton, and B. N. Walker. “Comprehension of Sonified Weather Data Across Multiple Auditory Streams,” in *Proceedings of International Conference on Auditory Display*, 2014.
- [7] V. Goudarzi, K. Vogt, and R. Höldrich. “Observations on an Interdisciplinary Design Process Using a Sonification Framework” in *Proceedings of International Conference on Auditory Display*, 2015.
- [8] <http://www.bcairquality.ca/assessment/air-monitoring-data.html>
- [9] <http://airdata.aemera.org/>
- [10] <http://www.airqualityontario.com/history/index.php>
- [11] A. deCampo, “Toward a Data Sonification Design Space Map,” in *Proceedings of the 13th International Conference on Auditory Display*, Montreal, 2007.
- [12] J. H. Flowers, “Thirteen Years of Reflection on Auditory Graphing: Promises, Pitfalls, and Potential New Directions,” in *Proceedings of the International Conference on Auditory Display*, Montreal, 2007.
- [13] <https://soundcloud.com/sysonproject/toa-in-and-out-radiation>
- [14] <https://soundcloud.com/marcestpierre>
- [15] S. Carlile, *The Sonification Handbook*. Logos Publishing House, Berlin, 2011, ch. 3 – Psychoacoustics, pp. 41-61.
- [16] P. R. Cook, *The Sonification Handbook*. Logos Publishing House, Berlin, 2011, ch. 9 – Sound Synthesis for Auditory Display, pp. 197-235.
- [17] <http://www3.epa.gov/pmdesignations/faq.htm#0>
- [18] S. C. Peres and D. M. Lane, “Auditory Graphs: the Effects of Redundant Dimensions and Divided

LOST OSCILLATIONS: EXPLORING A CITY'S SPACE AND TIME WITH AN INTERACTIVE AUDITORY ART INSTALLATION

Jim Murphy, Dugal McKinnon, Mo H. Zareei

New Zealand School of Music, Victoria University of Wellington
PO Box 600, Wellington 6140, New Zealand

Jim.Murphy@vuw.ac.nz, Dugal.McKinnon@vuw.ac.nz, Mo.Zareei@vuw.ac.nz

ABSTRACT

Lost Oscillations is a spatio-temporal sound art installation that allows users to explore the past and present of a city's soundscape. Participants are positioned in the center of an octophonic speaker array; situated in the middle of the array is a touch-sensitive user interface. The user interface is a stylized representation of a map of Christchurch, New Zealand, with electrodes placed throughout the map. Upon touching an electrode, one of many sound recordings made at the electrode's real-world location is chosen and played; users must stay in contact with the electrodes in order for the sounds to continue playing, requiring commitment from users in order to explore the soundscape. The sound recordings have been chosen to represent Christchurch's development throughout its history, allowing participants to explore the evolution of the city from the early 20th Century through to its post-earthquake reconstruction. This paper discusses the motivations for *Lost Oscillations* before presenting the installation's design, development, and presentation.

1. INTRODUCTION AND MOTIVATIONS

Sound archives allow dedicated researchers access primary sources associated with the history of a place and the events which shaped it. By engaging in a longitudinal listening survey, as a kind of sonic archeology, researchers may unearth the development of a place, but also vicariously experience significant events that have affected an area over the course of its (phonographic) history. Those persons less able or inclined to peruse large amounts of recorded material, they may be largely unaware of the sonic stratigraphy upon which they live. One use of auditory display in an artistic installation context is to convey this sonic history to the public in an affectively compelling and aesthetically-motivated manner.

This paper explores one such sonic artwork, the interactive *Lost Oscillations* installation. Installed in Christchurch, New Zealand in October, 2015, *Lost Oscillations* allows participants to engage in a multi-sensory exploration of recorded events throughout the last eight decades of Christchurch's history.

In informal contemporary dialog, Christchurch's history prior to the 2010 and 2011 Canterbury earthquakes is muted: the earthquakes function as a sort of 'event horizon,' and there is little dis-



Figure 1: The purpose-built *Lost Oscillations* user interface, shown with two of its eight loudspeakers.

ussion of events preceding them. A key goal of *Lost Oscillations* is to slip past these traumatic events, allowing those interacting with the artwork to engage, through touch and listening, with Christchurch in a context that largely focuses on its pre-earthquake history. A tactile, multi-user interactive interface (shown in Figure 1) was chosen as the means by which participants may engage with an array of archival recordings. This decision was arrived at both to allow communal experience of sonic history and to emphasize embodied connection, through sound and touch, to sonic place. To further provide those visiting the artwork with an immersive, mixed reality, site-specific auditory experience, sounds triggered from the interface are output to an eight-channel loudspeaker array. This octophonic array spatializes the sounds relative to the installation's central-Christchurch location: sounds recorded in locations to the north of the installation space, for example, are output to northward-located loudspeakers.

The remainder of this paper provides a technical and aesthetic overview of *Lost Oscillations*. It begins with a discussion of related works, presenting a number of pieces whose usage of interface design and artistic aesthetic were influential in the artwork's initial conception and development. Following the review of related works, an overview of the design, development, and construction of *Lost Oscillations* is provided. Section 3.3 discusses



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

the purpose-built interface, its connection to a host computer, and the computer’s audio output configuration. Similarly, Section 3.4 details the user interface’s microcontroller firmware as well as the audio playback software used on the audio host PC. Finally, Section 3.2 discusses the audio used in *Lost Oscillations*, including both archival recordings and contemporary field recordings. After discussing the physical design of the artwork, the installation and use of *Lost Oscillations* is presented, focusing on its interactivity and engagement by the public. Finally, the paper concludes with a discussion of future avenues for similar works as well as a number of potential improvements that may be undertaken in future iterations of *Lost Oscillations*.

2. RELATED WORKS

Lost Oscillations fuses three sonic arts subdisciplines together, including elements pertaining to the development of new interfaces for musical expression, spatio-temporal sound art, and acousmatic composition approaches. As such, key works from each of the three subdisciplines served as inspirations for *Lost Oscillations*. This section details a number of such works.

2.1. Interface Design and Diffusion Performance

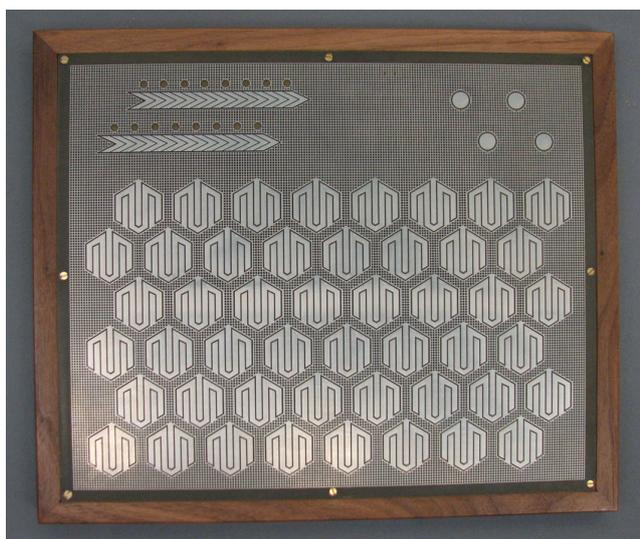


Figure 2: The Manta touch-sensitive audio interface: an array of touch-sensitive electrodes that may be configured for audio interfacing applications. Photo courtesy <http://www.snyderphonics.com/>

There is a rich history of new musical controller interfaces featuring touch-sensitive electrodes. Whether capacitive or resistive, these interfaces allow a user to affect an auditory output through the use of a touch event. An early example of such interfaces is the Cracklebox, a small analog soundmaker module developed in 1973 by Michel Waisvisz at STEIM¹. The Cracklebox’s influence has extended to modern MIDI-enabled musical performance tools, including the recent Snyderphonics Manta². The Manta (shown in

¹<http://steim.org/product/cracklebox/>

²www.snyderphonics.com/manta.htm

Figure 2) is a touch-sensitive MIDI controller whose configuration was a key influence during the design of the *Lost Oscillations* interface: capable of USB-based MIDI communication, the Manta may be configured to trigger and affect audio events on a PC.

Further, the second author’s prior audio interfaces were also used as reference points in the design of the *Lost Oscillations* interface: the decision to pursue touch input with no tactile feedback was chosen in part due to the corporeal engagement afforded by the membrane potentiometer-based *Helio* interface, described in [1].

In addition to turning to notable general purpose performance interfaces for inspiration when developing *Lost Oscillations*, a number of multichannel-specific interfaces were examined. [2] provides a history of such interfaces, many of which are used for live ‘diffusion’ performance and feature audio mixing desk-style user interaction schemes, containing arrays of linear potentiometers for adjusting loudspeaker gains. In contrast to these general-purpose interfaces, intended to be used across a number of different music genres, the *Lost Oscillations* interface allows users to explore a predefined range of samples. As such, a sample-set specific interface may be used, allowing for a close coupling between interface and audio.

2.2. Spatio-Temporal Sound Art

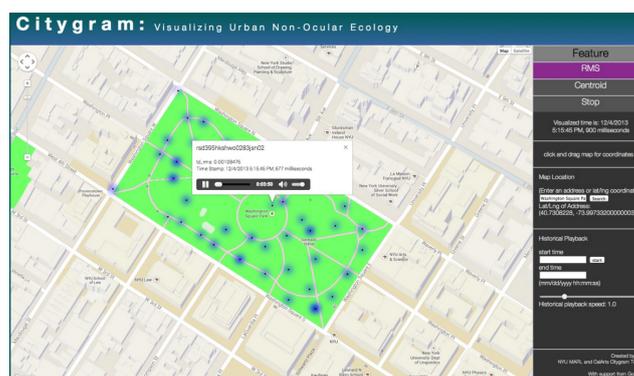


Figure 3: The Citygram interface, showing a map with audio features spatially illustrated. Screenshot courtesy <http://cds.nyu.edu/projects/citygram-sound/>

In developing *Lost Oscillations*, existing artworks that allow participants to explore the spatiotemporality of a city or place were examined. Third author Mo H. Zareei’s *Complex*, a “physical re-sonification of urban noise” (described in detail in [3]) was a key prior inspiration to *Lost Oscillations*: with a number of kinetic sound sculptures positioned on a map in locations relative to microphone locations in the city, Zareei’s work made use of the Citygram locative sonification dataset (described in [4] and illustrated in Figure 3) to convey a city’s temporally-morphing sonic feature-set.

Lost Oscillations shares this spatial coupling, using its map-like interface to allow participants to explore sounds recorded at locations upon which electrodes are placed on the interface. Where *Complex* is a self-running kinetic sound sculpture, *Lost Oscillations* focuses on participatory engagement with the piece, requiring people to engage with the interface to remain in physical contact with it in order for audio elements to be played back.

3. DESIGN, DEVELOPMENT, AND USER EXPERIENCE

The development of *Lost Oscillations* focused on three areas: the system's hardware, its software, and the accompanying audio files to be played back. After providing an overview of the piece's functioning, this section discusses each of these three and details the means by which participants interact with the artwork.

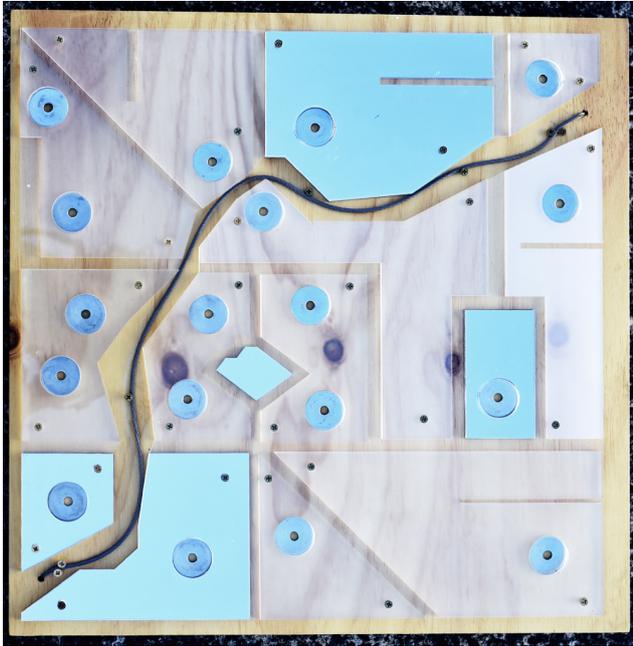


Figure 4: Top view of the *Lost Oscillations* user interface, showing electrodes placed throughout a map-like representation of the city center of Christchurch, New Zealand.

3.1. User Experience

Upon approaching the user interface at the center of a ring of loudspeakers, a participant sees a sculptural interface (shown in Figure 4) made of wood and plastic and interspersed with a number of metal pads. Upon touching the pads, an ambient soundscape produced by the loudspeaker array fades out to be replaced by an audio recording emanating from the loudspeakers in the direction of the sound file's location of recording. The participant may touch more than one metal pad, resulting in multiple playback events, each spatialized relative to their real-world recording location. After moving away from the contact points, the recordings fade out and are slowly replaced by ambient audio emanating from all directions. Another significant aspect of the interaction with audio materials is that users cannot access the same temporal point in any audio file as the audio is constantly, but silently, playing in real-time and only becomes audible when an electrode is touched. Here the intention was to highlight time, and in particular the evanescent of sound and the vital role of memory in auditory experience given its invisible, intangible sound objects.

The following subsections provide specific details about the development of the project, with a focus on the audio files used as well as the hardware of the user interface and the software controlling audio playback.

3.2. Audio

The development of *Lost Oscillations* began with the selection and creation of audio files to be played back in response to user input. 110 sound files were selected, spanning the years 1935 to 2015. These sound files were obtained both from Ngā Taonga Sound and Vision, New Zealand's archive of film and video, and also include self-made field recordings that were recorded in Christchurch in the weeks prior to the installation of *Lost Oscillations*.

The archival audio samples from Ngā Taonga Sound and Vision were chosen to equally represent everyday events and historically-significant events. Further, the audio was selected with consideration given to the location of its recording: by selecting recordings made at the same locations over the course of many decades, a sort of stratigraphic column of sound material is formed. Sitting at the top of this column are the most recent recordings, created by the third author during 2015 in the weeks leading up to the piece's installation. These contemporary field recordings serve to provide a direct coupling between the "real-world" sonic cityscape, in which the participants are immersed, and the recorded soundscapes played through the loudspeaker array and thereby maximize the mixed-reality ambiguity described below.

After selecting the sound files from Ngā Taonga Sound and Vision and creating a number of field recordings, the sound files were organized according to the location of their recording. All of the archival and field recording made in a single location were then merged together into a single audio file, creating a temporal stream of audio anchored to a single point in space.

3.3. Hardware

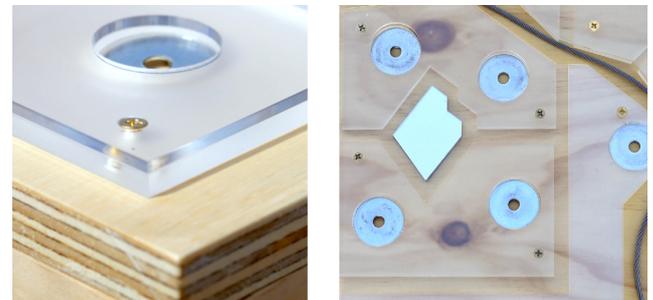


Figure 5: Detailed views of the user interface's electrodes. At left, an electrode mounted within a plastic panel; at right, a number of electrodes made from metal washers and wire braid.

After selecting, organizing, and compiling the sound files into discrete streams, the work on *Lost Oscillations* shifted to the realization of the user input device and the interfacing schemes between the input device and the audio playback software.

In essence, the user interface in *Lost Oscillations* is a MIDI input device that transmits MIDI messages in response to user interaction events. Physically, the interface consists of a large wooden box that serves both as enclosure for transducer control and communications electronics and as a mounting chassis for the interface's electrodes.

The purpose-built physical enclosure is constructed from materials chosen as representative of the post-earthquake

Christchurch rebuild, containing plywood surfaces, plastic ornamentation, and metal electrodes made from wire braid and large washers of the type used in building construction projects (shown at right in Figure 5). These materials were selected with the intention of affording tangible symbolic engagement with the city. Such a focus on the affective properties of materiality (as explored in Jane Bennett’s *Vibrant Matter* [5]) serves to further couple *Lost Oscillations* to the location in which it is installed. Further, the layers of materials on the interface (shown at left in Figure 5) provide a visual connection to the sort of aural stratigraphic layering present in the audio files (discussed in Section 3.2).

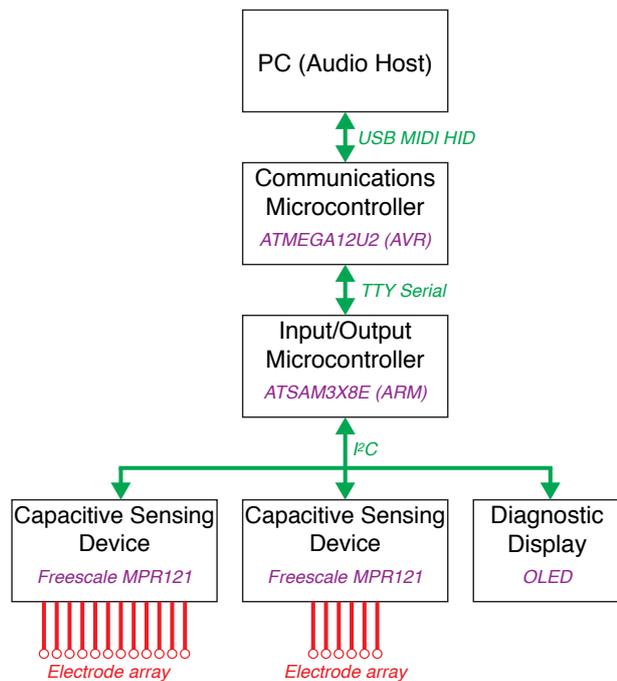


Figure 6: A block diagram illustration the *Lost Oscillations* user interface circuit.

The *Lost Oscillations* electronics consist of a number of subassemblies that are interconnected, acting together to allow touches to be detected and result in audio playback on a host PC. Figure 6 shows a block diagram of the electronics, which are described below.

Each of the 19 electrodes is connected to a Freescale MPR121 capacitive sensing device. The MPR121 is intended to allow a number of electrodes to easily be connected to a microcontroller. To simplify communications between a host microcontroller and the capacitive sensing device, the MPR121 uses the I²C protocol (described in more detail by [6]). The I²C protocol allows multiple MPR121 devices to be connected together on a shared bus, with each device assigned its own address. As the MPR121 allows for a maximum of 12 electrodes to be connected, a second MPR121 is used in the *Lost Oscillations* interface to allow for all of the 19 electrodes to be scanned.

The I²C bus employed by *Lost Oscillations* allows for two-way communication: the MPR121 devices may be digitally configured at startup in addition to subsequently reporting their electrode states (discussed below in Section 3.4). This configuration

capability allows the electrode sensitivities to be individually set, decreasing the chance of false positive electrode sensing events. In addition to using I²C to control and communicate with the MPR121 devices, a small I²C OLED diagnostic display is also connected to the bus. This display allows for rapid debugging and status checks to be conducted while the artwork is installed in the field.

The *Lost Oscillations* user interface employs an Arduino Due microcontroller development board to handle its input, output, and communications. The Arduino Due consists of two separate microcontrollers: the primary microcontroller is a 32-bit AT-SAM3X8E ARM device; the secondary microcontroller is an 8-bit ATMEGA32U4 AVR device. The primary microcontroller is used to communicate with the MPR121 capacitive sensors. It communicates with the secondary microcontroller via TTY serial messages. The secondary microcontroller’s role is to convert the TTY serial messages from the primary microcontroller into USB MIDI HID messages that may be read by a digital audio workstation on a host PC. To serve as a USB MIDI HID device, the secondary microcontroller is programmed with the HIDUINO firmware, allowing for driverless MIDI communications [7].

The I²C devices along with the two microcontrollers make up the purpose-built electronics assembly for the *Lost Oscillations* user interface. Due to the universal compatibility of the HIDUINO-equipped MIDI HID device, any computer with an operating system designed to handle HID systems may be used for receiving MIDI touch events for *Lost Oscillations*. In practice, a Mac Mini was chosen due to its relative low cost, small size, and compatibility with the digital audio workstation software discussed in the following subsection.

3.4. Software

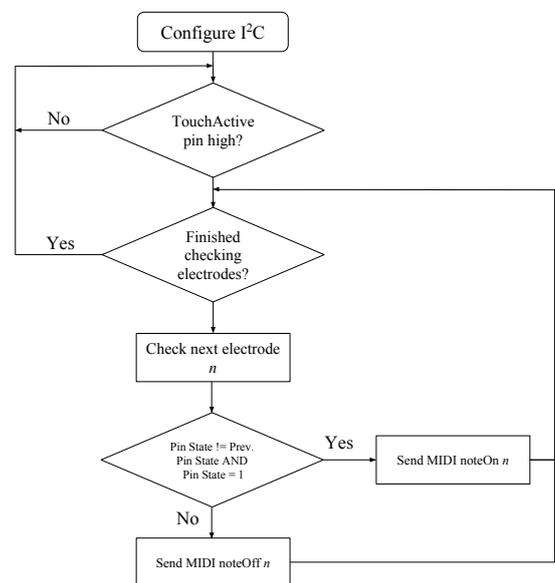


Figure 7: A program flow diagram of the capacitive sensor and MIDI communications scheme deployed on the ATSAM3X8E.

Lost Oscillations makes use of two software subsystems: one on the microcontroller assembly presented in the preceding subsection and a second on the audio host PC.

As discussed above, the microcontroller assembly consists of two separate microcontrollers, one to handle MIDI and one to handle user input and output. While the MIDI communications microcontroller is equipped with a standard HIDUINO firmware (as presented in [7]), the input/output microcontroller contains a custom-developed firmware.

Figure 7 is a program flow diagram illustrating the means by which the MPR121 capacitive touch sensor arrays are read and, in the event of a touch or release event, MIDI messages sent over TTY serial to the HIDUINO-equipped communications microcontroller. As shown in the figure, the firmware is quite simple, and (following configuration of the MPR121 devices) simply loops through the incoming messages from the MPR121 devices, checking to see whether an electrode's current state is different than its previous state. If the state is different and the electrode is currently sensing a touch event, then a MIDI NoteOn message corresponding to the electrode's number is sent to the ATMEGA32U2; if the state is different and the electrode is not sensing a touch event, a MIDI NoteOff event is sent. After the ATMEGA32U4 receives the TTY serial message, it is converted into a USB MIDI HID event and is sent over USB to the host PC.

The host PC makes use of the Ableton Live digital audio workstation software to read incoming MIDI messages and adjust the gain of pertinent audio clips. As discussed in Section 3.2, an ambient drone plays continually in the background, attenuating only when an electrode touch event occurs. Therefore, any incoming MIDI NoteOn message attenuates the drone track and, concurrently, increases the gain on the audio clips whose electrodes were just touched. After one minute of no input events, the drone track begins to increase in gain, reaching a set volume and looping until a MIDI NoteOn event occurs.

4. INSTALLATION AND USE

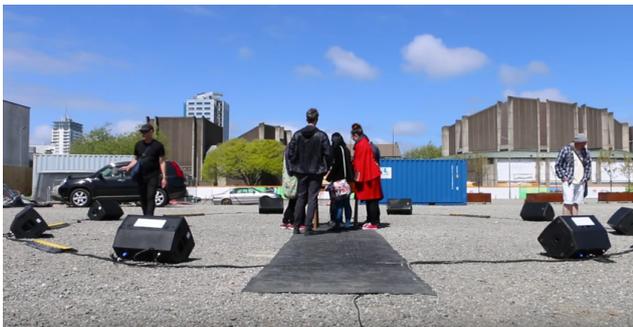


Figure 8: *Lost Oscillations* installed in Christchurch as part of the 2015 Audacious sound art festival. Visible here are a number of the installation's eight loudspeakers, arranged in a ring around the central user interface.

Lost Oscillations debuted at Christchurch's 2015 Audacious Festival of sonic art works. As a site-specific installation, the piece's location and on-site configuration are significant to the work's outcome. As such, these two elements are discussed in detail in the following two subsections.

4.1. Location

When installed, the user interface and speaker array were positioned outdoors in the center of Christchurch, surrounded by the locations at which the piece's sound materials were recorded or which are referred to in spoken-word materials. The virtual soundscape therefore has a strong situated audio aspect, using sound spatialization in combination with topographic cues provided by the stylized map and the location of the sensors on it, to alert users to the location of historical events and materials in terms of the actual soundscape around them. Furthermore, the ambiguity between the actual and virtual (8-channel) soundscapes create a strong mixed-reality experience, thereby sonically and experientially connecting the past and present soundscapes of the city, and blending its virtual sounds with the present-day landscape of the city. It was intended that such blending would result in spatiotemporal ambiguity, requiring that participants determine whether a sound was created in the "here and now" of the city or, rather, in the "there and then" of the virtual sound world.

To couple with the vital materiality of the interface's physical components, a rock-strewn vacant lot in central Christchurch (shown in Figure 8) was selected as the installation location. As the work was sited at the location the Christchurch Art Gallery prior to its demolition in the aftermath of the 2010 Canterbury Earthquake, the vacant lot served as a tabula rasa (surrounded by the city and its ambient noises) from which the contemporary cityscape's sounds could be combined with user-triggered historical audio events.

4.2. Configuration and Use

After selecting a location, the on-site configuration of *Lost Oscillations* was considered. To allow for relatively accurate placement of phantom audio sources, an eight-channel loudspeaker array consisting of large weather-resistant monitor speakers was chosen as the means of audio playback. The loudspeakers are positioned around the centrally-located user interface: users approach the user interface by stepping inside the ring of loudspeakers, an act which physically and metaphorically immerses them in the situated mixed-reality audio-space of the piece's situated audio.

After placing the speakers, the individual audio files' gains were adjusted *in situ*, with the objective of balancing the audio files' loudness with that of the ambient noise of the Christchurch cityscape. After completing the level balancing, many of the field recordings were difficult to distinguish from real-world sounds occurring in the city at large; such ambiguity between audio file and ambient noise indicated that the pursuit of inconspicuous situated audio was a successful one.

Once the installation was opened to the public, participants began to engage with the artwork. It was observed that the multi-touch and multi-user capabilities of the audio interface led to interesting and unanticipated interactions between different users. When more than one user was touching the interface's various electrodes, some participants began to explore different rhythmic and timbral means of interlocking the audio files that they each controlled. Such open-ended user interaction schemes indicate that the interface developed for *Lost Oscillations* is a flexible device that may be re-used or extended in future pieces with different contexts.

During the installation of *Lost Oscillations*, video footage was captured in order to allow the inherently

transient site-specific work to be viewed by those unable to visit it. This documentation may be viewed at <https://www.youtube.com/watch?v=eajBHvfvfNU>.

5. FUTURE WORK

Lost Oscillations is the first in an intended series of interactive sound art installations focusing on the spatio-temporal exploration of sound. After installing and qualitatively evaluating *Lost Oscillations*, a number of aspects will be altered in future iterations of the piece.

The electrodes in *Lost Oscillations* are fixed in space, attached to the surface of the interface's enclosure. While touching the electrodes results in spatially relevant audio, there is no accompanying physical analog to the temporality of the audio. Future versions of the interface will feature electrodes mounted to sensors that allow vertical displacement to be transduced, letting users press down to explore "deeper," older sounds. Such depth-related control over the sound will further explore the underlying theme of sonic archeology, allowing users to metaphorically excavate a certain area's sound, beginning by exploring recent events and, by pressing down and lowering the electrode, proceeding through to listen to older sounds. Additionally, the depth element is envisaged as increasing the affective affordance of the interface, as interaction requires increased effort and control from users, encouraging them to invest greater attention to, and corporeal engagement with, the installation.

After exploring enhancements to the transducers, additional versions of *Lost Oscillations* will be developed, each one pertaining to a different city or region. A further area for development in these regionally-discrete interfaces will be the employment of regionally relevant materials in the construction of their interfaces, further coupling an area's physical materiality with the sounds that are explored by those using the interface.

6. CONCLUSIONS

As an interactive auditory installation, *Lost Oscillations* combines technical and aesthetic sonic arts subdisciplines in order to create a piece that requires the human touch to explore the contemporary and historical soundscapes of Christchurch, New Zealand.

Technically, *Lost Oscillations* serves as a case study in user interface design for interactive auditory artworks, demonstrating that a relatively simple, low-parts-count, driverless interface may be effectively used to allow participants a meaningful way to interact with situated audio. It is hoped that the design and development of the physical interface for *Lost Oscillations* may serve as a starting point for future auditory installation interfaces.

Aesthetically and conceptually, the outcome of *Lost Oscillations* may be viewed as a success: participants were able to engage in a physical and aural exploration of the "sonic archeology" of the city, experiencing the city's auditory past within the greater context of the present-day soundscape of the city. It is anticipated that the technical, conceptual, and aesthetic outcomes of *Lost Oscillations* will be the first in a series of related artworks focusing on the layered spatio-temporal sonic stratigraphy of cities and other spaces.

7. ACKNOWLEDGMENTS

(Section Removed for Anonymous Review Process)

8. REFERENCES

- [1] (Reference Removed for Anonymous Review Process)
- [2] Johnson, B. (2014). Emerging Technologies for Real-Time Diffusion Performance. In *Leonardo Music Journal*, Volume 24. MIT Press.
- [3] (Reference Removed for Anonymous Review Process)
- [4] Park, T. H. et al. (2013). Locative Sonification: Playing the World Through Citygram. In *Proceedings of The 2013 International Computer Music Conference (ICMC-2013)*, Perth, Australia.
- [5] Bennett, J (2010). *Vibrant Matter: A Political Ecology of Things*, Duke University Press.
- [6] NXP Semiconductors N.V. (2014). *I²C Bus Specification and User Manual*, NXP Semiconductors,.
- [7] Diakopoulos, D and Kapur, A. (2011). HIDUINO: A firmware for building driverless USB-MIDI devices using the Arduino microcontroller. In *Proceedings of the 2011 Conference on New Interfaces for Musical Expression*, Oslo, Norway.

ORAL PAPERS

Body and Mind

TEMPO-FIT HEART RATE APP: USING HEART RATE SONIFICATION AS EXERCISE PERFORMANCE FEEDBACK

Steven Landry, Yuangjing Sun, Darnishia Slade, and Myounghoon Jeon
Michigan Technological University,
Mind Music Machine Lab,
1400 Townsend Drive, Houghton, MI 49931
{sglandry, ysun4, dslade, mjeon}@mtu.edu

ABSTRACT

Physical inactivity is a worldwide issue causing a variety of health problems. Exploring novel ways to encourage people to engage in physical activity is a topic at the forefront of research for countless stakeholders. Based upon a review of the literature, a pilot study, and exit interviews, we propose an app prototype that utilizes music tempo manipulation to guide users into a target heart rate zone during an exercise session. A study was conducted with 26 participants in a fifteen-minute cycling session using different sonification mappings and combinations of audiovisual feedback based on the user's current heart rate. Results suggest manipulating the playback speed of music in real time based on heart rate zone departures can be an effective motivational tool for increasing or decreasing activity levels of the listener. Participants vastly preferred prescriptive sonifications mappings over descriptive mappings, due to people's natural inclination to follow the tempo of music.

1. INTRODUCTION

Physical inactivity is a worldwide issue, causing health problems such as obesity, high blood pressure, diabetes, and other psychosocial problems [1]. Finding new and effective ways to motivate and guide users in their exercise sessions could help solve this epidemic of inactivity. Music is known to distract listeners from the monotony of exercise, and music tempo has a direct link to both perceived and actual physical output [2] of a work out session. This paper explores the viability of using music tempo manipulation as a means to guide and motivate users to exercise more effectively. The mission of the *Tempo-Fit Heart Rate App* is to offer users a simple, easy to use, effective, and motivating approach to improving their physical activity and achieving target health goals through the use of musical tempo feedback.

Previous studies have examined the viability of sonifying physiological data to guide runners into a predefined optimal heart range. For instance, Wärnegård developed and tested an app that provided auditory warnings (Earcons) when the user's heart rate fell outside of the predefined optimal range. This sonification strategy helped users maintain a consistent heart rate inside the optimal range

compared to a control condition [3]. Instead of discrete Earcons, the proposed Tempo-Fit app modifies the playback speed (tempo) of the user's music to either reflect current heart rate (descriptive, or "what you are doing") or guide the user (perspective, or "what you should be doing") back to the optimal zone. Using this type of feedback system allows listeners to continuously monitor their changing HR while also providing more interesting and motivating auditory cues than simple Earcons.

Auditory research has shown time and time again the relationship between music tempo preference and exercise intensity [4]. Multiple studies have observed that when exercising, participants vastly prefer to listen to music with medium to high tempos (< 120 bpm). Even for low intensity exercise (40% max HR_{reserve}, or around 80 bpm for a 20 year old), participants showed a significant preference for both medium (120bpm) and high (140bpm) tempo music over slow tempos (80bpm). The same preference was found for higher intensity exercises (75% max HR_{reserve}, or around 150bpm for a 20 year old), suggesting the relationship between HR during exercise and music tempo preference is not perfectly linear. Multiple exercise motivational applications (such as MusicalHeart, RockMyRun, Mptrain, etc.) analyze heart or step rate of the exerciser to suggest songs from a database with matching BPM's [5, 6, 7].

The Tempo-Fit heart rate application presented in this paper takes a different approach to using the BPM of music for exercise motivation. The music tempo (or in this case, speed of playback) is manipulated in real time to give cues to the listener's current activity level. Since it is not necessary to match the exact music BPM to heart rate, and there already exists a general preference to medium to high music BPM's regardless of exercise intensity level [4], users are allowed to select their own music. The application speeds up or slows down the playback speed of the current song as an auditory cue that the user's heart rate has dropped below or above the recommended (or preset) HR range. Playback of the audio file is temporarily manipulated to either 125% or 75% speed. This manipulation gradually increases or decreases over five seconds, emphasizing the change in tempo, as opposed to the actual BPM of the song. Once the user's HR returns to the desired range, the playback speed of the audio file immediately returns to 100% (normal) speed. Our hypothesis is that the gradual change in tempo will be a more effective motivational feedback cue than the actual tempo of the music at any point in time. To test this hypothesis, and to gauge preference for either prescriptive (what you should be doing) or descriptive (what you are doing) sonification mapping, we



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

invited participants to try out our system during a fifteen minute exercise session on a cycling machine.

2. METHODS

2.1. Participants

All participants ($N = 26$, 4 females, $M_{age} = 20.1$, $SD_{age} = 1.3$, age range: 18-22 years) were recruited from the local university undergraduate participant pool. Each participant was compensated with two psychology course credit points for an hour long study. The majority of participants (20/26) rated themselves as fairly athletic and reported going to the gym or working out semi-regularly (at least once per week). All participants were screened for medical conditions associated with increased risks of complications in relation to physical exertion. No other demographic information was collected.

2.2. Stimuli and equipment

Figure 1 depicts the laboratory setting where the cycling session took place. A 20 inch computer monitor was positioned three feet in front of the participant for television viewing. A smartphone (Galaxy S3) was placed on a table approximately two feet away from the participant. The smartphone app displayed a visual HR (approximately two squared centimeters in size) slightly outside of the participant's field of view, forcing the participant to turn their head away from the television monitor to visually check their heart rate readout from the smartphone. A Monark 818E cycling machine (with participant-chosen resistance weight; normally between .25 and 2.0 KG) was used as the exercise equipment. The music stimuli were played from computer speakers positioned 3 feet away from the participant at a volume level averaging around 80 dB. All participants heard the same four songs (see Table 1) in the same order approximately twice during the exercise session. All songs fit into the music genre of dance music, varying from 117-130 BPM.

Table 1: Track titles and BPM of each song used for auditory feedback

Artist – Song title	BPM
Adam Lambert – For Your Entertainment	130
Daft Punk – One More Time	123
Casio Kids – Fot I Hos	128
Justin Timberlake – Sexyback	117

2.3. Conditions

A between-subjects design was implemented where participants were randomly assigned to one of the following 4 conditions:

Control – No tempo manipulation (music played at normal speed regardless of participant's HR). HR was visually displayed on the smart phone (updated once every five seconds) positioned next to the participant.

Prescriptive – “What you should be doing” mapping: Music tempo is gradually *increased* (over five seconds) to 125% speed when the participant's HR was *below* the target range. Music tempo is gradually *decreased* (over five

seconds) to 75% speed when the participant's HR was *above* the target range. Once the participant's HR returned to the target range, the music tempo immediately jumped back to 100% (normal) speed. Visual feedback was also provided in the same manner as the control group.

Descriptive – “What you are doing” mapping: Opposite to the prescriptive condition; the music tempo *increased* to 125% speed when the participant's HR was *above* the target range. Music tempo *decreased* to 75% when the participant's HR was *below* the target range. Music returned to 100% (normal) speed immediately when the participant's HR was within the target range. Visual feedback was provided in the same manner as the above conditions.

Music Only – Auditory only, no visual feedback: Auditory feedback was provided in the same manner as the prescriptive condition (“What you should be doing” mapping). No visual HR information was provided.

2.4. Procedure

Following the consent form and screening questionnaire, the participant equipped the Equivital vest [8] which sends the participant's HR data over Bluetooth to the smartphone app once every five seconds. The participant was then instructed to select a TV episode to watch from Netflix.com to watch on the computer monitor. The TV sound was muted and subtitles were enabled to more closely resemble a gym environment and to provide incentive to use the auditory display over the visual display. All but four participants chose comedic cartoon shows such as “Family Guy” or “Futurama”.



Figure 1: Actual setup of experiment, including: TV monitor for Netflix.com, smartphone (Galaxy S3), Monark 818E cycling machine. Music played from experimenter controlled PC speakers.

After selecting the show on Netflix, the target HR range was calculated for each participant using the formula: $\min = [(220 - \text{age}) * .5] + 5$, $\max = [(220 - \text{age}) * .6] - 5$ [8]. Five was added to the minimum and subtracted from the maximum to decrease the target zone from a range of 20 to a range of 10 to prevent a ceiling effect on performance. The majority of the participants' personally calculated target heart rate ranges were around 125-135 BPM. The participant was then instructed on his or her group's particular sonification mapping.

After confirmation that the participant understood the instructions, he or she would begin pedaling on the Monark 818E cycling machine. Once the participant's HR was within

the target range for 20 consecutive seconds, the experimenter would begin the Netflix TV show, the music, and a 15 minute timer. The experimenter manually adjusted the playback speed of the music in a “Wizard of Oz” fashion. Tempo manipulation was done via a custom Max/MSP patch. After the 15 minute cycling period, the experimenter conducted a semi-structured interview before releasing the participant to assess user preferences.

To operationalize performance of the different feedback strategies, the percentage of time the user’s heart rate fell within the target range was calculated for each participant (heart rate samples within target range / total samples = percentage of time in target range).

3. RESULTS & DISCUSSION

Each participant experienced one condition, and the percent of HR samples (1 HR sample every 5 seconds) was used as the dependent variable. From Figure 2 we interpreted no significant differences ($p > .05$) between group performances due to the overlapping 95% confidence intervals and extremely small sample sizes within groups. However, this suggests that the music only condition performed as well as all other conditions including visual feedback.

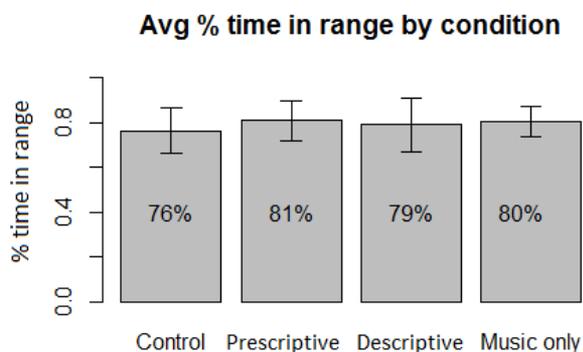


Figure 2: Average percent in range by condition. All error bars represent 95% confidence intervals.

All but three participants said they would use the app if available. Participants reported the level of activity of the cycling session to be equivalent to a brisk walk or light jog. All participants (except those in the control condition) considered the music helpful. Participants in prescriptive and descriptive conditions reported a strategy of watching TV for 20-30 seconds, checking the visual read out for 1-3 seconds, then back to the TV. Participants in the control condition compensated without auditory feedback by ignoring the TV show and allocating attention primarily on reading the small HR digit. Even with this compensation strategy, participants in the control condition stayed within the target range for 76% of the session (lowest of all four conditions) compared to 80% in the music only condition. Those participants in the descriptive condition often reported that the mapping of HR to tempo was not intuitive, or confusing. For instance, whenever a participant in this condition’s HR increased above the target range, the tempo of the music would increase to 125%. This increase in tempo was sometimes associated with an involuntary 3-5 BPM spike in the participant’s HR even though the participant knew that an increase in the tempo of the music was used to suggest that they should decrease their activity level to return their HR to

the target range. Figure 3 shows the sampled HR from a participant in the descriptive condition, and figure 4 shows the sampled HR from a participant in the prescriptive condition.

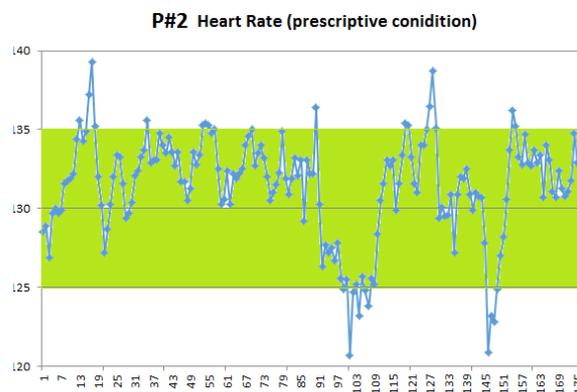


Figure 3: One participant’s HR (bpm) over time. The shaded area represents the target HR range. The participant was in the prescriptive condition.

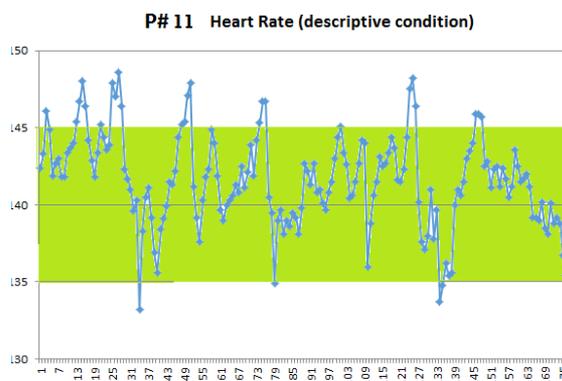


Figure 4: One participant’s HR (bpm) over time. The shaded area represents the target HR range. The participant was in the descriptive condition.

Figure 3 shows that the participant easily manages to return back into the target range within 5-10 seconds or 1-2 HR samples in the prescriptive condition. Compare this to Figure 4 where each time the participant’s (in the descriptive condition) HR exceeds the target range, there is an additional 3-5 bpm spike in HR, and it takes longer for that participant to return to the target range (on average, 3-4 HR samples or 15-20 seconds). To further analyze this trend, the average time per target zone departure was calculated for both above the target range. For each time a participant left the target HR zone, the number of samples was counted until the participant returned to the target zone. The sum of samples in each departure divided by the number of departures in a session gives an estimate of the average time spent per departure for that individual.

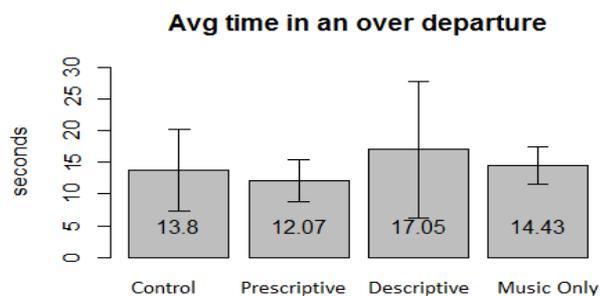


Figure 5: Average time spent in an “over departure” per condition.

Analyzing the data in this fashion gives an estimate of reaction time when exposed to different types of feedback. Reaction time in this sense is how quickly a participant reduces their heart rate back into the target range after exposure to a tempo manipulation. Again, the 95% confidence intervals in each group are overlapping, suggesting no significant differences. However, there is a trend of participants in the descriptive condition needing an extra 5 seconds on average to bring their heart rate down to the target range compared to participants in the prescriptive condition. In other words, participants found the prescriptive sonification mapping technique to be more intuitive than the target matching descriptive approach. This lends further support to the idea that it is people’s natural inclination to follow the rhythm of the music, as opposed to descriptive mapping where the tempo reflects the user’s current HR.

4. CONCLUSIONS

The results of the study suggest two main points. The first is that the change in music tempo served an effective motivational element, perhaps more than the actual tempo of the songs we selected. The second conclusion is that prescriptive sonification mapping (what you should be doing) is more intuitive than the descriptive mapping (what you are doing). This is probably explained by people’s natural inclination to synchronize their movements (or generally, their activity level) to the speed of the music, not the other way around. Subjective feedback from the semi structured interviews indicated that participants enjoyed the novelty of hearing familiar songs at unfamiliar speeds. Additionally, participants reported that they enjoyed the “game aspect” of controlling the playback speed of the music with their physical activity output.

The main limitation of this study is the small sample size (six per condition). Perhaps a significant difference could be found between conditions if more participants were included (if there truly is one to be found). Level of athleticism could have biased our results as a confounding variable that was not controlled for. The self-reported athletes in the study seemed to have an easier time controlling their heart rate (due to prior experience with target heart rate zones). Either way, a follow up study should be employed utilizing a within-subjects design with more participants, including a baseline condition with no heart rate feedback whatsoever to compare to.

Based on this small study, the use of music tempo manipulation as feedback on physical activity has shown to be as effective (and more enjoyable) than visual feedback. All 26 participants agreed that for guiding a user towards a target

heart rate range, the mapping utilized in the prescribing condition was most appropriate. This type of auditory feedback would be most helpful for exercises such as running or cycling, where it may disrupt the user’s flow to pull out a smartphone or watch and visually check for current heart rate. Some participants suggested that the opposite mapping (the type used in the describing condition) could be effective for achieving short bursts of high activity, as required for sprinting and weight lifting. If people’s intuitive reaction to an increase in tempo is to increase physical output, perhaps this mapping could result in a positive feedback loop, where the sensor detects an increase in heart rate (associated with the start of a sprint or end of a lifting set), and the system increases the tempo of the music resulting in an extra burst of physical output by the user. Alternatively, the same mapping could create a negative feedback loop where a decrease in physical activity is required, such as yoga or meditation. Further research is needed to validate this type of sonification in these specific circumstances. For instance, a few participants reported that they listened to podcasts or audiobooks while exercising. A future usability study could use similar sonification mapping techniques applied to spoken word audio to see if it would have a similar effect on exercise performance.

A version of the Tempo-Fit Heart Rate application is in the process of being made for android smartphones using the libPD library for playback speed manipulation. It will communicate with a variety of popular wearable HR monitors that use the BluetoothHealth API (Polar, Garmin, etc.). Once completed, it will be uploaded to the Google Play market for free and its source code released on GitHub.

5. REFERENCES

- [1] <http://www.cdc.gov/physicalactivity/data/facts.html>
- [2] J. Edworthy and H. Waring, “The Effects of music tempo and loudness level on treadmill exercise,” *Ergonomics*, 2006, pp. 1597-1610.
- [3] H. Wårnegård, “Heart rate sonification-using sound to monitor heart beat when running,” *student thesis*, 2012.
- [4] C. Karageorghis, I. Costas, L. Jones, and D. Low. “Relationship between exercise heart rate and music tempo preference,” *Research Quarterly for Exercise and Sport*, 77.2 (2006): 240-250.
- [5] S. Nirjon, R. Shahriar, R. Dickerson, Q. Li, P. A. John, A. Stankovic, D. Hong, B. Zhang, X. Jiang, G. Shen, and F. Zhao. “Musicalheart: A hearty way of listening to music,” *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. ACM, 2012.
- [6] N. Oliver, and F. Flores-Mangas. “MPTrain: a mobile, music and physiology-based personal trainer,” *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*. ACM, 2006.
- [7] C. Bauer, and A. Kratschmar. “Designing a Music-controlled Running Application: a Sports Science and Psychological Perspective,” *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2015.
- [8] <http://www.equivital.co.uk/products/tnr>
- [9] <http://www.acsm.org/access-public-information/articles/2012/01/13/the-heart-rate-debate>

THE TRIPLE TONE SONIFICATION METHOD TO ENHANCE THE DIAGNOSIS OF ALZHEIMER'S DEMENTIA

*Letizia Gionfrida*¹, *Agnieszka Roginska*², *James Keary*², *Hariharan Mohanraj*², *Kent P. Friedman*³

¹School of Computer Engineering
Nanyang Technological University
Singapore
lgionfrida@ntu.edu.sg

²Music and Audio Research Lab
Music Technology
New York University
New York, NY, USA
{roginska, jpk349, hari.m}@nyu.edu

³Division of Nuclear Medicine
Department of Radiology
NYU School of Medicine
New York, NY, USA
kent.friedman@nyumc.org

ABSTRACT

For the current diagnosis of Alzheimer's dementia (AD), physicians and neuroscientists primarily call upon visual and statistical analysis methods of large, multi-dimensional positron emission tomography (PET) brain scan data sets. As these data sets are complex in nature, the assessment of disease severity proves challenging, and is susceptible to cognitive and perceptual errors causing intra and inter-reader variability among doctors. The Triple-Tone Sonification method, first presented and evaluated by Roginska et al., invites an audible element to the diagnosis process, offering doctors another tool to gain certainty and clarification of disease stages. Audible beating patterns resulting from three interacting frequencies extracted from PET brain scan data, the Triple-Tone method underwent a second round of subjective listening test and evaluation, this time on radiologists from NYU Langone Medical Center. Results show the method is effective at evaluation PET scan brain data.

1. INTRODUCTION

Medical imaging techniques such as X-rays, magnetic resonance imaging (MRI), and positron emission tomography (PET) have drastically aided the diagnosis of medical conditions. Such techniques provide physicians with multi-dimensional and time-varying data sets that continue to increase in precision and reso-

lution. These techniques rely heavily on visual displays and visual based analysis for the principal method of evaluation and subsequent diagnosis. The techniques listed above convey large amounts of complex data via medical imaging. Although diagnostic accuracy is higher than ever, clinicians continue to struggle when the visual analysis of the information provided leads to imperceptible differences between health and disease. It is a possibility that for the diagnosis of AD, visual imaging techniques of such complex data need supplementary information to aid diagnosis. The Triple Tone method aims to enhance clarity in the diagnosis process, particularly in regards to diagnostic accuracy and inter-observer variability.

2. BACKGROUND

The Triple Tone method proposed by Roginska et al. (2013) generates a sonification represented by three separate tones that correspond to three regions of the brain: the sensorimotor cortex (SMC), frontal lobe (FL), and parietal lobe (PL). Metabolic activity of each region maps to frequencies that create the tone. The SMC has a fixed tone of 440 Hz while the FL and PL tones are the relative standard deviation with respect to the SMC in the positive and negative directions. This difference in tones creates a beating pattern, which would be more perceptible for the listener to hear. Resulting sonification models were initially tested for accuracy on listeners without medical backgrounds. Participants categorized the sonification samples with both coarse and fine grained levels, which represented the severity of the different cases of AD. For instance, coarse categorization, on one side, involved four levels that ranged from severities (normal, mild, moderate, and severe), on the other, fine categorization involved twice as many levels by



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

splitting each of the coarse levels in half. This test showed that listeners could more accurately categorize the correct diagnosis with finer gradation.

2.1. Current issues in the diagnosis of Alzheimer's Dementia

Alzheimer's disease (AD) is a neurodegenerative disease that affects the frontal and parietal regions of the brain. Recently, a multiplicity of imaging techniques, including positron emission tomography (PET) studies of cerebral metabolism have shown characteristic changes in the brain of patients with AD. Particularly, when a PET scan is used on a brain affected by AD, a decrease in metabolic activity of the above cited regions can be found, as shown in Figure 1. PET scan datasets are displayed visually with MIM software, throughout which a radiologist can compare the brightness of the affected regions to the sensorimotor cortex, a region unaffected by the disease. The greater is the difference in brightness the more severe will be the diagnosis. To be fair, this is an oversimplification of the process; however, such a technique does rely heavily on the visual process. The downsides of this can be seen in the intra and inter reader variability when it comes to spotting the early stages of the disease's progression.

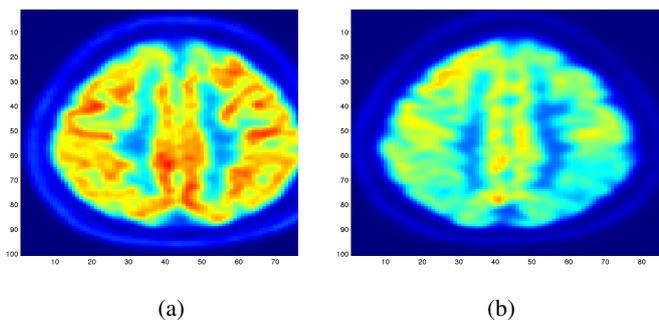


Figure 1: Example of orthogonal PET scan brain images of (a) a brain not affected by AD and (b) a brain severely affected by AD.

2.2. A review of Triple-tone sonification

The Triple-Tone Sonification method leverages the ear over the eye. For example, if presented with two shades of red one frequency apart on the color spectrum, the human eye will have a hard time determining the subtle difference between them. However, the human ear has a distinct advantage in that it hears two close frequencies beat against each other in a phenomenon known as beating. As thoroughly outlined by Roginska et al. (2014), complete with mathematical equations, the triple tone sonification method takes PET scan data-sets and assigns a single audible frequency value to the frontal, sensorimotor cortex and parietal regions. The sensorimotor cortex is given a base frequency of 440 Hz, and the other two regions are deviations from that frequency. These deviation values are determined by how far the overall metabolic regions activity deviates from the sensorimotor cortex. When these 3 frequencies are sonified, the differences between the regions are audibly heard as rhythmic beating patterns.

3. METHODOLOGY

3.1. Research structure and goals

As shown by Roginska et al. (2014) preexisting works, the previous round of testing validated the technique, proving its capability to provide accurate categorizations of the beating patterns. However, that testing round was done on the trained musical ears of audio professionals. The goal of this second round of testing was to determine how the Triple-Tone sonification technique affected the diagnoses process of those who would theoretically use it: physicians in the field of radiology. Thus, the methodology by which we tested the radiologists was informed accordingly. The goals of this round of testing were to determine the following:

- The accuracy and consistency of diagnosis by musically untrained ears (physicians).
- The extent of influence of the sonification in the diagnosis of a brain scan, for both experienced and inexperienced physicians.

The learning curve associated with this sonification technique on untrained ears, for both experienced and physicians with less experience.

3.2. Dataset preparation

The datasets used were the same as those in the first round of testing: 32 de-identified PET/CT scans of human brains diagnosed with varying stages of AD. Obtained from the Radiology Department of New York University Langone Medical Center, these 32 brains scans consisted of 8 brain scans in each of the four categories of diagnosis of AD; Normal, Mild AD, Moderate AD, and Severe AD. Only one slice per brain scan data set was used for sonification. The datasets were prepared in the same manner as the first round of testing as well. All datasets were spatially warped to a standard brain model using the MIM software package (www.mimsoftware.com). The spatial normalization process was necessary so that all 32 datasets used in testing were spatially consistent. This made it possible to sonify the same two dimensional subset of data points within each three dimensional dataset by simply choosing the same lateral slice each time for sonification. The next step was lobe segmentation. The MIM software was used to determine the boundary coordinates of the lobes within the dataset. The datasets were then output from the MIM software as DICOM standard RTSTRUCT file format, and input into SoniScan, our primary data analysis tool developed for the purpose of this research. SoniScan chose the 30th lateral slice of the spatially normalized datasets, segmented the lobes into the three regions to be sonified based on the given boundary coordinate points, and removed all irrelevant data points from the sub datasets. The 30th slice was chosen because it passes through representative regions of the frontal lobe, parietal lobe, and sensorimotor cortex. Thereafter, the 3 regions were mapped to a single frequency value each. The full frequency mapping process is outlined in detail by Roginska et al., after this process, SoniScan output 9 frequency values, 3 for the 3 regions in the left hemisphere, 3 for the 3 regions in the right, and 3 for both left and right together. AD can be asymmetrical in where it acts, it can happen in the left, right or both hemisphere. Thus it was necessary to dissect these regions by hemisphere, and choose the 3 values that demonstrated the most deviation, drawing attention to the abnormalities. Otherwise it was

possible that, for example, an asymmetrical severe case would be heard as mild or moderate when the left and right hemispheres were sonified together. A sonification software tool was developed for this research project using the sound synthesis engine and programming language SuperCollider. The 9 frequency values were input into this program for interactive testing of the sonifications and output audio recordings. The final results were 32 30 second audio files, one for each PET scan.

4. EVALUATION

4.1. General procedures

The Triple-Tone sonification technique was tested by one highly experienced and one less experienced radiologist from the New York University Langone Medical Center. Testing took place over four sessions, each session included diagnoses of the same 32 de-identified PET scan cases. The order of presentation of the sonifications was consistently randomized across diagnoses sessions to minimize recall bias. The sessions were set up to mimic the diagnosis process that these radiologist would normally undertake in which the radiologists would diagnose the left hemisphere's frontal lobe, the left's parietal lobe, the right hemisphere's frontal and the right's parietal. The radiologists were given no information about the test case patient's gender, age or medical history. For each session, the diagnoses were recorded as well as the time it took for the full diagnosis of each patient. The first session was a basic visual diagnosis where the radiologists used the MIM software to search through the 2D slices of each test case. The second session was the advanced visual diagnosis session, which saw the addition of the 3D projection views of the brain. 3D projection views are generated by the MIM software. The data points in the 3D projection view represent the average intensity of voxels along the line of view from an external perspective, directly left of and directly right of the brain, to a depth of three millimeters. Physicians utilize this view of the brain to assess the overall picture regarding the diagnosis of a brain. The third session was the basic visual diagnosis with sonification, and the fourth session was the advanced visual diagnosis with sonification. All sonifications were played through Sennheiser HD650 headphones.

4.2. Sonification Training and Testing

The sessions with sonifications saw the addition of a training session. The training was exactly the same for both sessions. The first step of the training was to verbally introduce the radiologists to the general idea of sonification. The next step was to present them with specific examples of PET scan sonifications. To do this, 2 training case sonifications were played for each of the 4 AD diagnosis categories, 2 normal, 2 mild, 2 moderate and 2 severe. These sonifications were played for 30 seconds while the radiologist looked at the lateral slice related to that particular sonification. This was done to introduce them to the sound of each AD diagnosis category and to closely match the form of the test itself. The radiologists were informed that the sounds they were hearing were sonifications of a 2D lateral slice of the region of the brain where the greatest deviation of metabolic activity occurred. The next part of the training was the introduction of a graphical user interface tool that the radiologists could use during the diagnosis process. Created in Matlab, the tool consisted of an interface of 40 buttons divided into 4 columns; Normal sonifications, Mild AD

Questions	Physician 1 (High Exp.)	Physician 2 (Low Exp.)
The sonification was tiring to listen to and induced fatigue.	2	1
The sonification was pleasant.	2	2
The sonification provided additional information that the visual display did not.	3	5
The sonification was helpful in discerning between:	-	-
a) normal and mild cases	5	4
b) mild and moderate cases	5	4
c) mild and severe cases	4	4
The sonification made me more certain about my diagnosis.	4	4

Table 1: Physician questionnaire results

sonifications, Moderate AD sonifications and Severe AD sonifications. Each column had 10 buttons that would each play a sonification from that column's category. Ten sonifications were generated for each category, resulting in a total of 40 training cases. The 40 training cases were generated from the spectra of the deviation values of the 32 test cases. Specifically and for example, the 8 severe test cases had deviation values that went across a range. The 10 severe training cases were generated on average from that range. The ranges of the categories were displayed in order through the Matlab tool, where the smallest deviation in that category was at the top of the column and the largest was at the bottom. The radiologists were given a few minutes to test out the interface in order to familiarize themselves with more sonifications from each category as well as familiarize themselves with the tool they would be using in the diagnosis process. Directly after the training session, the radiologists were told to go about their diagnosis process as they would normally do, searching through the 2D slices and 3D projections visually. Only this time, the 30 second sonification would play along with the visuals. The radiologist could control the playback and volume of the sonification. They were also given the ability to play the training cases, by clicking the buttons of the Matlab tool to do AB comparisons with the brain scan sonification.

5. RESULTS

5.1. Physician questionnaire

After the final session, the radiologists were given a brief survey. They were asked a short series of questions on if they found the sonification helpful in the diagnoses process. Their answers were given on a scale from 1 to 5, where 1 = strongly disagree, 2 = disagree, 3 = somewhere in the middle, 4 = agree, and 5 = strongly agree. As shown in Table 1, the radiologists found the sonification

Physician 1 (High Experienced)

Sess.	LF	LP	RF	RP	W
0 Basic Visual	0.64***	0.84***	0.56**	0.8***	0.88***
1 Adv. Visual	0.39*	0.76***	0.37*	0.74***	0.81***
2 Basic Visual + Sonif.	0.69***	0.94***	0.66***	0.94***	0.98***
3 Adv. Visual + Sonif.	0.68***	0.79***	0.67***	0.81***	0.86***

Physician 2 (Low Experienced)

Sess.	LF	LP	RF	RP	W
0 Basic Visual	0.58***	0.70***	0.54**	0.64***	0.87***
1 Adv. Visual	-0.15	0.00	-0.01	0.12	0.03
2 Basic Visual + Sonif.	0.61***	0.89***	0.51**	0.83***	0.93***
3 Adv. Visual + Sonif.	0.58**	0.85***	0.48*	0.79**	0.95***

Table 2: Pearson’s correlation for association with the ground truth, ***significant at $p < 0.001$, ** significant at $p < 0.005$, * significant at $p < 0.5$

technique, although not pleasant, helpful in the diagnosis process, especially when it came to discerning between degrees of AD and in solidifying the certainty of their overall diagnosis.

5.2. Statistical Analysis results

During each evaluation, both physicians took part in four distinct sessions, basic and advanced visual diagnosis with and without sonification. For each one of the 32 unidentified brain scans of patient, eight cases per each severity level (normal, mild, moderate, severe), physicians analyzed four part of the brain: left frontal (LF), left parietal (LP), right frontal (RF), right parietal (RP). The results were then compared to the ground truth provided by Dr. Kent P. Friedman.

5.2.1. Regional Scores

For each combination of two physicians and four sessions there is a significant positive correlation between the ground and the

Physician 1 (High Experienced)

Sess.	LF	LP	RF	RP
Basic Visual	0.68	0.86	0.64	0.84
Advanced Visual	0.57	0.97	0.55	0.91
Basic Visual + Sonification	0.74	0.99	0.71	0.96
Advanced Visual + Sonification	0.87	0.96	0.85	0.92

Physician 2 (Low Experienced)

Sess.	LF	LP	RF	RP
Basic Visual	0.64	0.86	0.59	0.75
Advanced Visual	0.57	0.97	0.55	0.91
Basic Visual + Sonification	0.61	0.96	0.48	0.84
Advanced Visual + Sonification	0.52	0.85	0.44	0.77

Table 3: Pearson’s correlation for association with worst scores

frontal, parietal and worst regions. This is excepted for the scores reported by the less experienced physician during the advanced visual diagnosis (shown in Table 2), where there is no significant correlation between the ground truth provided and the regional or worst. Comparing basic and advanced inspections with and without sonification a significant increase of the correlation is shown. Since the highest significant correlation was always between the ground truth and the worst scores, accuracy was assessed in terms of concordance between the truth and worst results. Therefore

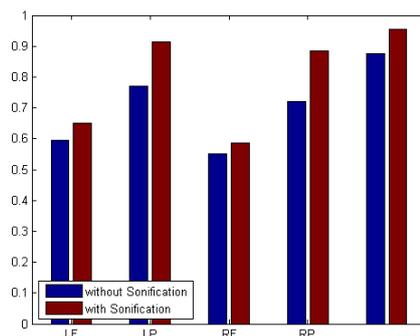


Figure 2: Basic visual with and without sonification for the two physicians.

the correlation for the association of the regional scores with worst scores is presented in Table 3. All correlations were significant ($p < 0.01$). The worst scores were most highly correlated with the

Sessions :	<i>Phys.1</i>	<i>High Exp.</i>	<i>Phys.2</i>	<i>Low Exp.</i>
	Lower	Upper	Lower	Upper
Basic Visual	41.9%	74.7%	41.9%	74.7%
Adv. Visual	41.9 %	74.7%	13.0%	42.6%
Basic Visual + Sonification	78.2%	98.4%	64.1%	91.3%
Adv. Visual + Sonification	64.1%	91.3%	67.5%	93.3%

Table 4: Concordance percentage of ground truth and worst scores and left parietal scores

scores from left parietal region, highlighted in bold. According to these results, looking at the parietal elements highlights more efficient scores. Thus, this result suggest that the worst scores may be approximated using only parietal region scores.

5.2.2. Accuracy Percentage

Logistic-Regression was used to characterize and compare sessions in terms of the accuracy of the worst scores for each subject. Accuracy was defined as the percentage of times the worst scores were concordant with the ground truth. Generalized estimating equations (GEE) is generally used to estimate the parameters of a generalized linear model with a possible, unknown, correlation between outcomes and it is based on binary logic regression. In this work it has been used as a measure of agreement between the ground truth and the worst and left parietal region. The lack of statistical independence among results from the same subject (i.e., results for the 4 sessions) was accounted for by assuming results symmetrically correlated when acquired from the same subject and independent when derived for different subjects. Output included a 95% confidence interval (CI), an estimate that calculate interval-*sof* possible values of unknown parameters, for the percentage of times each session produced worst scores that were concordant with the ground truth. The percentage % and number (#) of times the worst scores and the left parietal scores were concordant with the ground truth, is shown in Table 4.

6. DISCUSSION

A sonification method throughout which medical researchers have access an additional way of interpreting data has been presented. This method can assist more physicians detecting something that may be missed with traditional visual inspection. To evaluate this technique, the correlation between each evaluation method has been proposed in this study and the correct diagnosis based on the ground truth has been computed to measure whether sonification helped improve accuracy in the diagnosis. Results from the statistical analysis show a higher correlation between tests using sonification and the ground truth. During the basic visual testing, one physician achieved an accuracy of 56-88% while the other performed at an average of 50-55%. For each session, the physicians performed more accurately with an average improvement of about 10% using the added sonification method. For the more severe

cases, the improvement in accuracy was as high as 30% with sonification added. The results of the evaluation show that subjects with untrained ears are able to categorize the sonifications generated by the triple-tone technique with accuracy.

7. FUTURE WORK

Future purpose for the Triple-tone sonification method will extend its data scope to spatial representation of the audio. It is theorized that by sonify by one slice to a larger subset of the brain, a more accurate description of the brain will be reached. Furthermore, by extending the spatial scene it may provide more information from the data. Each DICOM data set is a 3 dimensional matrix that contains 60 slices portraying the full brain. In the current triple tone sonification method one subset of each dataset was chosen for sonification: the 30th lateral slice (from the top) of the spatially normalized dataset. This slice was segmented into three lobes of interest (frontal and parietal) and the reference lobe (sensorimotor cortex). After data was extracted from each relevant lobe and frequencies per lobe were allocated, physicians were presented with a mono audio file, representative of the 30th slice of DICOM data, sonified by three frequencies. The beating patterns created by the three frequencies reveal the severity of the case. To extract more slices from the DIACOM data set, modifications to the SoniScan program were implemented. SoniScan is the C++ program used to perform slice selection, lobe segmentation and aid calculations for frequency allocations. Due to the physical makeup of the brain, coordinates for lobe segmentation will be individual to each slice. Segmentation information was performed with the aid of the MIM software. Each slice and its relevant coordinates were placed into a 2D matrix for the SoniScan program to read as per user defined slice number. The contours were approximated to straight lines connecting the coordinates as in the original program. Currently, physicians are presented with a mono audio file containing the three frequencies. To better represent this information, binaural renditions of case studies have been created. The binaural examples allow for multi-slice playback, and individual lobe isolation, all of which are user defined via a GUI. The frontal lobe is spatially placed at azimuth: -5 to +5 and parietal lobe at azimuth 175 to -175. Elevation of each slice is determined by how many slices there are during playback. The more slices rendered, the closer each slice will be to one another in elevation. Distance of each sound source is 1 meter. The Head Related Impulse Responses used to for the convolution process were extracted from the MARL-NYU file format for storing HRIRs.

8. CONCLUSIONS

Initial validation for the the triple-tone sonification method has been presented in this paper. The promising results obtained from this study highlight more possibilities in helping AD diagnosis. In fact, through spatialization analysis of the resulting slices and their sonification using this technique, radiologists can get more information to investigate a larger region of the brain. Given a 3D representation aurally also presents an opportunity to create a virtual space for the brain with which the listener can interact. The next steps involve possibly working with 3D audio and enhancing the analysis and/or experience of diagnosing AD. Furthermore this sonification technique can be enlarged to existing visualization imaging techniques, such as MRI, in order to bring considerable improvement to medical diagnosis area. Therefore, the proposed

method is broadly ready and now applicable to a number of different methods with a promising future.

9. ACKNOWLEDGMENT

This work was supported by the NYU CTSI Award Number NIH/NATS UL1 TR000038.

10. REFERENCES

- [1] G. Baier, T. Hermann, The Sonification of Rhythms in Human Electroencephalogram, in Proc. of the 10th Int. Auditory Display (ICAD 2004), Sydney, Australia, July 2004.
- [2] G. Baier, T. Hermann, U. Stephani, Multi-channel sonification of human EEG, in Proc. of the 10th Int. Auditory Display (ICAD 2007), Montreal, Canada 2007.
- [3] R. S. Frackowiak, C. Pozzilli, N. J. Legg, G. H. Du Boulay, J. Marshall, G. L. Lenzi, Regional Cerebral Oxygen Supply and Utilization in Dementia, *Brain* (104), 753-778.
- [4] J. W. Piper, Validation of an Artificial Neural Network Computer-Aided Diagnosis Scheme for Alzheimer's Disease using FDG-PET, Wake Forest University, Department of Computer Science, 2007.
- [5] J. Oh, Y. Wang, A. Apte, J. Deasy, A Statistical and Machine Learning-Based Tool for Modeling and Visualization of Radiotherapy Treatment Outcomes, *Medical Physics* (2012), 39 (6), 3763.
- [6] A. Roginska, K. P. Friedman, H. Mohanraj, Exploring Sonification for Augmenting Brain Scan Data, in Proc. of the 19th Int. Conf. on Auditory Display (ICAD 2013), Lodz, Poland, July 2013.
- [7] T. Kagawa, S. Tanoue, H. Kiyosue, H. Nishino, A Sonification Method for Medical Images to Support Diagnosis Imaging, in Proc. of the 7th Int. Conf. on Complex, Intelligent, and Software Intensive Systems (CISIS), Taichung, China, July 2013.
- [8] A. Roginska, H. Mohanraj, M. Ballora, K. P. Friedman, Immersive sonification for displaying brain scan data, in HEALTHINF 2013, 6th International Conference on Biomedical Electronics and Device, Barcelona, Spain, February 2013.
- [9] A. Roginska, H. Mohanraj, J. Keary, K. P. Friedman, Sonification Method to Enhance the Diagnosis of Dementia, in Proc. of the 20th Int. Conf. on Auditory Display (ICAD 2014), New York, New York, June 2014

MUSICAL ROBOTS FOR CHILDREN WITH ASD USING A CLIENT-SERVER ARCHITECTURE

Ruimin Zhang

Jaclyn Barnes

Joseph Ryan

Michigan Technological University Michigan Technological University Michigan Technological University
Houghton, MI 49931 Houghton, MI 49931 Houghton, MI 49931
ruiminz@mtu.edu jaclynb@mtu.edu jdryan@mtu.edu

Myounghoon Jeon

Chung Hyuk Park

Ayanna M. Howard

Michigan Technological University George Washington University Georgia Institute of Technology
Houghton, MI 49931 Washington, DC 20052 Atlanta, GA 30332
mjeon@mtu.edu chpark@gwu.edu ayanna.howard@ece.gatech.edu

ABSTRACT

People with Autistic Spectrum Disorders (ASD) are known to have difficulty recognizing and expressing emotions, which affects their social integration. Leveraging the recent advances in interactive robot and music therapy approaches, and integrating both, we have designed musical robots that can facilitate social and emotional interactions of children with ASD. Robots communicate with children with ASD while detecting their emotional states and physical activities and then, make real-time sonification based on the interaction data. Given that we envision the use of multiple robots with children, we have adopted a client-server architecture. Each robot and sensing device plays a role as a terminal, while the sonification server processes all the data and generates harmonized sonification. After describing our goals for the use of sonification, we detail the system architecture and on-going research scenarios. We believe that the present paper offers a new perspective on the sonification application for assistive technologies.

1. INTRODUCTION

One of the most important characteristics of people with Autism Spectrum Disorders (ASD) comes from difficulties in social communication. ASD may not be diagnosed until adolescence or later, whereas the symptoms appear in early childhood. Children with ASD often have trouble with fundamental social skills, such as turn taking, eye contact, and joint attention. Particularly, their problems in emotional communication are a

serious barrier to their social inclusion. Given that the use of interactive robots is known to have high potential for enhancing social skills of children with ASD [1-5] and that music and sound are known to be effective to induce, deliver, and regulate emotions [6], we aim to design robotic sonification for children with ASD. While there has been active research on voice communication between children with ASD and robots [7-8], little empirical research has focused on sonification robots. Sonification is defined as the use of non-speech audio signals to convey information [9]. By definition, sonification can include all types of non-speech sounds, such as musical sounds, natural sounds, sound effects, even noise, etc. With all these sounds in mind, we have investigated the effectiveness of robotic sonification for children with ASD in a more systematic way. We envision not only interaction between a child and a robot, but also interactions between multiple children and multiple robots. To this end, we have developed a client-server architecture. Each robot communicating with a child acts as a client and a sonification server processes all the data and returns the action command to the clients. Our robotic sonification system is expected to facilitate emotional and social communications of children with ASD using music and real-time sonification, generated based on a child's emotional states and physical interaction patterns with robots.

2. APPROACH TO ASD INTERVENTION

While waiting for more effective treatment for ASD, many remedies with diverse disciplines have been applied, including pharmaceutical research, behavioral therapy, and complementary and alternative medicine [10-11]. Here, we focus on an interactive robot approach and music therapy approach.



Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** First author, Second author, Third author.

Web Audio Conference WAC-2016, April 4-6, 2016, Atlanta, USA

© 2016 Copyright held by the owner/author(s).

2.1. Interactive Robot Approach

Some people with autism prefer to communicate with and through computers because they are predictable and place some control on the otherwise chaotic social world [12-13]. While keeping the advantages of using computers, researchers have attempted to use robots for children with ASD to help social interaction skills [1-5]. In addition to their basic computing ability, interactive robots provide a sensing and detecting environment, and save interaction data. Since robots play a role as a toy, we can also expect that psychological “rapport” between children and robots may be well formed.

From the previous robot approach, we can learn metrics they successfully used. For example, research has shown that an interactive robotic dog (AIBO) enhances social interaction more than a mechanical toy dog (Kasha), which has no ability to detect or respond to its physical or social environment [4]. Researchers found that children with autism (aged 5-8 years) spoke more words to the AIBO, in comparison to the Kasha. AIBO also elicited more social behaviors such as verbal engagement (e.g., salutations, valedictions, and general conversation), reciprocal interaction (e.g., motioning with arms, hands, or fingers to give direction to the artifact), and authentic interaction (e.g., the speed, tone, and volume of the child’s voice is exceptionally well modulated for the circumstances, or the child’s body is in a state of repose oriented toward the artifact as a social partner). Moreover, as the children interacted more with AIBO, they engaged in fewer autistic behaviors such as rocking back and forth and flicking fingers or hands. Scassellati and his colleagues [8] have been working on affective prosody production learning of children with autism using Pleo, a dinosaur robot. Their pilot study was conducted with 9 to 13-year-old children with High Functioning Autism. The children had to use an encouraging tone of voice to have the robot move across areas of “water”, which they are told the robot fears. Children in these trials have shown increased appropriate prosodic production with the robot trials as opposed to equal sessions with a human instructor.

2.2. Music Therapy Approach

Children with ASD often show enhanced musical ability [14], though we acknowledge that it is not always the case. Music therapy is a broad category that encompasses diverse techniques that are frequently mixed and combined including receptive, recreative, compositional, improvisational, and musical activity therapies [15]. This variety demonstrates that drawing a general conclusion about the effectiveness of music therapy is not simple, but meta-analysis of existing studies has shown positive effects of music therapy [14-15]. We expect the use of various types of sonification (including music) in interactive robotic sessions will result in as effective outcomes as the traditional music therapy sessions or better outcomes than those.

3. GOALS OF ROBOTIC SONIFICATION

A human brain has four different anatomical functions of emotions, which include appraisal, reactivity, emotional understanding, and regulation [16]. We believe our robotic

sonification can enhance all of these emotional aspects of children with ASD.

3.1. Enhancing Appraisal

Individuals with ASD typically show limited engagement in expressing their own emotional states. This might be because they have some issues in appraisal of their states and/or they are not good at expressing their states actively. Serving as external stimuli, emotional music and sound will help children with ASD appraise their emotional states. Thus, our robotic sonification system is expected to boost children’s ability to appraise their own emotional states.

3.2. Enhancing Reactivity/ Expression

Once they are more capable to assess their emotional states, we can expect they are likely to express their emotions. However, appraisal and reactivity might be an independent process. Once our emotion detection system can estimate the children’s emotional states, the sonification system can easily help them express their emotions because music and sound deliver mood and emotions very well. It could be done in a conscious way (e.g., happy music or laughing sounds for the happy state), but it could also be done in an ambient way (e.g., slowly reflecting their mind state using the soundscape).

3.3. Enhancing Emotional Understanding

We anticipate that our robotic sonification system will also serve as a positive reinforcement of the mapping between social situations (e.g., happy or disappointed) and the sounds (e.g., laughter or a sigh) in the session. Therefore, children with ASD can make more strong associations between the situations and appropriate situational sounds. It can further help them learn and estimate the robot’s (or other person’s) emotional states, which might lead to forming a general capability of empathizing others. To this end, we have designed over 600 non-speech sound cues with about 30 emotional states [17]. Our plan is to construct personalized sound-emotion mapping dimensions based on subsequent mapping experiments.

3.4. Enhancing Emotion Regulation

Children with ASD have a variety of symptoms (e.g., tantrum vs. social isolation). Given that music is frequently used as an intervention for emotion regulation, if appropriately used, sonification might improve some children’s emotion regulation. However, for other children with ASD, specific stimulus can make their symptoms more severe. This is why we need to create an individualized sensory stimuli map.

3.5. Enhancing Monitoring of Interaction

From the perspective of a family member, teacher, or clinician, adding sonification can provide an additional way of monitoring of children’s interaction in the session. If the sonification is mapped so that the children are uniquely identifiable, they can monitor each child’s emotional states. Additionally, the

sonification can provide a representation of the interaction between robots and children overall, which will be helpful assessing the entire class or therapy session.

4. SYSTEM OVERVIEW

To obtain these therapeutic goals, we have designed robotic sonification platforms. We use a client-server architecture in our system (Figures 2 and 4) that can be connected with any robots and other sensing devices via TCP/IP sockets. The first prototype system can detect children’s emotional states during the interaction and play sonification adaptive to their emotions. The second system assumes that a child can interact with the multiple robots at the same time. All the activities and relationships between the child (or children) and multiple robots can be integrated in the sonification server and reflected in the sonification pattern. In this procedure, the sonification server can process different data, and control and harmonize different layers of the sonification in real-time. Depending on the purpose of the sonification, the server can decide which robot will generate the sounds or the server can also generate sonification (e.g., providing feedback to children vs. monitoring the emotional state of the child).

4.1. Prototype System Configuration

4.1.1. Non-humanoid Robot as a Client



Figure 1. Romo and facial expression detection app.

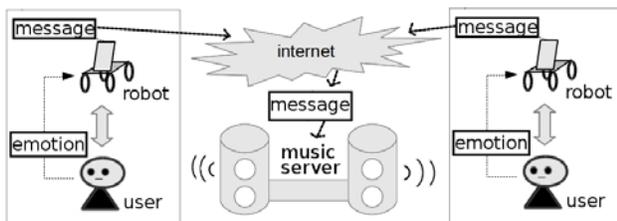


Figure 2. A client-server structure with Romo.

Our first prototype system uses Romo (Figure 1), which is composed of an iPhone and a trackmounted base. Romo can detect the child’s emotions at background and generate sonification on the fly in the robot or the data can be sent to the sonification server through the Internet and the server can generate sonification. The emotion detection system was implemented through an iOS application. This application

activates the camera included in the iPhone, which captures a video in real-time at a frequency of 15 frames per second, when a child interacts with Romo. When an emotional state is detected, a pink rectangle marks the relevant region in the video and a text label of the emotion is displayed on the left top corner of the rectangle (Figure 1). The detected emotions are recorded with the time they are captured. They can be encoded as messages sent to the server or rendered as a visual graph when the interaction finishes. We used the cascade training method provided by OpenCV [18] (an open-source library that focuses on real-time image processing) to train our customized classifiers for basic emotional states, including happy, surprised, and angry. The training images were from Cohn-Kanade AU-Coded Facial Expression database v2, which includes 592 sequences from 123 posers. Each sequence begins with a neutral expression and proceeds to a peak expression. When the child plays with Romo, our classifiers scan every frame of the video captured by the iPhone camera and send the data to the server.

4.2.2. Sonification Server

The sonification server was implemented with JFugue, which is an open-source Java API for programming music (e.g., MusicString, MIDI, audio file I/O, etc.). It allows us to specify notes, chords, musical instruments, and tracks to play a music piece based on our own algorithms. JFugue provides a class called Pattern. We used Pattern to create a block of sound/music (e.g., four or eight measures) for each interaction unit (e.g., an emotional state, physical activity, etc.), which we are interested in. The sonification server has an unlimited loop to listen to messages from the clients until it is manually stopped. It processes every message it has received and plays the corresponding sonification piece.

4.2. Current System Configuration

4.2.1. Humanoid Robots as Clients

While Romo was developed as the first prototype system, our long-term research plan involves using various robot platforms, including humanoid robots, non-humanoid robots, and animal robots, each of which requires different hardware and software. Again, this is why we use a client-server structure. We have been working with two robot platforms: ROBOTIS OP2 (formerly known as “DARWIN 2” or DARWIN-OP2”) and Nao (Figure 3), but here we focus on the description of ROBOTIS OP2.



Figure 3. ROBOTIS OP2 and Nao robots for this project.

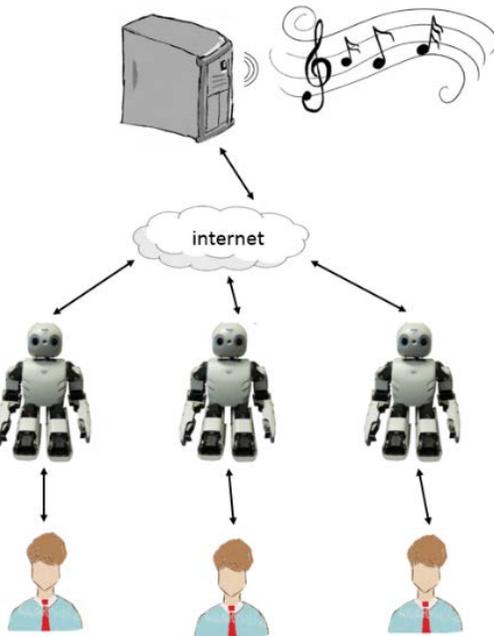


Figure 4. A client-server architecture with ROBOTIS OP2.

Our second version client is ROBOTIS OP2, which is a PC-based humanoid research robot. ROBOTIS OP2 has convenient interfaces, such as a camera and microphone, which enable natural interactions between children and robots. ROBOTIS OP2 has two 6-DOF legs, two 3-DOF arms and one 2-DOF neck. These joints allow ROBOTIS OP2 to simulate a number of human movements, such as walking, falling, standing up, looking around, or waving. ROBOTIS OP2 has a built-in PC with an Intel Atom N2600 @1.6 GHz dual core CPU, 4G DDR3 RAM and 32 GB mSATA, and runs Linux operating system. It provides programmers with a familiar environment and decent computing power to developing applications utilizing ROBOTIS OP2's interfaces. However, compared to other computing devices, such as personal laptops, desktop computers, or commercial servers, ROBOTIS OP2 is low powered and has a particular problem for process that requires intensive computation. Such tasks may involve graphical or audio signal processing, artificial intelligence, and multi-threaded applications.

Therefore, we have designed a human-robot interaction system, using ROBOTIS OP2 as an autonomous terminal to interact with children through natural interfaces including sounds (speech and non-speech) and visuals, and passing computation-intensive tasks to the remote server. The result is returned by the server and used as a reference by ROBOTIS OP2 to decide its next interaction.

4.2.2. Sensing Devices as Clients

Even though each robot has multiple cameras, it is hard to monitor the overall interaction scene between children and robots, using those embedded cameras. Therefore, we have incorporated an RGB-D depth sensor (Microsoft Kinect) to monitor the physical activities of a child and a robot to estimate the social engagement. We also plan to use multiple Kinects and other sensing devices to estimate their interactions more effectively.

4.2.3. Integrated Sonification Server

Our system includes a depth camera, one or more terminal robots (e.g., Romo, ROBOTIS OP2, and the Nao, see Figure 3), and a sonification server. Again, the clients and the server are connected using the reliable TCP/IP protocol. One of the salient differences between the first version prototype (Romo) and the current version system (ROBOTIS OP2) is the enhanced and integrated functions of the server. While the iPhone of Romo processes image data and performs facial expression detection, ROBOTIS OP2 just plays a role as a terminal. The server uses multiple threads to handle connections with each terminal robot and the depth camera. For example, when interacting with children, the terminal robot transfers the data captured by its cameras to the server as a sequence of images. The server performs an analysis of each image to detect a face. A detected face is marked with a rectangle and the position of its center is returned to ROBOTIS OP2. When ROBOTIS OP2 is notified by the server that a face is detected, it rotates its head to the direction where the face is detected and greets the person by a pre-recorded script to introducing itself. If ROBOTIS OP2 does not detect any faces, it will search for a face by rotating its head from side to side.

On the other hand, to evaluate children's physical activities and social interactions based on the data from the depth camera, we have incorporated metrics from physical therapy and rehabilitation [19]. However, the metrics are usually defined verbally based on physical rehabilitation conditions, and the underlying equations for deriving the metrics are not clearly defined. Thus, for assessing participating children's gestures and small motions, we have determined from the literature that the best approach for our problem is to use the following metrics: range of motion (ROM), path length (PATH), peak velocity (PV), average velocity (AvgV), and movement units (MUs). Besides these parameters, we will also utilize the child's specific motions or gestures to train the robot to attract more attention from the child.

The integrated sonification server was programmed in Java, which enables us to access functionalities of our existing

JFugue platform easily. For example, the server can have a further analysis on the data it receives from the terminal robots and the depth camera, and uses the statistical data to generate an ambient music on-the-fly. For further research using diverse mappings, we specifically developed a sonification mapping platform in the integrated sonification server.

Sonification Mapping Platform

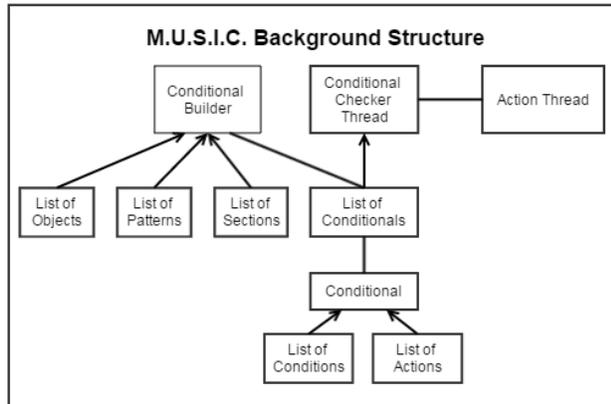


Figure 5. The M.U.S.I.C. platform’s background structure.

To make a more flexible, robust, and easy to use interactive environment for sonification mappings with the clients’ real-time data (e.g., children’s emotion data and physical movement data), we have developed the Musical Utility Software for Interactive Creations (M.U.S.I.C.) platform. The M.U.S.I.C. platform allows researchers to perform actions when the data meet created criteria in the form of conditions. The M.U.S.I.C. allows researchers to create simple one-on-one mappings between tracked emotion and movement data, and pre-defined MIDI notes or melody patterns. In the M.U.S.I.C. structure, researchers can make this mapping both dynamic and static, to load the program from launch or to make changes in real-time. The structure in the diagram (Figure 5) has been simplified to only describe the conditionals rather than the program as a whole. A conditional is a combined set of conditions and actions. When all conditions are met, all actions are performed in a separate thread. This allows for multiple conditionals to be run concurrently as well as maintaining a high speed. The conditionals are built in a conditional builder. The conditional builder has a reference to all of the current objects, patterns, and sections as well as a list of previously made conditionals.

All of these references are acting independently of the conditional builder, and thus, allowing changes to occur at any point in the development process. Conditionals have a flag that determines if the conditional is enabled or disabled. The conditional checker is a separate thread that has a reference to the conditional list and is checking all conditionals that are enabled. Therefore, the process looks as follows:

```
private void runConditionals()
for(int c = 0; c < conditionals.size(); c++){ //For each conditional
    if(conditionals.get(c).isEnabled()){ //Is the conditional enabled?
        boolean doActions = true; //Assume that we can do an action

        for(int con = 0; con < conditionals.get(c).getConditions().size(); con++){
            if(!conditionals.get(c).getConditions().get(con).testCondition()){ //Check to see if //the condition //did not get met
                doActions = false;
                break;
            }
        }

        if(doActions){ //Do all actions in the conditional list ...
            ...
        }
    }
}
```

Figure 6. The pseudo code of the conditional checker.

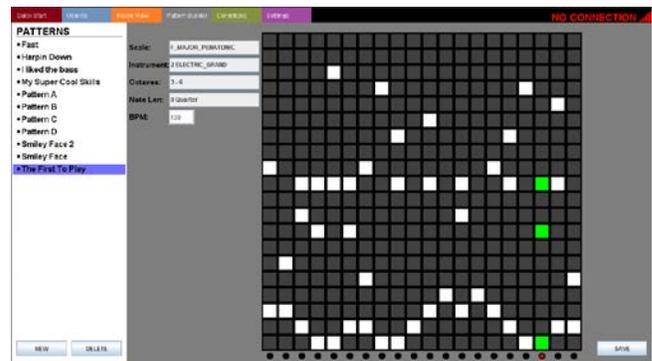


Figure 7. The pattern view tab.

This is the generalized pseudo code of the conditional checker (Figure 6). It checks every enabled conditional, and then checks all conditions within that conditional. Notice how the flag doActions is set true. This is so that in situations where there are no conditions, all actions can still be performed. In turn, this allows for chosen actions to always be performed. This code does run at $O(n^2)$ time. In the future, this conditional checker will be run by multiple threads so that runtime can be improved.

This structure is the reason the program works and allows for dynamic adjustments of conditionals. Although this structure may seem complex, the actual GUI is easy to understand (Figure 7). Researchers can make any patterns for each emotion or each movement of each child. Also, they can easily map the entire data stream with a certain sonification pattern. The structure makes it very convenient to add new conditions and actions. For example, when the algorithms for a new movement recognition are created (e.g., drawing a circle in a 3D space), it is just a matter of adding a new condition option for researchers to utilize.

We plan to optimize the code to further improve performance and run usability tests, such as heuristic evaluation or cognitive walkthrough to improve the research environment. Since this application is currently built in Java, we are exploring the option of moving this application to Android as well for better mobility. Being able to use the program on the go, and with extra additions of more robots and sensing devices to connect with the sonification server, could really take this program to the next level of usability and accessibility for the researchers and clinicians without a programming background.

5. RESEARCH SCENARIOS

To conduct therapeutic sessions with children with ASD, we have developed relationships with Autism Awareness Groups, local elementary schools, and the Center for Autism in three different sites (Michigan, Washington DC, and Georgia). As mentioned, we have developed our research scenarios based on the phases in music therapy: receptive, recreative, compositional, improvisational, and musical activity therapies. Some of our initial research scenarios are as follows.

5.1. Receptive Therapy

“Musical walk with a robot” is a type of receptive activity. For example, we selected Antonio Vivaldi’s “Four Seasons” as a thematic music, because it is a well-known classical piece with various themes and a sensational storyline, which can be smoothly related to different emotional sensations. Throughout the scenario, the robot will take the child to the four physical stages, each of which consists of different sounds and images evoked during spring, summer, fall, and winter seasons.

5.2. Musical Activity Therapy

As a musical activity scenario, we consider a traditional children’s game, “Dance Freeze”. With the child and the robot facing each other, music will be played and stopped at random (or pseudorandom) intervals. While the music is going, the players dance or move around the room. When it stops, they must freeze in place. If they move before the music starts again, the robot will play a disappointing sound. If they do not, a happy reward sound will be played, and then the music will start again for more dancing. The music can be varied to regulate the speed of the game and the arousal of the child. Depending on the aptitude of the child, it could be made more complicated by dictating what dance moves should be done and alternating who controls the game.

5.3. Recreative and Compositional Therapy

In addition to receptive or reactive scenarios, we also plan to conduct a recreative or co-creational sonification activity session. To illustrate, the robot will initially play a portion of music and perform actions (e.g., move or dance) correlated to the music. Then, the child can respond to the robot’s actions. The task will be interactive, such as kicking a ball toward each other or mimicking behaviors of each other. Each action sequence of the robot and the responses of the child will be translated into harmonized sonification. The music and the robotic actions will be built a priori according to our music-action mapping algorithm. The child’s approximate movement and action can generate melodious sounds. If the child does not respond to the stimulus, the robot can continue to act while dynamically changing the music based on the behavior, distance, or interaction patterns of the child. If the child responds to the stimulus, then the robot will take turns and lead the child by suggesting a gesture or task to enhance the sonification outcomes and the social interactions.

6. CONCLUSION AND FUTURE WORKS

To promote social and emotional interactions of children with ASD, we have designed an orchestration robot platform using a client-server architecture. Given that children with ASD literally show a large “spectrum”, we need to be cautious in selecting musical variables and making sonification. Even though the participating child prefers auditory stimuli in general, we need to empirically construct an optimized sonification zone to avoid the plausible reverse effect.

Another aspect to consider is the selection of an appropriate robot. Some children might prefer animal robots, whereas others might prefer humanoids. To this end, we are conducting robot acceptance research simultaneously. Our client-server structure will allow us to flexibly adapt to the environments of using different types of robots for each child. We believe this multifaceted approach will ultimately lead to more efficient and effective interventions for children with ASD.

7. ACKNOWLEDGMENTS

This project is supported by the National Institutes of Health under grant No. 1 R01 HD082914-01.

8. REFERENCES

- [1] Juslin, P. N., and Vastfjall, D. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, 31, 559-621.
- [2] Dautenhahn, K. 1999. Robots as social actors: Aurora and the case of autism. In *Proceedings of the Third International Cognitive Technology Conference, August, San Francisco*. Vol. 359. CT’99
- [3] Scassellati, B. 2009. Affective prosody recognition for human-robot interaction. *Microsoft Research’s External Research Symposium*. Redmond, WA, USA. 2009.
- [4] Kramer, G., Walker, B. N., Bonebright, T., Cook, P., Flowers, J. and Miner, N. e. a. 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 (Santa Fe, NM: International Community for Auditory Display (ICAD)).
- [5] Prelock, P. A., and McCauley, R. J. 2012. *Treatment of Autism Spectrum Disorders*. Baltimore: Brooks Publishing.
- [6] Hanson, E., Kalish, L. A., Bunce, E., Curtis, C., McDaniel, S., Ware, J., and Petry, J. 2007. Use of complementary and alternative medicine among children diagnosed with autism spectrum disorder. *Journal of autism and developmental disorders*, 37, 4, 628-636.
- [7] Moore, D., McGrath, P., and Thorpe, J. 2000. Computer-aided learning for people with autism—a framework for research and development. *Innovations in Education and Training International*, 37, 3, 218--228.

- [8] Hailpern, J. 2007. Encouraging speech and vocalization in children with autistic spectrum disorder, *SIGACCESS NEWSLETTER*, 89, (Sept. 2007).
- [9] Feil-Seifer, D. and Mataric, M. 2008. Robot-assisted therapy for children with Autism Spectrum Disorders, *IDC Proceedings – Workshop on Special Needs*, June 11-13, Chicago, IL.
- [10] Michaud, F. and Theberge-Turmel, C. 2002. Mobile robotic toys and autism. In Dautenhahn, K. (Ed.), *Socially Intelligent Agents – Creating Relationships with Computers and Robots*, Springer, 125-132.
- [11] Scassellati, B. 2007. How social robots will help us to diagnose, treat, and understand autism, In S. Thrum, R. Brooics, H. Durrant-Whyte (Eds.), *Robotics Research*, STAR 28, 552-563.
- [12] Stanton, C. M., Kahn, Jr., P. H. Severson, R. L., Ruckert, J. H., and Gill, B. T. 2008. Robotic animals might aid in the social development of children with autism, In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (Amsterdam, Netherlands, March, 12-15), HRI '08*. ACM.
- [13] Werry, I., Dautenhahn, K., and Harwin, W. 2001. Investigating a robot as a therapy partner for children with autism. In *Proceedings of the European Conference for the Advancement of Assistive Technology (Ljubljana, Slovenia. Sept.) AAATE'01*.
- [14] Accordino, R., Comer, R., and Heller, W. B. 2007. Searching for music's potential: A critical examination of research on music therapy with individuals with autism. *Research in Autism Spectrum Disorders*, 1, 1, 101-115.
- [15] Gold, C., Wigram, T., and Elefant, C. 2006. Music therapy for autistic spectrum disorder. *The Cochrane Library*.
- [16] Rosen, H. J., and Levenson, R. W. 2009. The emotional brain: Combining insights from patients and basic science. *Neurocase*, 15, 3, 173-181.
- [17] Sterkenburg, J., Jeon, M., and Plummer, C. 2014. Auditory emoticons: Iterative design and acoustic characteristics of emotional auditory icons and earcons. In M. Kurosu (Ed). *Human-Computer Interaction, Part II, HCII 2014, Lecture Notes in Computer Science (LNCS) 8511*, 633-640. Springer International Publishing Switzerland.
- [18] Liao S., Zhu, X., Lei, Z., Zhang, L., and Li, S. Z. 2007. Learning multi-scale block local binary patterns for face recognition. In: Lee, S.-W., Li, S. Z. (eds.) *ICB 2007. LNCS. 4642*. 828–837. Springer, Heidelberg.
- [19] Brooks, D., and Howard, A. M. 2010. A computational method for physical rehabilitation assessment. In *Proceedings of the 3rd IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics. BioRob '10* (pp. 442-447). IEEE.

LIFEMUSIC: REFLECTION OF LIFE MEMORIES BY DATA SONIFICATION

Ridwan A. Khan¹, Ram K. Avvari¹, Katherine Wiykovics¹, Pooja Ranay², and Myounghoon Jeon¹

¹Computer Science,
Michigan Technological University,
Houghton, MI 49931, USA
{ridwank, javvari, krwiykovi, mjeon}@mtu.edu

²Electrical and Computer Engineering,
Michigan Technological University,
Houghton, MI 49931, USA
pranay@mtu.edu

ABSTRACT

Memorable life events are important to form the present self-image. Looking back on these memories provides an opportunity to ruminate meaning of life and envision future. Integrating the life-log concept and auditory graphs, we have implemented a mobile application, “LifeMusic”, which helps people reflect their memories by listening to their life event sonification that is synchronous to these memories. Reflecting the life events through LifeMusic can relieve users of the present and have them journey to the past moments and thus, they can keep balance of emotions in the present life. In the current paper, we describe the implementation and workflow of LifeMusic and briefly discuss focus group results, improvements, and future works.

1. INTRODUCTION

Past events are essential part to form the present self. On one hand, remembering these events can make some people nostalgic. On the other hand, people may evaluate their lives through reflecting past events. To help this self-reflection process, life-logging research and products have been introduced. How can we facilitate this reflection process to be more effective rather than just piling up the data? What if we sonify our life events and listen to them as a music piece? Given that sonification can provide not only cognitive information, but also strong emotional information [1], it might be a plausible approach. Our research started to address these questions. Past memories can be either pleasant or unpleasant and thus, we are able to plot the life events on the Cartesian coordinate with the mapping of the X axis as time and Y axis as emotions (e.g., positive-negative valence). Leveraging the life-logging concept and auditory graphs, we have developed an Android app, “LifeMusic”, to help people reflect their memories more deeply and remember precious life moments more vividly. People are not able to change the past, but they may be able to connect it to the present and better realize the value of their present lives.

2. LIFE-LOGGING

Life-logging is not a new concept. Researchers have tried to trace people’s geographical positions using GPS, WiFi [e.g., 2], or wearable camera [3] to support their memory about personal life events. Attempts such as LifeLog [4] or Microsoft MyLifeBits came out with a focus on life-long

information capture. This type of applications concentrated on events, states, and relationships of an individual and aggregated them as raw data into the person’s timeline. Some of them pursued the role of growth memories in the intentional cultivation of good life [5]. Another study with older adults [6] focused on remembering what life events (both positive and negative) shaped their lives. From this study, it seems that negative events are less likely to have shaped peoples’ lives than positively scripted events, such as marriage, raising kids, etc. Overall, most research has represented the data visually, whereas there has been little research on representing the data auditorily. Based on this background, we posit the hypothesis that providing auditory representations of one’s life events (in addition to visuals) will facilitate this reflection process more effectively.

3. AUDITORY GRAPHS

The ICAD community has a long history of auditory graph research [e.g., 7-10]. For example, Davison and Walker [10] tried to make a software standard for auditory graphs, “Sonification Sandbox”. Developed in Java, it provides a cross-platform tool for auditory display researchers. The Sonification Sandbox converts tabular information into a descriptive auditory graph with various sound profiles, including pitch, timbre, polarity, pan, and volume. It also provides graph contexts, such as tick marks and labels. Given that a graph itself is a complex construct, an auditory graph can be more easily presented to the listeners with this type of context. Smith and Walker [8] empirically showed that an addition of the context to auditory graphs better represented the data than without the context. Nees and Walker [9] also identified additional essential considerations of the auditory graph characteristics: data, mappings, scaling, polarities, context, temporal characteristics of auditory graph stimuli, and multiple data series, all of which we can further consider for our app.

Representing one’s life events with an auditory graph has some advantages over just visual representation: (1) Given that life has a time dimension, an auditory modality fits to displaying it effectively; (2) Since sound and music are deeply related to emotions, the sonification of life will promote people’s affective contemplation process of their lives; (3) We aim to develop a smartphone app and so sound is an appropriate channel to compensate for its small display. Integrating these ideas, we have developed our application, LifeMusic, which records users’ life events and their mood, and represent those events with visual and auditory graphs according to different memory nodes.



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License.

The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

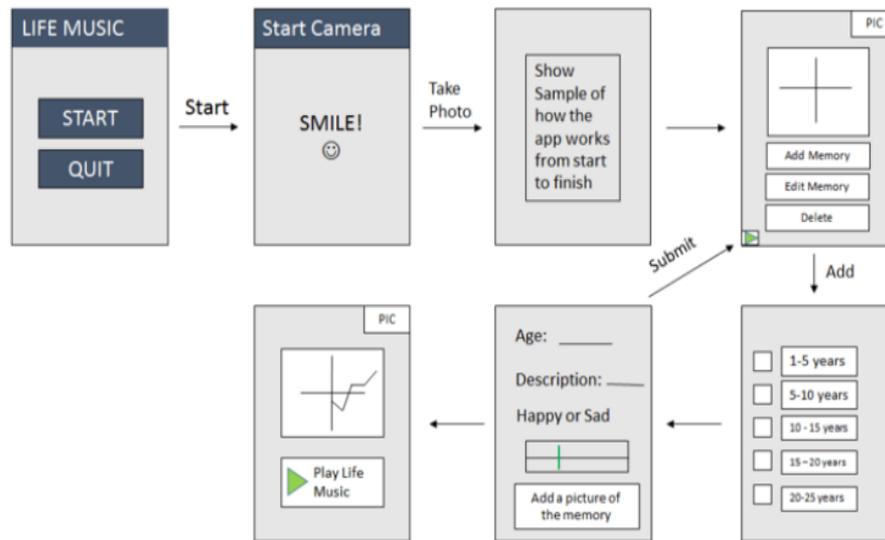


Figure 1. Proposed Model of the App Flow

4. LIFE MUSIC APP

LifeMusic application was developed as a smartphone app based on Android OS so that a large number of population could get access to this application and use it on the go. The application has a simple design which starts with a welcome screen (Figures 1-2). Pressing the start button in the welcome screen leads users to the picture capture screen. Users can take a photo or select one from their photo gallery, which is connected to their memory node. After the short demo video of how to use the app, users can see the core part of the application – the plot screen where they can plot their memories as nodes in a visual graph. In this screen, there is an “Add memory” button, which leads to a form screen. In the form screen, users can write about the specific memory. Users can give a title of the memory and write some description and time of the memory. This writing process can induce a specific emotional state linked to their memory [11]. Then, users rate this memory on a sad-happy scale between one to ten where one being absolutely sad and ten being absolutely happy. After saving the data in the form, users are redirected to the plot screen where the memory is added to the graph as a node. Again, in this way, users can add several memory nodes in the graph. In the graph, the X axis represents the time or age when the memory event happened and the Y axis represents the sad-happy scale.

We have used a third party API, GraphView [12] to show the graph. There is a button, “Play life music,” in the plot screen. When users press this button, a sonification piece is produced and played according to the plotted nodes of the graph. Initially, we have three types of different piano sounds for playing memories: one for happy, one for neutral, and the last one for sad memory events. The sound for each emotional node lasts relatively long (around 7-10 seconds) to facilitate user contemplation and reflection. For our working prototype, we used Android SDK’s multimedia framework (MediaPlayer API) to play the sound.

5. PRELIMINARY EVALUATION

To receive initial user feedback, we demonstrated our application to eight (4 female) graduate student participants in a single focus group. Their age was between 21 years to 30 years. They have different nationality, including India, China, USA, and Iran. First, the application was demonstrated to the participants and then, each of them experienced the application one by one. After getting used to the application, they could easily use the application to log their life events. Next, they provided their feedback on LifeMusic app. Our participants understood the concept of our app and how it worked very well. Some participants wanted to have more options of emotions (e.g., circumplex model [13]) rather than just happiness and sadness. Participants also suggested that the timeline of a memory can be more specific (e.g., time of the day, week, month, etc.) rather than just a year with a zooming function. Another suggestion was to make the music more rhythmic and varied. Overall, participants were excited about the use of our application and provided considerable ideas and improvements.

6. IMPROVED SONIFICATION CONCEPTS

Based on the comments and feedback from the focus group, we are improving our app in a number of aspects. Here, we focus on the description of the sonification part. First iteration is to include a melody contour of the familiar music pieces [e.g., 14]. For example, we can have users mark their events as four seasons: spring, summer, fall, and winter as people often use an analogy of four seasons to describe their lives. Then, we can play each season’s theme of Antonio Vivaldi’s “Four Seasons” for each life event. The second iteration could be to sonify the life events just as auditory graphs do with more granularity (e.g., using the Sonification Sandbox or other sound engine such as JFugue in Java) beyond just positive and negative valence. We can also include many other variables. For example, we will have users’ input on the strength and effect size of each life event

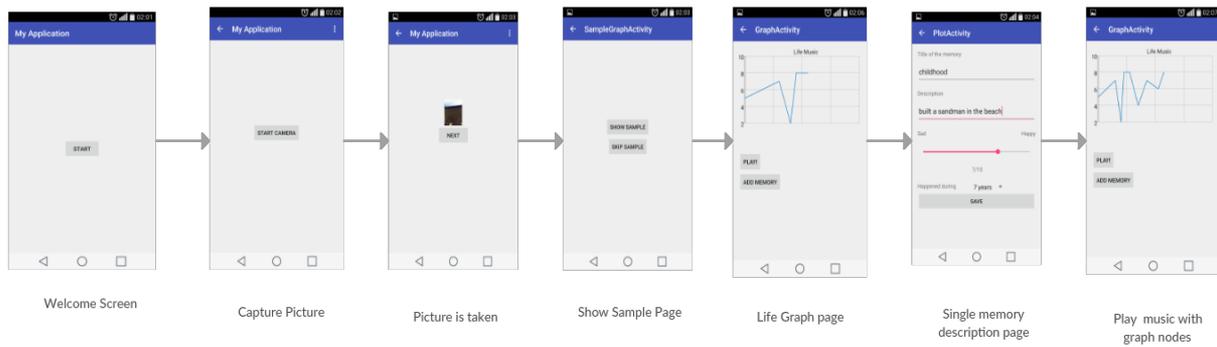


Figure 2. Screen Capture of the Actual LifeMusic App

and those data could be mapped onto musical variables, such as volume or duration of the sound. Depending on people's conceptualization of their life events, the same melody contour can vary and play a different genre of music (e.g., Rock & Roll vs. Classic vs. Jazz) or different cultural music.

7. CONCLUSION & FUTURE WORKS

LifeMusic is a small step towards a novel application with many possibilities. The aim of this application is to help people trigger their past nostalgic memories and reflect them more effectively. It allows users to gain a new perspective of their past and plan their future. Both happy and sad moments of the past constitute the "music of life". The current application can be improved by adding more subtle emotions rather than the simple valence dimensions. Another extension for this application is to create it for multiple users so that one user can compare and understand his or her own graph with others' by orchestrating them together. Of course, we will also use different visual representations of the graph and improve the sound generation by making it more accessible and customizable. Overall, the project is still a work-in-progress, but we are able to demonstrate it in the conference and hope to get some feedback from the ICAD community. We hope that a simple life graph system like LifeMusic can be a little reminder for people that life is short and we should try to make wonderful memories everyday regardless of being happy or sad.

8. ACKNOWLEDGMENT

The authors would like to thank all the students who came and spent their precious time to view and provide feedback about the application.

9. REFERENCES

- [1] M. Jeon, "Two or three things you need to know about AUI design or designers," in *Proc. of Int. Conf. (ICAD2010)*, Washington, D.C., June, 2010.
- [2] J. Rekimoto, T. Miyaki, and T. Ishizawa, "LifeTag: WiFi-based continuous location logging for life pattern analysis." In *LoCA* (Vol. 2007, pp. 35-49).
- [3] A. J. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, and K. Wood, "Do life-logging technologies support memory for the past?: an experimental study using sensecam," in *Proc. (CHI2007)* (pp. 81-90). ACM Press. 2007
- [4] K. O'Hara, M. M. Tuffield, and N. Shadbolt, "Lifelogging: Privacy and empowerment with memories for life", *Identity in the Information Society*, vol. 1, no. 1, pp. 155-172, 2009.
- [5] J. J. Bauer, D. P. McAdams, and A. R. Sakaeda, "Interpreting the good life: Growth memories in the lives of mature, happy people", *Journal of Personality and Social Psychology*, pp. 203-217, 2005.
- [6] D. Bertsen, D. C. Rubin, and I. C. Siegler, "Two versions of life: Emotionally negative and positive life events have different roles in the organization of life story and identity," *Emotion*, vol. 11, no. 5, pp. 1190-1201, 2011.
- [7] M. L. Brown, and S. A. Brewster, "Drawing by ear: interpreting sonified line graphs," in *Proc. of Int. Conf. (ICAD2003)*, 2003.
- [8] D. R. Smith and B. N. Walker, "Tick-marks, axes, and labels: The effects of adding context to auditory graphs", in *Proc. of Int. Conf. (ICAD 2002)*, 2002.
- [9] M. A. Nees, and B. N. Walker, "Listener, task, and auditory graph: Toward a conceptual model of auditory graph comprehension", in *Proc. of Int. Conf. (ICAD 2007)*, 2007.
- [10] B. Davison, B. N. Walker, "Sonification Sandbox overhaul: Software standard for auditory graphs," in *Proc. of Int. Conf. (ICAD 2007)*, Montreal, Canada, 26-29 June. pp. 509-512, 2007.
- [11] G. V. Bodenhausen, L. A. Sheppard, and G.P. Kramer, "Negative affect and social judgment: The differential impact of anger and sadness," *European Journal of Social Psychology*, vol. 24, pp. 45-62.
- [12] <http://www.android-graphview.org>
- [13] J. A. Russel, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.
- [14] B. N. Walker, J. Kim, and A. Pendse, "Musical soundscapes for an accessible aquarium: Bringing dynamic exhibits to the visually impaired," in *Proc. of Int. (ICMC 2007)*, Copenhagen, Denmark, August 27-30, 2007.

ORAL PAPERS

Mobiles and Wearables

TOWARDS AN IN-VEHICLE SONICALLY-ENHANCED GESTURE CONTROL INTERFACE: A PILOT STUDY

Jason Sterkenburg, Steven Landry, Myounghoon Jeon, and Joshua Johnson

Mind Music Machine Lab

Michigan Technological University,
1400 Townsend Ave. Houghton, MI, USA
{jtsterke, sglantry, mjeon, jocjohns}@mtu.edu

ABSTRACT

A pilot study was conducted to explore the potential of sonically-enhanced gestures as controls for future in-vehicle information systems (IVIS). Four concept menu systems were developed using a LEAP Motion and Pure Data: (1) 2x2 with auditory feedback, (2) 2x2 without auditory feedback, (3) 4x4 with auditory feedback, and (4) 4x4 without auditory feedback. Seven participants drove in a simulator while completing simple target-acquisition tasks using each of the four prototype systems. Driving performance and eye glance behavior were collected as well as subjective ratings of workload and system preference. Results from driving performance and eye tracking measures strongly indicate that the 2x2 grids yield better driving safety outcomes than 4x4 grids. Subjective ratings show similar patterns for driver workload and preferences. Auditory feedback led to similar improvements in driving performance and eye glance behavior as well as subjective ratings of workload and preference, compared to visual-only.

1. INTRODUCTION

Touchscreens in vehicles have increased in popularity in recent years. Touchscreens provide many benefits over traditional analog controls like buttons and knobs. They also introduce new problems. Touchscreen use requires relatively high amounts of visual-attentional resources because they are visual displays. Driving is also a visually demanding task. Competition between driving and touchscreen use for visual-attentional resources has been shown to increase unsafe driving behaviors and crash risk [1]. Driving researchers have been calling for new infotainment system designs which reduce visual demands on drivers [2]. Recent technological advances have made it possible to develop in-air gesture controls. In-air gesture controls, if supported with appropriate auditory feedback, may limit visual demands and allow drivers to navigate menus and controls without looking away from the road. Research has shown that accuracy of surface gesture movements can be increased with addition of auditory feedback [3]. However, there are many unanswered questions surrounding the development of an auditory supported in-air gesture-controlled infotainment system: What type of auditory feedback do users prefer? How can auditory feedback be displayed to limit cognitive load? What type of menu can offer an easily navigable interface for both beginners and experienced users? More importantly, do these displays reduce

the eyes-off-road time and frequency of long off-road glances? Does the system improve driving safety overall when compared to touchscreens or analog interfaces? These are among the many questions that we attempt to address in this project, of which, this study is a first step. This study describes our efforts to develop an in-vehicle sonically-enhanced gesture control interface. The development of the prototypes draws from research in movement science, human-computer interaction (HCI), and auditory display research to develop prototype that improves on the safety of touchscreen interfaces.

2. DRIVING

2.1. Multi-tasking in Vehicles

In-vehicle information systems (IVIS), such as navigation devices, mobile phones, and radios often require manual input from drivers. If a driver wants to use an IVIS, he/she must balance the demands of the driving task with the demands of using the IVIS. Multiple Resource Theory [4] models how the demands of multi-tasking influence the performance on each of the tasks being completed. It suggests that while multi-tasking, performance on two or more tasks is dependent on their overlap in demand for resources. If two tasks share demands for similar resources, then performance on one, or both tasks will suffer. Both driving and IVIS use are primarily visual-manual tasks. Multiple Resource Theory predicts that driving performance may suffer as drivers attempt to use IVISs, as long as those IVISs require visual-manual resources to use. Auditory feedback has potential to facilitate IVIS use by providing driver with information without introducing competition for visual resources. Indeed, auditory feedback has been shown to improve menu navigation in IVISs [e.g., 5].

2.2. Eye Glances and Driving

Not all off-road glances are equal in their impact on driving performance. Compared to normal, baseline driving, short glances away from the road pose little or no risk to driving safety. Long glances away from the road – 2 seconds or more – increase near-crash/crash risk by at least two times normal driving [6]. The National Highway Traffic Safety Administration (NHTSA) has developed guidelines for IVIS design that suggest limits for permissible visual demands of IVIS use [7] which state that a driver should be able to complete tasks while driving with glances away from the road of 2 seconds or less. These guidelines and principles informed the design and analysis of the pilot study and will inform future iterations of the prototype design and future evaluations of the prototype effectiveness.



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License.

The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

3. MOVEMENT SCIENCE

3.1. Fitts' Law

Paul Fitts' first quantified a movement task's difficulty, known as the index of difficulty (ID) [8, 9]. The original Fitts' Law equations describe movement along one dimension (1).

$$ID = \log_2 \left(\frac{2A}{W} \right), \quad (1)$$

Here, A is the amplitude, or distance, from the start of the movement to the target and W is the target width. The Shannon Formulation of Fitt's law (2) is generally preferred now because of its improved fit to observations while still adhering to Fitts' Law and because it ensures a positive value for ID.

$$ID = \log_2 \left(\frac{2A}{W} + 1 \right), \quad (2)$$

This equation can help us predict the difficulty of completing movement tasks in our different systems. For example, when comparing movements toward similarly positioned targets in the two different grid sizes, such as target A in the 2x2 grid and target A in the 4x4 grid (Figure 1), if the amplitude is 50 cm for both grids (approximately true), and the target size in the 2x2 grid is 12.6 cm and 6.3 cm in the 4x4 grid, then the calculated ID for the 2x2 is 1.79 and the ID for 4x4 is 2.5. This predicts that selecting targets on the 4x4 will be more difficult. We do not suggest that Fitts' Law provides a complete description of the nature of these complex cognitive, visual-manual search tasks, but it does give us a foundation from which to make simple predictions about relative difficulty of using systems with different target sizes.

3.2. Auditory Feedback and Fitts' Law

Fitts' Law, and most of the related work done in the area of movement science have assumed that feedback about movement was obtained through the visual and proprioceptive modalities [10]. Research has shown that proprioceptive cues alone lead to reduced accuracy in movement tasks [11]. Since the in-vehicle gesture interface is intended to be used by drivers who are simultaneously driving a vehicle, visual resources may not be available. Proprioceptive cues alone may be insufficient to aid in movement toward targets. It is currently unclear how other feedback modalities, like auditory or haptic, can be best utilized to facilitate visually-unaided movement tasks while minimizing workload and unnecessary system noise.

4. PILOT STUDY

4.1. Objectives and Hypotheses

The purpose of this study was to evaluate the impact of two major design features on driving performance and driver glance behavior: the size and number of target boxes, and the presence of auditory feedback.

Hypothesis 1: We hypothesized that the larger target sizes would reduce the secondary task difficulty and result in better driving performance (lower lane deviations) and eye glance behavior (fewer glances, less eyes-off-road time, fewer long glances) compared to smaller target sizes.

Hypothesis 2: We also hypothesized that auditory feedback would decrease secondary task difficulty and result in better driving performance and eye glance behavior compared to conditions without auditory feedback.

4.2. Participants

A total of seven participants were recruited from Michigan Technological University undergraduate psychology student pool. Among the participants one was male and seven were female.

4.3. Equipment

4.3.1 In-vehicle Sonically-Enhanced Gesture Control Interface

The in-vehicle gesture interface is comprised of two major components. A LEAP Motion, an infrared sensor designed to recognize hand features, was used to detect the hand position of the driver. Data from the LEAP Motion is sent to Pure Data, a free, open-source, real-time graphical programming environment for audio and visual processing. Within the customized Pure Data program there are audio and visual displays generated from the LEAP Motion data. The LEAP Motion tracks the center of a user's palm and counts the number of visible fingers and relays that information to Pure Data, which contains a visual grid display (Figure 1).

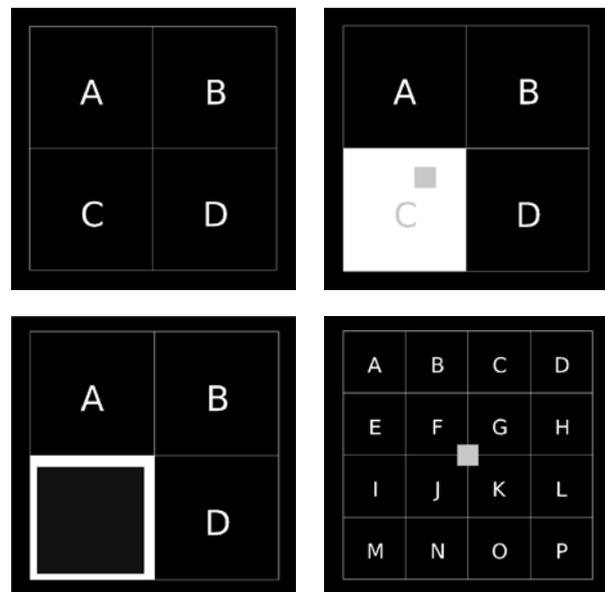


Figure 1: 2x2 grid (Top Left), 2x2 grid with visualization of hand position and highlighting box C (Top Right), 2x2 grid showing visualization of a selection (Bottom Left), and Graphical display of 4x4 grid with hand position (Bottom Right).

A graphic is displayed on a 1280x1024 monitor (Figures 1 & 2). The graphic shows a grid (2x2 or 4x4). Each box contains a letter. As the user holds his/her hand over the LEAP Motion, the visual display shows a box representing the position of his/her hand within the grid. If the center of the user's hand is within one of the boxes, that box is highlighted. For design concepts which have audio feedback, the same action will cue a text-to-speech .wav file for the letter in the box that is highlighted. Navigation through the system was

intended to be completed along a horizontal plane, with controls working analogously to a computer mouse. Target selection is dependent on the number of fingers visible to the LEAP Motion. If the system detects five fingers, then it will select the target which is highlighted at that moment. For the concept designs that have audio feedback, a selection action is followed by a confirmatory auditory icon which contains two “raindrop” tones, the first low followed immediately by a second higher frequency note. This is intended to provide an indication of selection.

4.3.2 Driving Simulator

A National Advanced Driving Simulator (NADS) MiniSim medium-fidelity driving simulator (Figure 2) was used for all driving scenarios. The driving scenario consisted of a single circuit through a residential area with many left and right curves. There were no other cars in the scenario. The simulator automatically records lane deviations and vehicle speed, along with many other variables.



Figure 2: Driving simulator setup, visual display monitor with webcam, and LEAP Motion.

4.3.3 Eye Tracking

Eye glance behaviors were recorded by a webcam placed on top of the visual display monitor. The eye glances were later coded by a researcher and placed into three categories based on the estimated length of the glance duration: short (<1 second), medium (1 second \leq t \leq 2 seconds), and long (>2 seconds).

4.4. Experimental Design

The study was a within-subjects repeated measures factorial design. Each participant completed all four conditions in one session.

- 2x2 grid with auditory feedback (2x2 VA)
- 2x2 without auditory feedback (2x2 V)
- 4x4 with auditory feedback (4x4 VA)
- 4x4 without auditory feedback (4x4 V)

4.5. Procedure

4.5.1 Training

Before driving in the simulator participants were introduced to the gesture prototype system. Initially, participants were shown the system and given no instruction in order to observe their first assumptions about how the system is used. A brief training period followed, in which participants were instructed to navigate with a closed fist and select by showing all five fingers. Practice trials were completed until the participant

was comfortable with the system. Next, participants were introduced to the driving simulator. Participants were told to drive in the right lane, and maintain a speed between 30-40 mph. The participants were given no instructions about how they should balance the demands of the two tasks.

4.5.2 Concept Systems

The order in which participants used the concept systems was randomized. A total of 32 selection tasks, evenly divided between target boxes, were completed for each concept system, taking approximately five minutes to complete. Auditory cues instruct participants which target to select (e.g., “select option B”). The order of the auditory cues was randomly determined by the Pure Data program.

4.5.3 Questionnaires

After completing all of the selection tasks, the participants were asked to stop the car and put it in park. During that time, the experimenter asked participants about his/her first impressions. Qualitative notes were taken regarding participants first impressions. Next, participants were asked several questions about their workload [12], including: mental demand, physical demand, performance, effort, and frustration using the electronic version of NASA-TLX. This process was repeated for all four concept system designs.

4.5.4 Semi-structured Interview

Following completion of all concept system designs, a short interview was conducted to identify issues that participants noticed and to probe about experiences with various aspects of the system, including the target size and the presence of auditory feedback.

5. RESULTS

5.1. Driving Performance

Speed data indicate that participants were generally capable of maintaining a speed between 30-40 mph, as instructed, while using each of the concept designs. Lane deviation data show a pattern indicating that participants’ lane deviations were larger when using the systems with the smaller target sizes (4x4 grids) (Figure 3). Presence of auditory feedback appeared to have little or no effect on lane deviations.

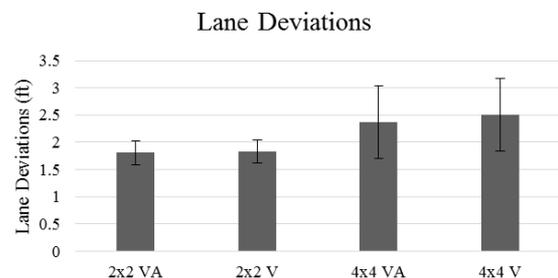


Figure 3: Mean lane deviations for each of the concept systems. Error bars denote 95% confidence intervals.

5.2. Eye Glance Behavior

Drivers made more frequent off-road glances for design concepts with smaller target sizes, and also for systems with no auditory feedback. This is true for all three glance durations (short, medium, long). The effect of both the target size and

the auditory feedback appears to be large. Target size and auditory feedback seem to act independently on glance durations, with no interaction occurring.

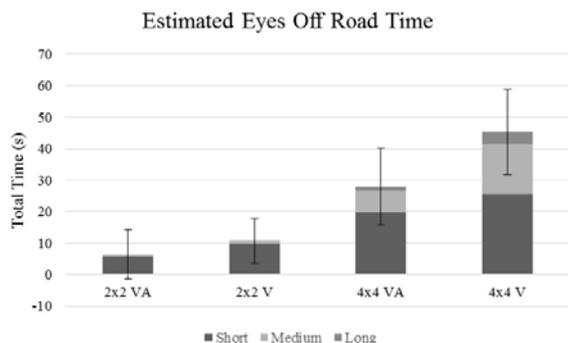


Figure 4: Cumulative eyes-off-road time for each of the concept systems. Error bars represent 95% confidence intervals.

5.3. Workload

NASA-TLX results show similar patterns for mental demand, effort, and frustration, each of which showed lowest scores for 2x2 VA, followed by 2x2 V, 4x4 VA, and 4x4 V. Perception of performance followed the reverse pattern, with the 2x2 VA grid resulting in highest perceptions of performance and the 4x4 V grid resulting in lowest perceptions of performance.

5.4. Semi-structured Interview

When participants were asked to rank-order their overall system preferences, they nearly unanimously favored systems in the following order: 2x2 VA, 2x2 V, 4x4 VA, 4x4 V. Two participants said that the auditory feedback was helpful for 2x2 grids but became more annoying than useful for 4x4 grids. Participants cited the ease of memorizing and acquiring the larger targets and the helpfulness of auditory cues (preview cues and confirmatory cues).

Researchers also observed that some participants initially attempted to control the device by moving vertically rather than horizontally. They stated that the vertical mapping was more intuitive to them. However, the current orientation mapping is used because movements tend to be faster along the x-plane than the y-plane [10]. Interestingly, participants would frequently move their hand down as they moved backwards, although no participants acknowledged conscious control over their downward movement.

6. DISCUSSION AND FUTURE WORKS

The trends for all of the dependent measures indicate that larger target sizes, such as those in the 2x2 grids, lead to improved driving safety outcomes including lane deviations, eye glance frequency, eye glance duration as well as subjective measures of workload.

It is possible that the 2x2 grid is easier because the proprioceptive and/or peripheral visual information is sufficient to guide a person within the target range. Conversely, the smaller targets in the 4x4 grid may require additional visual information because the smaller targets cannot be acquired with proprioceptive information alone.

These results suggest that selection tasks with difficulty indices (ID) of 2.58 or higher should not be considered if the

control space is located immediately in front of the in-vehicle center stack. Increasing target sizes and providing previewing and confirmatory auditory feedback can reduce secondary task difficulty and improve driving safety outcomes.

With the pilot study completed, we are developing custom software to allow us to test more refined designs. This software will come with configuration files allowing for a wider range studies. The new menu will be configurable to allow us to study the effects of variable menu layouts, different auditory displays for menu navigation (e.g., spoken titles, earcons, etc.), and record timing of participant actions. We will have predefined task sets defined within the software. Timestamps of each point of data from the start to the completion of the action will be recorded and will be later analyzed to better understand the relationship between a sonically-enhanced gesture controls and driving performance.

7. REFERENCES

- [1] W. Horrey, and C. Wickens, "In-vehicle glance duration: distributions, tails, and model of crash risk" *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2018, pp. 22-28, 2007.
- [2] P. Green, "Crashes induced by driver information systems and what can be done to reduce them," In *Sae Conf. Proc. SAE*; 1999, 2000.
- [3] B. Hatfield, W. Wyatt, and J. Shea, "Effects of auditory feedback on movement time in a Fitts task," *Journal of Motor Behavior*, vol. 42, no. 5, pp. 289-293, 2010.
- [4] C. Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, no. 2, pp. 159-177, 2002.
- [5] M. Jeon, T. M. Gable, B. K. Davison, M. Nees, J. Wilson, and B. N. Walker, "Menu navigation with in-vehicle technologies: Auditory menu cues improve dual task performance, preference, and workload," *International Journal of Human-Computer Interaction*, vol. 31, no. 1, pp. 1-16, 2015.
- [6] S. Klauer et al., "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," (2006).
- [7] P. Green, "Driver interface/HMI standards to minimize driver distraction/overload," (2008).
- [8] P. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *Journal of Experimental Psychology*, vol. 47, no. 6, pp. 381, 1954.
- [9] P. Fitts, and J. Peterson, "Information capacity of discrete motor responses," *Journal of Experimental Psychology*, vol. 67, no. 2, pp. 103, 1964.
- [10] J. Medina, S. Jax, and H. Coslett, "Two-component models of reaching: Evidence from deafferentation in a Fitts' law task," *Neuroscience Letters*, vol. 451, no. 3, pp. 222-226, 2009
- [11] S. Wallace, and K. Newell, "Visual control of discrete aiming movements," *The Quarterly Journal of Experimental Psychology*, vol. 35, no. 2, pp. 311-321, 1983.
- [12] S. Hart, and L. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Advances in Psychology*, vol. 52, pp.139-183, 1988.

SONIFICATION OF MOVEMENT FOR MOTOR SKILL LEARNING IN A NOVEL BIMANUAL TASK: AESTHETICS AND RETENTION STRATEGIES

John Dyer¹, Paul Stapleton² & Matthew Rodger¹

School of Psychology¹ & Sonic Arts Research Centre²,
Queen's University Belfast,
University Road,
BT7 1NN
jdyer01@qub.ac.uk

ABSTRACT

Here we report early results from an experiment designed to investigate the use of sonification for the learning of a novel perceptual-motor skill. We find that sonification which employs melody is more effective than a strategy which provides only bare timing information. We additionally show that it might be possible to 'refresh' learning after performance has waned following training - through passive listening to the sound that would be produced by perfect performance. Implications of these findings are discussed in terms of general motor performance enhancement and sonic feedback design.

almost entirely on research employing transformed or abstracted visual feedback. When sonification has been compared directly to such feedback, the guidance effect fails to materialize; task performance remains at a high level [9]. Retention tests without live sonification are crucial in this domain, as the goal is to learn an underlying movement skill; sonification is the vehicle for getting there more quickly, or learning the skill more accurately. Both everyday and sport-related skills (the usual targets of this treatment) should ideally *not* be dependent on immersion within a feedback system. Findings such as the above suggest that sonification could in fact be a more appropriate type of augmented feedback for skill learning than the same information provided via a visual display (for more on this idea, see [10]).

1. INTRODUCTION

Sonification of human movement is slowly becoming a more commonly-used strategy for the provision of augmented perceptual feedback for motor skill learning [1]–[4]. Typically, this entails some variable of motor performance being tracked by a sensing system (e.g. accelerometers, optical motion capture, touchpads, force-plates) and fed back 'live' to the moving individual in the form of synthesized sound [5]. Making movement information available through sound where it would otherwise be difficult to perceive can allow a learner to exert much finer control over their actions, ideally resulting in better task performance [6]. In this report, we aim to explore the benefits of making sonification musical (as compared to sonification concerned purely with providing temporal sonic information), and test a strategy for improving long-term retention of new skills learned with sonification.

2. THE VALUE OF SONIFICATION FOR PERCEPTUAL-MOTOR LEARNING

In the psychology of motor skill learning and feedback, a major concern is the transfer of learned motor skills beyond the feedback environment [7]. Traditionally, the consensus has been that overuse of augmented feedback leads to dependence, as learners come to over-rely on the guidance it provides [8]. This "guidance effect" has been assumed to apply to all types of augmented feedback, independent of sensory modality or form, however this assumption is based

3. AESTHETIC ISSUES

Where the efficacy of sonification has been tested experimentally, feedback systems have sometimes made use of aesthetically impoverished movement-sound mapping strategies. Pitch-mapping is the most common strategy in sonification generally [11], and the same can be said for the more narrow subdomain of sonification for perceptual-motor feedback. Konttinen et al., [12] for example, mapped deviation from a target to sine-tone pitch in a shooting task to provide feedback for use in controlling rifle stability. Schaffert and Mattes [13] mapped boat acceleration to the pitch of discrete tones in a MIDI note scale, while Powell and Lumsden [14] employed tone pitch to allow drivers to perceive their lateral g-force relative to a set limit in a motorsport racing task. More complex and interesting sounds have also been used to provide information about human movement, including vowel-like sounds (in golf swinging and jumping [5], [15]) and physical modelling of real-world noisy interactions (in handwriting [1]). Direct comparisons of basic vs. pleasant (but structurally similar) mapping strategies for motor skill learning have rarely been explored experimentally.

For a novel motor task which has not been sonified before, it can be difficult to know the extent to which one should focus on aesthetics when designing sound as feedback. Simple approaches to sonification which provide basic temporal information to help organize performance have been shown to be effective for learning new tasks [3], [9], however there could yet be potential benefits unlocked through use of a more interesting mapping. Motivational factors are seldom considered in perceptual-motor learning studies, despite their importance for task engagement and therefore, performance [16]. Sonification as feedback presents a unique opportunity to provide augmented perceptual information which is



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

pleasant, evocative and interesting to use; to carry over information-design habits from classic visual feedback experiments (which have historically provided performance information as moving lines on a screen e.g. [17], [18]) might be a tremendous waste of potential. This is one of the dilemmas we aim to probe with the current investigation - by comparing one type of sonification which provides only temporal information, to another which also employs melody.

4. PROLONGING RETENTION

As mentioned in section 2, good performance after the removal of sonified feedback is the goal of our research. It is not always practical to provide a live auditory display during motor performance, thus we require learning which is not dependent on the immediate presence of augmented feedback, i.e. learning that generalizes - and is not subject to the 'guidance effect' [8]. In a previous study by the authors similar to the current investigation (manuscript under revision at time of writing), we observed better performance in a custom bimanual shape-tracing task by participants learning under sonification conditions relative to control (i.e. silence). This task requires participants to trace a triangle with the left index finger while simultaneously tracing a diamond with the right (see Figure 1). When the task is performed correctly, participants move between corner zones with a timing ratio of 4:3. This type of bimanual dual-task is difficult to perform, but can be learned quickly with the use of augmented perceptual feedback [17]. Such feedback works by integrating perception of both hands into a single perceptual stream, which is more easily controllable [19]. When we removed sonification to test retention-without-feedback, the boosted performance by participants in that group remained; there was no evidence of a 'guidance effect'. However, in a second retention-without-feedback session 24 hours later, performance had declined and there was no longer a difference in scores between sonification and control.

4.1. Listening for retention

It may be possible to temporally extend the advantage of sonification by allowing participants to hear the sound of good performance before no-feedback retention-testing. It has been shown in musical instrument learning that listening to a learned piece of music elicits activations in neural areas associated with performing the piece [20]. It has been argued that this and similar such findings represent a mechanism of 'common coding' for perception and action in the brain [21]. In other words, perceptual experience of learned action is neurally very similar to active performance. This could be exploited to enhance recall of new motor skills in sonification, as has been demonstrated in keyboard learning [22]. If this strategy works, it could have implications for how sonification-based training should be implemented in real-life skills. Playing a recording is much less onerous than providing live sonification. For example, a sporting skill, say, a golf swing [15] can be trained using sonification in a lab setting. Sonification might enhance performance of the swing in the lab by making temporal information about bodily rotation more perceivable, and the learner may come to understand their action in terms of its sonic outcome. Through practice, it is expected that the learner would come to know the sound of a good swing and purposely act so as to

produce it. On the golf course, where it may be impractical to use live sonification (perhaps due to cumbersome equipment), the learner could listen to the sound of a good swing through headphones, and thus re-experience (part of) what it is like to produce a good swing, thereby enhancing motor sequence retention.

In the current experiment, we test this strategy by re-exposing participants to the sound of good performance prior to a 24-hour retention test, with the expectation that doing so should improve performance.

5. METHOD

Participants were recruited from the university undergraduate population (currently $N = 45$) and randomly allocated to one of three independent conditions.

All were required to learn the same bimanual shape-tracing task (Figure 1). Participants were instructed (via an animated demonstration) to trace two shapes (a triangle and a diamond) in an anticlockwise direction starting from the top corners, and to arrive at corner zones at regular intervals on each hand. When done correctly, the fingertips of both hands would complete a cycle (i.e. return to the top corner) at the same time. Task performance required continuous repeated cycles of the shapes.

Movements were tracked using reflective markers attached to a pair of modified golfing gloves which were picked up by four optical motion-capture cameras. Sonification was provided (where necessary) by streaming Cartesian coordinate data corresponding to the position of the fingertip marker of each hand into *Max/MSP 6.0* at 20Hz.

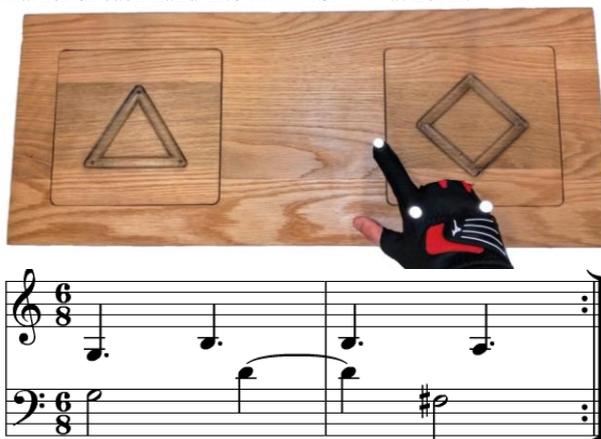


Figure 1: Custom bimanual shape-tracing apparatus used in the reported experiment (top) and notes produced by the sonification patch in the 'Melodic' experimental condition (bottom). This melody was composed by the authors for the purpose of the experiment.

One group of participants ($N = 15$) was required to learn the task without sonification of any kind (the 'Control' condition). This group listened to pink noise through headphones during practice. Another group ($N = 15$) practiced with basic sonification of fingertip corner arrivals (the 'Temporal' sonification condition). When a fingertip reached a corner, a short (200ms) burst of white noise was triggered. Correct performance thus produced a 4:3 rhythm. A third group ($N = 15$) practiced with melodic sonification of fingertip arrivals (the 'Melodic' sonification condition). In this condition, correct performance of the task produced a simple melody (right - left hand notes occurring in a 4:3 rhythm) on a synthesized plucked stringed instrument in the

key of G major (see Figure 1). Sounds were presented through headphones.

The procedure of this experiment consisted of three main stages: pretest, practice and retention testing.

5.1. Pretest

Participants completed a pretest trial at the beginning of the session. A demo animation was played to participants prior to the pretest which showed corner arrivals occurring as they would if the task were performed perfectly. The demo lasted 9 seconds and consisted of three cycles of the shapes. For the pretest, all participants heard constant pink noise to obscure potential task-intrinsic auditory feedback. The movement phase consisted of a 26-second window in which participants attempted to match the demonstration. No artificial feedback was provided.

5.2. Practice

Following the pretest, participants underwent fourteen, 26-second-long practice trials – the nature of which varied depending on condition assignment. The demo animation was played prior to every practice trial. Participants in the Control condition heard pink noise during demo presentation and the movement phase. Participants in the Temporal sonification condition heard 200ms bursts of white noise coincident with corner arrivals while the demo played and subsequently with their own corner arrivals on the shapes. Participants in the Melodic sonification condition heard the notes shown in Figure 1 coincident with corner arrivals in the demo and their own on the shapes. Participants did not commence movement until the demo had concluded. For participants in the sonification conditions, engaging in this task was thus instantiated as an unfolding sonic performance, and practice trials as repeated attempts to ‘play’ the task correctly. Participants in all groups received terminal (post-trial) feedback in the form of their inter-manual timing ratio plotted over time.

5.3. Retention testing

Following the practice phase, all participants immediately underwent a retention test with no demo, terminal (graph) feedback or sound except for constant pink noise during the movement phase.

Another retention test under exactly the same conditions was administered the following day.

Participants in the two sonification conditions were then re-exposed to the sound of perfect task performance. Note, they did not see the demo animation again, only the sound it produced during the practice phase the day before. Participants then completed another retention test. To control for potential practice effects of multiple-retention tests, participants in the control condition also completed this additional retention test, but did not hear any sound other than constant pink noise. Participants also completed a questionnaire asking about their experience of the experiment (enjoyment, interest and strategies used) and musical experience.

6. RESULTS AND DISCUSSION

The main measure of performance in the current task is the bimanual timing ratio produced by participants over time.

Within each trial, the absolute difference between produced and ideal (4:3) timing ratios was averaged to produce a single error score for each trial for each participant. As learner performance after practice is our primary interest, we here present a preliminary analysis of data from trial 14 and the following three retention tests (i.e. the final four stages shown in Figure 2 for all 3 conditions and all 45 current participants). A mixed ANOVA with trial and feedback group as factors revealed a significant main effect of feedback group: $F(2, 39) = 3.579, p = 0.037$, no significant main effect of trial: $F(2.147, 83.744) = 2.593, p = 0.077$ and no significant interaction: $F(4.295, 39) = 0.572, p = 0.696$. Our analysis does not currently go further because we are still in the process of collecting data, with the aim of $N = 60$.

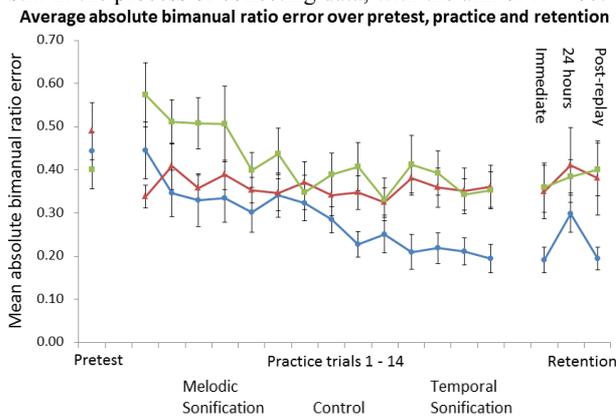


Figure 2: Rates of error for the three experimental groups over time. Live sonification and terminal feedback were provided only on practice trials. The three retention means show no-feedback error rates following practice, after 24 hours, and lastly, after listening to the sound of perfect performance. Error bars are standard error of the mean.

Figure 2 shows different rates of reduction in error for experimental groups over time. Participants in the Melodic sonification condition reached lower average error scores than the Temporal sonification and Control conditions. This indicates that Melodic sonification was most useful for acquisition of the bimanual skill. The same pattern is observed at the first retention test, in which no sonification feedback was available. This indicates that participants in the Melodic sonification condition were not dependent on the presence of augmented feedback for good performance. On the second retention test (after 24 hours), average ratio error in the Melodic sonification condition increases to levels similar to the Temporal sonification and Control conditions. However, subsequently re-exposing participants to the sound of good performance appears to have had the desired effect, at least in the Melodic condition – error reduces in line with performance on the previous day. No benefit of hearing the sound of good performance seems evident in the Temporal condition, and there is little if any practice effect for the Control condition on day 2.

Limited conclusions can be drawn from this incomplete dataset and the preliminary analysis we have conducted here. The lack of an enhancement effect of sonification in the Temporal condition is surprising, but may be related to motivation (the sound is very dull), or informational-structural factors (the melody specifies the ordering of bimanual movements, making the task relatively easier in that condition). This may become clear with further analysis including questionnaire data.

7. CONCLUSION

We have presented preliminary evidence which indicates potential for improving motor task retention with passive listening. This means that if a new motor skill is learned with sonification, it may be possible to effectively ‘refresh’ learning through listening, rather than placing learners back in an augmented feedback environment.

The value of melody and engaging sound vs. purely temporal sonic information for learning in this task may be partly motivational, but could perhaps be related to the extra, relevant information provided by the use of different tones. Feedback designers should consider using melodic movement sonification for either or both of these reasons.

8. REFERENCES

- [1] J. Danna, M. Fontaine, V. Paz-Villagrán, C. Gondre, E. Thoret, M. Aramaki, R. Kronland-Martinet, S. Ystad, and J.-L. Velay, “The effect of real-time auditory feedback on learning new characters.,” *Hum. Mov. Sci.*, vol. 43, pp. 216–228, Dec. 2015.
- [2] R. Sigrist, G. Rauter, L. Marchal-Crespo, R. Riener, and P. Wolf, “Sonification and haptic feedback in addition to visual feedback enhances complex motor task learning.,” *Exp. brain Res.*, vol. 233, pp. 909–925, Dec. 2014.
- [3] F. T. van Vugt and B. Tillmann, “Auditory feedback in error-based learning of motor regularity.,” *Brain Res.*, vol. 1606, pp. 54–67, May 2015.
- [4] F. Sors, M. Murgia, I. Santoro, and T. Agostini, “Audio-Based Interventions in Sport,” *Open Psychol. J.*, vol. 8, no. 3, pp. 212–219, 2015.
- [5] A. O. Effenberg, “Movement Sonification: Effects on Perception and Action,” *IEEE Multimed.*, vol. 12, no. 2, pp. 53–59, Apr. 2005.
- [6] J. Stienstra, K. Overbeeke, and S. Wensveen, “Embodying complexity through movement sonification,” in *Proceedings of the 9th ACM SIGCHI Italian Chapter International Conference on Computer-Human Interaction: Facing Complexity*, 2011, pp. 39–44.
- [7] R. Sigrist, G. Rauter, R. Riener, and P. Wolf, “Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review.,” *Psychon. Bull. Rev.*, vol. 20, no. 1, pp. 21–53, Feb. 2013.
- [8] D. I. Anderson, R. A. Magill, H. Sekiya, and G. Ryan, “Support for an explanation of the guidance effect in motor skill learning.,” *J. Mot. Behav.*, vol. 37, no. 3, pp. 231–8, May 2005.
- [9] R. Ronsse, V. Puttemans, J. P. Coxon, D. J. Goble, J. Wagemans, N. Wenderoth, and S. P. Swinnen, “Motor learning with augmented feedback: modality-dependent behavioral and neural consequences.,” *Cereb. Cortex*, vol. 21, no. 6, pp. 1283–94, Jun. 2011.
- [10] J. Dyer, P. Stapleton, and M. W. Rodger, “Sonification as concurrent augmented feedback for motor skill learning and the importance of mapping design,” *Open Psychol. J.*, vol. 8, no. 3, pp. 1–11, 2015.
- [11] G. Dubus and R. Bresin, “A systematic review of mapping strategies for the sonification of physical quantities.,” *PLoS One*, vol. 8, no. 12, p. e82491, Jan. 2013.
- [12] N. Konttinen, K. Mononen, J. T. Viitasalo, and T. Mets, “The effects of augmented auditory feedback on psychomotor skill learning in precision shooting,” *J. Sport Exerc. Psychol.*, vol. 26, pp. 306–316, 2004.
- [13] N. Schaffert and K. Mattes, “Effects of acoustic feedback training in elite-standard Para-Rowing.,” *J. Sports Sci.*, vol. 33, no. 4, pp. 411–8, Jan. 2015.
- [14] N. Powell and J. Lumsden, “Exploring novel auditory displays for supporting accelerated skills acquisition and enhanced performance in motorsport,” in *Proceedings of The 21st International Conference on Auditory Display (ICAD-2015)*, 2015, pp. 173–180.
- [15] M. Kleiman-Weiner and J. Berger, “The sound of one arm swinging: A model for multidimensional auditory display of physical motion,” in *12th International Conference on Auditory Display (ICAD), London, UK, June 20-23, 2006*.
- [16] G. Wulf, C. H. Shea, and R. Lewthwaite, “Motor skill learning and performance: a review of influential factors.,” *Med. Educ.*, vol. 44, no. 1, pp. 75–84, Jan. 2010.
- [17] A. J. Kovacs and C. H. Shea, “The learning of 90° continuous relative phase with and without Lissajous feedback: external and internally generated bimanual coordination.,” *Acta Psychol. (Amst.)*, vol. 136, no. 3, pp. 311–20, Mar. 2011.
- [18] D. W. Vander Linden, J. H. Cauraugh, and T. A. Greene, “The effect of frequency of kinetic feedback on learning an isometric force production task in nondisabled subjects.,” *Phys. Ther.*, vol. 73, no. 2, pp. 79–87, Feb. 1993.
- [19] E. A. Franz and R. McCormick, “Conceptual unifying constraints override sensorimotor interference during anticipatory control of bimanual actions.,” *Exp. brain Res.*, vol. 205, no. 2, pp. 273–82, Aug. 2010.
- [20] A. Lahav, E. Saltzman, and G. Schlaug, “Action representation of sound: audiomotor recognition network while listening to newly acquired actions.,” *J. Neurosci.*, vol. 27, no. 2, pp. 308–14, Jan. 2007.
- [21] W. Prinz, “Perception and Action Planning,” *Eur. J. Cogn. Psychol.*, vol. 9, no. 2, pp. 129–154, Sep. 2010.
- [22] A. Lahav, T. Katz, R. Chess, and E. Saltzman, “Improved motor sequence retention by motionless listening.,” *Psychol. Res.*, vol. 77, no. 3, pp. 310–9, May 2013.

Satellite Gamelan: microtonal sonification using a large consort of mobile phones

Greg Schiemer

composer, instrument-builder and software developer

greg.schiemer@protonmail.com

ABSTRACT

This paper describes an approach to sonification based on an iPhone app created for multiple users to explore a microtonal scale generated from harmonics using the combination product set method devised by tuning theorist Erv Wilson. The app is intended for performance by a large consort of hand-held mobile phones where phones are played collaboratively in a shared listening space. Audio consisting of handbells and sine tones is synthesised independently on each phone. Sound projection from each phone relies entirely on venue acoustics unaided by mains-powered amplification. It was designed to perform a microtonal composition called *Transposed Dekany* which takes the form of a chamber concerto in which a consort of players explore the properties of an microtonal scale. The consort subdivides into families of instruments that play in different pitch registers assisted by processes that are enabled and disabled at various stages throughout the performance. The paper outlines Wilson's method, describes its current implementation and considers hypothetical sonification scenarios for implementation using different data with potential applications in the physical world.

Author Keywords

collaborative sonification, mobile performance, microtonal composition, software instrument, pocket gamelan

1. INTRODUCTION

Ever since the first ICAD more than two decades ago the art of auditory display has become a tool to interpret complex data in many forms. [1] Its applications range from the use of audio signals to represent various kinds of natural phenomena such as tidal data collected over many years using conventional record keeping methods [2] to the design of air traffic control consoles for 'receiving, analyzing and acting upon complex information in a timely manner'. [3] The Listening to the Mind Listening project also showed how interpretations of the same data set by different sonification designers can vary in their musical characteristics [4] In its current form it is possible to argue that *Transposed Dekany*, the work described in this paper, is composition more than sonification. Stephen Barrass clarifies the distinction between the two:

When the primary intention of the composer shifts to the revelation of the source, the work crosses into the realm of sonification. With this crossing over comes a question of whether the listener can also understand the composer's intention to produce more than an experience of the music itself. [5]

Hermann describes Model-Based Sonification as a transformation from data to sound where

data is used to build an instrument or sound-capable object, while the playing is left to the user.

This almost describes the Satellite Gamelan, the app I created to perform *Transposed Dekany*, and potentially satisfies all four conditions in Hermann's definition of sonification - namely, that the sound reflects objective properties, that the transformation of data is systematic, that sonification is reproducible and that it can be used with different data.

2. PUBLIC SOUND

The most common presentation for auditory display involves sound projected from a fixed location such as stereo or multi-channel speakers or binaural headphones. With the improved audio signal processing capabilities and widespread uptake of mobile phone technology there is now an alternative framework for public listening which I call collaborative sonification.

Collaborative sonification allows many listeners to present and interact with complex data in a shared acoustic environment. Sound is projected via a large malleable speaker array consisting of hand-held battery-powered mobile phone speakers. Each phone becomes an independent sound source like any conventional hand-held musical instrument which results in the creation of ensemble sound albeit from a consort of electronic sound sources. The ideal built environment for such a presentation is a performance venue with high ceilings and reflective acoustics. Sound projection is no longer entrusted to one listener positioned at an electronic mixing desk trying (or perhaps not even bothering) to second-guess what listeners might expect to hear.

There are obvious limitations in projecting sound from a miniature loudspeaker with the power limitations and frequency response of a speaker phone. As with conventional consorts of viols, recorders or voices, strength comes from the size of the ensemble. While it is possible to produce audio covering the audible spectrum and project audio using mains powered speaker tether the phone to project I have chosen to work within the frequency response limitations of the speaker phone.

Vickers noted that using 'organising principles of tonal music' to present data can 'result in more aesthetically-coherent sonifications'. [6] If those principles were extended to include music with a non-equally tempered pitch spectrum based on harmonic ratios could this not also provide 'a system that is more open to reading than it is a musical style that is recognized as such'?

My focus has not been what to do about a limited frequency response, but rather, what to do with frequencies produced within a limited range of four octaves. In fact the focus on frequency is even more specific: it is about unequal relationships that occur between pitches of a scale derived from pure harmonics. In the ideal acoustic environment a mobile phone speaker will clearly reproduce signals derived from pure harmonics played at the appropriate frequency. For that reason the app does not use sampled audio but synthesises bell tones and chorus tones by adding sines tones. These are tuned in a variety of ways using a simple algorithm based on harmonics.

The world's musical traditions have many such scales where the size of intervals, i.e. the gap between two notes, is defined by points on the harmonic series. One of the most common scales are pentatonic. Some scales will sound more recognisably pentatonic than others that are tuned using different harmonics; this is especially true for audiences conditioned to hearing a pentatonic scale played on the black notes of a conventional music keyboard.

The Satellite Gamelan app was never intended to sonify anything more than abstract data i.e. harmonic numbers. Nevertheless it offers a rich and elegantly variable palette of harmonic flavours derived from numbers available in the harmonic series.

3.MOBILE COLLABORATION

The work has precedents in other battery powered mobile events such as Iain Mott's *Sound Mapping* in 1998 [7] and Golan Levin's *Dialtones (A Telesymphony)*[8] in 2001. The latter involved active participation by a concert audience. Mobile phones became a multi-channel sound system through which triggered sequences of ringtones are played. Mobile phone ensembles have sprung up at Stanford, Michigan and Helsinki Universities using apps that synthesise audio with players wearing battery-powered amplifier-speakers [9]. These, along with the Satellite Gamelan, are pre-dated by earlier mobile projects I developed for j2me phones [10] and purpose-built battery powered analogue circuitry called UFOs [11].

4.MICROTONAL TUNING

The app also builds on the microtonal legacy of composer, instrument-builder and theorist Harry Partch whose work recognised and celebrated the diversity of tuning that lies outside the western concert tradition [12]. Each phone is tuned to a dekany, a 10-note scale generated from harmonics using a method devised by tuning theorist and instrument-builder Erv Wilson.

Wilson theories include a system of keyboard mapping that provides a broader historical perspective for Partch's work and a trajectory that embraces both experimental and traditional tuning [13]. One of Wilson's tuning theories, known as combination product sets (CPS) is a way to generate many scales of various size and harmonic flavour.

4.1.Dekany

The dekany is a CPS scale where pitches are generated by multiplying combinations of two harmonics from a set of five harmonics: 1, 3, 7, 9 and 11.

Table 1 shows the relationship between pitches of the dekany and its five harmonic generators. Pitches are described in cents and ratios. Note that unison - 0 cents or 1/1 - is missing from the scale, the salient feature of CPS scales; nevertheless, for convenience, rational scale pitches are defined with reference to unison. In a CPS scale the numerator represents a product while the denominator represents the missing unison, or some octave above it.

Table1. Pitches of the 1-3-7-9-11 dekany with five generators

	cents	ratios	1	3	7	9	11
1	53.273	33/32		3			11
2	203.910	9/8	1			9	
3	320.144	77/64			7		11
4	470.781	21/16		3	7		
5	551.318	11/8	1				11
6	701.955	3/2	1	3			
7	755.228	99/64				9	11
8	905.865	27/16		3		9	
9	968.826	7/4	1		7		
10	1172.736	63/32			7	9	

The 1-3-7-9-11 dekany has two interleaved scales that are recognisably pentatonic. Their pitches are separated by two intervals, one small and one large, just like pentatonic scales played on an equal tempered keyboard. Pitches of the odd numbered pentatonic scale are separated by intervals of 231.174 and 266.871 cents while pitches of the even numbered pentatonic scale are separated by intervals of 213.598 and 284.447 cents; by comparison, pitches of pentatonic scales played on a conventional equal tempered 12-tone keyboard are separated by intervals of 200 and 300 cents.

Each pentatonic scale represents a different harmonic flavour, a result of differences in the size of their small and large intervals. These flavours produce strong consonance when heard separately and strong dissonance when heard simultaneously, a dissonance and consonance stronger than any produced in 12-tET. By organising the 1-3-7-9-11 dekany as two pentatonic scales, I expected to make it easier for musicians to focus on pitches of one scale and selectively ignoring interference from pitches of the other.

This assumption can be tested in a laboratory. Using randomly selected tones interleaved with tones of a familiar melody, Alan Bregman demonstrates how auditory streaming makes it possible for listeners to segregate a stream of melodic tones from interfering tones that camouflage the melodic stream [14].

The 1-3-7-9-11 dekany has two scales each with a different melodic contour. Each tends to camouflage the other and each are recognisably pentatonic, a property related to the harmonic generators used to create the dekany.

4.2.Transposition

Transposition is a process for playing a scale in a different register starting on each note of the original scale. On a 12-tone equal-tempered keyboard, a transposed scale always fall on other notes on the keyboard. By contrast, transposing a scale with pitches that are generated from harmonics will result in additional harmonically related pitches that are not present in the original scale.

New harmonic flavours are created when different transpositions of the dekany are heard simultaneously. In performance players are divided into five groups, each playing in a different transposition.

The five transpositions are shown in Table 2, rearranged as interleaved odd and even pentatonic scales represented here in cents. Transposition results in 26 new pitches. Each transposition has some pitches in common with other transpositions creating harmonic connections between them.

	dekany	t1	t2	t3	t4	t5
odd pentatonic		+83.273	+320.144	+551.318	+701.955	+968.826
1	53.273	106.546	373.417	604.591	808.501	1022.099
3	320.144	373.417	640.288	871.462	1075.372	88.970
5	551.318	604.591	871.462	1102.636	106.546	320.144
7	755.228	808.501	1075.372	106.546	310.456	524.054
9	968.826	1022.099	88.970	320.144	524.054	737.652
even pentatonic		+203.910	+470.781	+701.955	+905.865	+1172.736
2	203.910	407.820	674.691	905.865	1109.775	176.646
4	470.781	674.691	941.562	1172.736	176.646	443.517
6	701.955	905.865	1172.736	203.910	407.820	674.691
8	905.865	1109.775	176.646	407.820	611.73	878.601
10	1172.736	176.646	443.517	674.691	878.601	1145.472

Table2. 1-3-7-9-11 dekany with five transpositions (in cents)

5.MOBILE MUSIC

5.1.Satellite Gamelan

The Satellite Gamelan is an app for eighty players. It configures a hand-held phone as an instrument that is played as part of a performing ensemble. The instruments are easy to play, quick to learn and enable a large consort of players to collaborate in the discovery of a new microtonal language. In the process of using the app players perform a composition called *Transposed Dekany*.

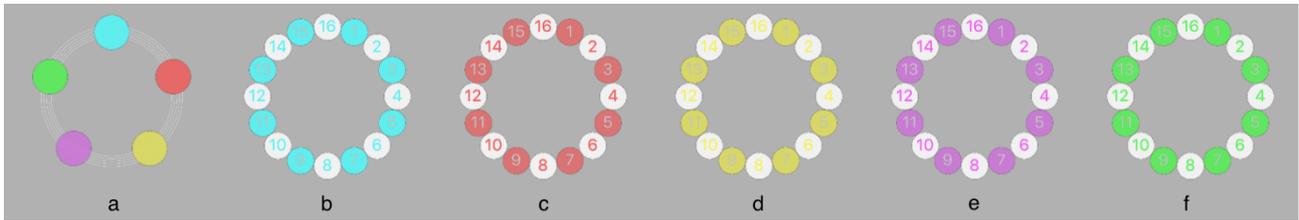


Figure 1. The left view [a] appears when the app is launched. Players select a family by touching 1-of-5 five coloured buttons. This takes them to another view [b, c, d, e or f] where they select 1-of-16 player settings. Each player has a different setting.

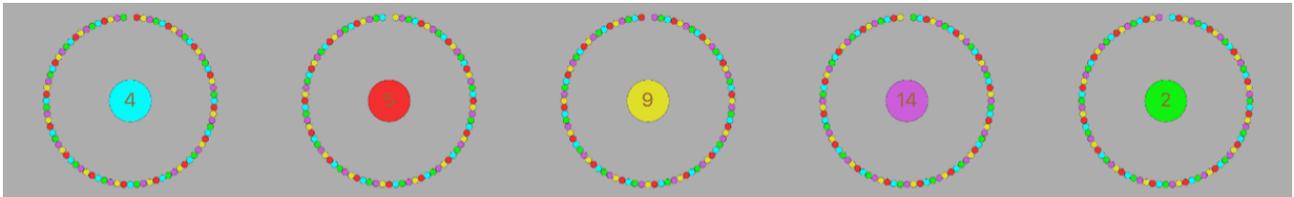


Figure 2. These views appear once players have selected player settings. Players wait for the rest of the consort to make their selection. On an agreed hand signal sign from a designated lead player every player hits the centre button on their screen.

5.2. Instrument Configuration and Tuning

Each player configures their phone as a mobile instrument that can be played in one of two ways. Firstly players can shake it like a handbell to produce bell sounds. Alternatively, players can touch buttons to produce a chorus of sine waves.

The full consort subdivides into five families each consisting of sixteen players. Players must first decide what part they will play in the consort, i.e. what family they join and which player they will be. Players then configure their instruments by selecting the family; this is done by touching one of the coloured buttons shown in Figure 1a; then selecting a player number; by touching one of sixteen buttons that appear as rosettes in Figures 1b, 1c, 1d, 1e or 1f.

Each family is tuned to a different scale transposition. This is selected automatically when players join a family (see Figure 1a). Once selected, an instrument remains tuned that way until the end of the performance. Selecting the player number chooses one of sixteen pitch and octave settings available within each family. Settings are uniquely assigned to each player. For example, every player has two bells each with its own pitch; every player has a pallet of chorus sounds played in uniquely assigned pitch and octave combinations.

5.3. Composition - *Transposed Dekany*

Once instrument configuration is complete, the app presents a screen view with a button surrounded by small circles as shown in Figure 2. All players are required to hit the centre button together. This starts a programmed sequence of states that automatically enables or disables the user interface during various states throughout the performance of *Transposed Dekanies*. The composition has 31 states in total as shown in Figure 3; the duration of each state is 24"; the total duration is 12'24".

The sequence of states generates every combination of the five families. Every family has one state where those players are heard on their own. There is also one state in which every family is heard playing together. Within those boundaries every combination of two, three or four families playing together is heard once and once only.

The sequence is displayed as an animated graphic score that updates whenever a state changes allowing players to monitor their progress in relation to their own family members and other families.

In any given state even numbered players enabled in every family create bell sounds while odd numbered players create chorus sounds. On alternate states players switch roles, i.e. odd players playing bells instead of chorus and vice versa.

Bell tones are played by shaking the phone like a conventional handbell; the centre circle is the bell clapper. Chorus tones are played by touching one of five points on the perimeter circle, or bell rim. These tones are synthesised using some of the first software instruments created by Jean-Claude Risset at Bell Labs [16].

Cues to play bell are shown in Figure 4a (odd) and 4b (even); the number in the centre is the selected player number. Cues to play chorus sounds are shown in Figures 4c for playing odd numbered notes of the dekanay and 4d for playing even numbered notes. The animated graphic score is shown below the cue in each figure.

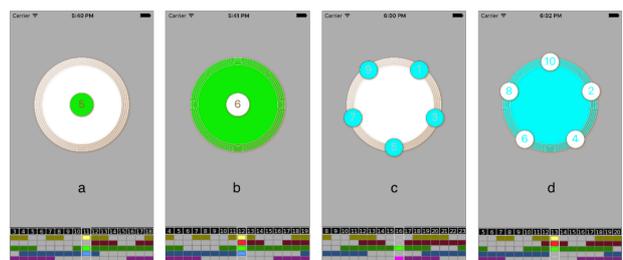


Figure 4. Odd and even bell cues (a,b) and chorus cues (c,d)

When the cue appears on the screen players may respond in their own time by ringing bells or playing chorus sounds within each 24 second window. Bell and chorus decay times are calculated on the fly from the start of the triggered event. They are timed to last until the end of the final active state for that family of instruments or until another event is initiated by the player, whichever occurs first.



Figure 3. The 5-bit sequence of 31 states enables every phone in each different family. Duration of each state is 24 seconds.

6. EXTENDED SCENARIOS

In its current form the app is a collaborative tool for manipulating a specific abstract data set, namely, the 5 harmonics that generate the scale. With changes the app could include the use of physical data.

6.1. Harmonic possibilities

Wilson's CPS method can generate many different dekanies scale each with its own specific harmonic flavour. To do this one might change the harmonics 1 3 7 9 and 11 to another set of values e.g. 1 3 7 13 and 17. Increasing the number of harmonic generators would also produce scales larger than 10 notes per octave.

With further modification the app would allow a user to change harmonics using a slider or touch menu to select new values; with a looped note sequence playing, this would allow a player to make a seamless transition from one harmonic flavour to the next. With further modification to the app, changes made by a user on one phone would also be broadcast to other phones.

Significantly, any change in the harmonic flavour is instantly recognisable irrespective of the musical expectations of the listener.

6.2. Bicycle flotilla

The app would lend itself to an event involving a flotilla of cyclists each using an iPhone as the mobile sound source, i.e. without headphones. The event would be an extension of the first Concert on Bicycles in 1983 when about 130 cyclists using ghetto-blasters tuned to a 1-hour broadcast radio program and cycled round Lake Burley Griffin, Canberra; half the cyclist travelled in a clockwise direction, the other half in a counter-clockwise direction, thereby passing one another at double the normal bicycle speed. The mono broadcast was transformed by the mass movement of multiple sound sources to create spatial artefacts discernible only to participating listeners. [16]

In this scenario participants install the app prior to the event. Instead of a ghetto-blasters, cyclists fasten an iPhone to their bicycle frame thereby making it responsive to accelerometer data. With some modification to the app, the handbell becomes a bicycle bell that responds to corrugations in the surface of the track while GPS data transforms the harmonic properties as cyclists enter new terrain.

7. CONCLUSIONS

The Satellite Gamelan app was first used on Nov 30 2012 as part of the Space Time Concerto Competition. On that occasion the performance involved a hook up of several concert venues spanning several continents and interconnected via a satellite link. The app has since been modified to support 80 players in the same venue. [17]

My ultimate objective is to take advantage of venue acoustics enjoyed by performers of conventional concert music. The Satellite Gamelan app used by a large ensemble can augment a standard contemporary concert program. For an established professional orchestra or large choir the economics are straight forward. *Transposed Dekany* is easy to play and quick to learn. It is currently available as an app that runs on a phone widely used by many musicians. For an existing ensemble of seasoned players the work can be made concert-ready in a single one-hour rehearsal. A performance lasts less than thirteen minutes, requires no special concert amplification and every instrument can be set up by its player without technical support. The missing ingredient so far is a conductor with the vision to convince eighty musicians to put their regular instrument aside (or rest their vocal chords) in order to present a new kind of chamber concerto where 'playing is left to the user' [5] yet executed to the highest standards of ensemble musicianship.

Clearly the problem lies not with the definition of sonification as this has evolved in the ICAD community but rather with limitations we place on what constitutes the 'organising principles of tonal music'. [6] That is a problem I have always had, and will probably continue to have, with music in general.

8. ACKNOWLEDGMENTS

I wish to acknowledge Etienne Deleflie for developing the templates that brought me up to speed with Objective C; and Mark Havryliv who helped troubleshoot maths code for Risset bell synthesis; and thanks to the reviewers whose comments helped focus this paper.

9. REFERENCES

1. Kramer, G. and Walker, B. "Sound science: Marking ten international conferences on auditory display". in Journal of ACM Transactions on Applied Perception (TAP) Volume 2 Issue 4, October 2005 pp. 383-388
2. Marrin, D. 'Infrasound Sources in the Environment: Oceanic, Atmospheric and Terrestrial' in Proceedings of Acoustics 2004, November 3-5 pp. 1-7
3. Cabrera, D., Ferguson, S. and Laing, G. "Considerations arising from the development of auditory alerts for air traffic control consoles" in Proceedings of the 12th International Conference on Auditory Display, London, UK, June 20-23, 2006 pp. 242-245
4. Barrass, S. Whitelaw, M. and Bailes, F. "Listening to the Mind Listening: An Analysis of Sonification Reviews, Designs and Correspondences" in Leonardo Music Journal Vol. 16, Noises Off: Sound Beyond Music (2006), pp. 13-19
5. Hermann, T. "Taxonomy and Definitions for Sonification and Auditory Display", in Proc. 14th Int. Conference on Auditory Display (ICAD), Paris, France, June 24-27, 2008.
6. Vickers, P. and Hogg, B. "Sonification ab- stracte/sonification concrete: An 'aesthetic perspective space' for classifying auditory displays in the ars musica domain," in ICAD 2006 - The 12th Meeting of the International Conference on Auditory Display, London, UK, June 20-23 2006, pp. 210-216.
7. Mott, I. (with collaborators Raszewski and Sosnin) <http://www.sounddesign.unimelb.edu.au/web/biogs/P000329b.htm>
8. Levin, G. et al. "*Dialtones (A Telesymphony)*" 2001 <http://www.flong.com/projects/telesymphony/index.html/retrieved> January 3 2016
9. Wang, G., Essl, G. and Penttinen, H. "Do mobile phones dream of electric orchestras?" Proceedings of the International Computer Music Conference (ICMC-08). August 2008
10. Schiemer, G. and Havryliv, M. "Pocket Gamelan: Tuneable trajectories for flying sources in Mandala 3 and Mandala 4". Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06) June 2006 pp. 37-42
11. Jenkins, J. 22 Australian Composers 1988 (21. Greg Schiemer) <http://www.rainerlinz.net/NMA/22CAC/TOC.html> retrieved April 30 2016
12. Partch, H. "Genesis Of A Music: An Account Of A Creative Work, Its Roots, And Its Fulfillments" Da Capo Press, 1979
13. Narushima, T. "Mapping the microtonal spectrum using Erv Wilson's Generalised Keyboard" PhD Thesis (2012) Routledge Oxford June 2016
14. Bregman, A. Auditory Stream Analysis No.5. Segregation of a melody from interfering tones <http://webpages.mcgill.ca/staff/Group2/abregm1/web/downloadstoc.htm#05> retrieved April 28 2016
15. Risset, J. Introductory Catalogue of Computer Synthesised Sounds, Bell Telephone Labs Murray Hill 1969
16. Schiemer, G.. (1994). Interactive Radio. Leonardo Music Journal, 4, 17-22
17. Schiemer, G. Satellite Gamelan (2012) concept video created for submission in the Space Time Concerto competition <https://www.youtube.com/watch?v=gfaZly6dhQA> retrieved January 2 2016

ORAL PAPERS

Tasks and Attention

CAN LISTENING TO MUSIC MAKE YOU TYPE BETTER? THE EFFECT OF MUSIC STYLE, VOCALS AND VOLUME ON TYPING PERFORMANCE

Anna Bramwell-Dicks

Department of Theatre, Film and Television,
Department of Computer Science
University of York, UK
anna.bramwell-dicks@york.ac.uk

Helen Petrie and Alistair Edwards

Department of Computer Science
University of York, UK
helen.petrie@york.ac.uk
alistair.edwards@york.ac.uk

ABSTRACT

Music psychologists have frequently shown that music affects people's behaviour. Applying this concept to work-related computing tasks has the potential to lead to improvements in a person's productivity, efficiency and effectiveness. This paper presents two quantitative experiments exploring whether transcription typing performance is affected when hearing a music accompaniment that includes vocals. The first experiment showed that classifying the typists as either slow or fast ability is important as there were significant interaction effects once this between group factor was included, with the accuracy of fast typists reduced when the music contained vocals. In the second experiment, a Dutch transcription typing task was added to manipulate task difficulty and the volume of playback was included as a between groups independent variable. When typing in Dutch the fast typists' speed was reduced with louder music. When typing in English the volume of music had little effect on typing speed for either the fast or slow typists. The fast typists achieved lower speeds when the loud volume music contained vocals, but with low volume music the inclusion of vocals in the background music did not have a noticeable effect on typing speed. The presence of vocals in the music reduced the accuracy of the text entry across the whole sample. Overall, these experiments show that the presence of vocals in background music reduces typing performance, but that we might be able to exploit instrumental music to improve performance in tasks involving typing with either low or high volume music.

1. INTRODUCTION

Hearing music can affect people's behaviour. From the speed of drinking a can of soda [1], to the type of wine bought in a supermarket [2], to children's performance in arithmetic tasks [3], there is substantial empirical evidence that hearing music affects what people do and how well they do it. Our research aims to identify whether we can exploit music to positively influence people while they are working.

Many people spend a large proportion of their working lives using a computer. A study in 2007 collected objective data of computer-use at work across 95 organisations from Europe, North America and Australasia [4]. Highest levels of computer use were

identified in the UK, where over a 4 week period employees spent an average of 16.8 hours per week using a computer, i.e. approximately 40% of their working week. Given that people spend a large proportion of their working hours using a computer, we ask whether hearing music while interacting with a computer can positively affect performance in work-related computing activities?

Kallinen [5] showed that the speed of reading news stories on mobile devices while in a noisy café was affected by the tempo of the background music. Significantly faster reading speeds were achieved when listening to music with a higher tempo. Although a café is not a typical work environment and reading the news on a mobile device is not a typical work task, this result shows there is potential to exploit the impact of particular parameters of music to improve performance in work-related tasks. This outcome leads to a refinement of our research aims – we want to identify how different parameters of music affect people's performance with work-related computing tasks.

Typing is a fundamental method used to interact with a computer. In 1937, Jensen investigated how hearing music affected the transcription typing performance of skilled typists [6]. The task required participants to copy text presented visually using a typewriter. The skilled typists made significantly more errors when listening to Jazz music than without music, or when accompanied by slow, melancholy Dirge music. Typing speeds were significantly slower with a Dirge music accompaniment than with Jazz music or without music. Jensen made no attempt to explain why Jazz increased error rates, or why Dirge music decreased typing speed. There does not seem to have been any follow up work to Jensen's experiment, so it is interesting to investigate the impact of music on typing performance further, and within a modern context.

In this paper, we present two experiments investigating the effect of different parameters of music on transcription typing speed and accuracy. Although transcription typing is not a typical work-related computing task, it is clearly-defined allowing us to retain tight control of the experiment, ensuring construct validity. The first experiment focuses on the impact of the presence or absence of vocals on typing performance, using two pieces of Rock music from different styles. The second experiment considers if the volume of the music is a significant factor affecting performance, and how the difficulty of the task mediates the influence of music.

2. VOCALS AND MUSIC STYLE

Shaffer [7] performed a number of experiments with a single, skilled typist, investigating how different verbal stimuli affect tran-



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

scription typing performance. This typist was able to maintain high levels of performance when copying from a visual source whilst reciting nursery rhymes. Further, the typist was able to transcribe material that was presented aurally through headphones in one ear, while concurrently repeating words heard through the other ear. For this skilled typist, typing performance was maintained in the presence of potentially conflicting verbal material. But does this outcome extend to being able to perform transcription typing while hearing music containing vocals? And would less skilled typists be affected in a similar way?

2.1. Method

The aim of this first experiment is to investigate whether the presence of vocals in a piece of music affects transcription typing performance with two pieces of Rock music with different styles. Rationally, one might expect that hearing one verbal source whilst copying different visually presented verbal material would be a difficult task which compromised performance. But, this argument contradicts the outcome from Shaffer's work which has been influential in the typing literature (e.g. in [8]) and we have identified no other empirical evidence to support this hypothesis.

2.1.1. Experimental Design and Hypotheses

Two experimental design paradigms were used. The first was a 1 by 5 design with a single independent variable (IV) for music with 5 levels (Alt Rock With Vocals, Alt Rock Without Vocals, Pop Rock With Vocals, Pop Rock Without Vocals and Without Music). This design paradigm focuses on the impact of the pieces of music as a whole and permits inclusion of a Without Music condition. The second 2 by 2 design paradigm focuses on manipulating the presence of vocals (2 levels – with and without vocals) and the style of the music (2 levels – Alt Rock and Pop Rock).

Typing speed and accuracy were measured as dependent variables (DVs). The number of transcribed characters was established and uncorrected errors counted using the Levenshtein Minimum String Distance algorithm. Measures of Characters per Minute (CPM) for speed, and Error Rate for accuracy were calculated using and (1) and (2).

$$\text{CPM} = \frac{\text{Number of Characters}}{\text{Length of Task (in minutes)}} \quad (1)$$

$$\text{Error Rate} = \frac{\text{Number of Errors}}{\text{Number of Characters}} * 100\% \quad (2)$$

2.1.2. Participants

The participants were recruited (22 male, 6 female) via an advert sent to University mailing lists, which stated that participants must be native English speakers, should not be dyslexic or have a hearing disability. Participants were aged between 18 and 44, 53% aged 18 to 24. Twenty-two participants were studying Computer Science (11 undergraduate and 11 PhD), with 4 Social Science PhD students and 2 professional researchers from the Humanities. All participants received a £10 Amazon voucher.

2.1.3. Materials and Environment

The experiment took place in a quiet usability laboratory on the University campus. The room contained 1 desktop computer running Windows XP which the participant used. The typing tasks

were hosted on a bespoke website accessed through the FireFox browser with spellcheck disabled. The website displayed two text boxes side by side. The left hand text box was non-editable and displayed the text to be copied. The participants entered their transcription in the right hand text box, the contents of which was saved on the webserver once the task had been completed. The experimenter used a MacBook Pro to control the music playback via Audacity connected to a pair of Philips SPA 2210 2.0 speakers.

Two pieces of Rock music were used in this experiment, both taken from Cambridge Music Technology website (<http://www.cambridge-mt.com/>). The first song, "Atrophy" by The Doppler Shift, was described as Alt Rock style, the second, "Big Dummy Shake" by Moosmusic is Pop Rock. For the experiment to have strong internal validity, the two variations of the music stimulus needed to be the same with the exception of the presence of vocals. Multi-track recordings were mixed down into two separate versions of each song. One version included all the vocal tracks while the other was a mix of only the instrumental tracks. It was also important that the participants were equally familiar or unfamiliar with the music. None of the participants could recall hearing either song prior to the experiment.

There were 5 conditions in the experiment. The order of presentation of music style was alternated so that the participants did not hear the two Alt Rock or Pop Rock conditions back to back. The style of music and inclusion of vocals were both counterbalanced to avoid learning and fatigue effects. The position of the Without Music condition was also systematically varied. Five passages from different chapters of *The Outlaw of Torn* (Edgar Rice Burroughs) were displayed to participants as the text for transcription. The order of presentation of the passages was varied to avoid connections between experiment condition and each text passage.

2.1.4. Procedure

The experimenter began by telling the participants that they would complete a number of transcription typing tasks while listening to different pieces of music to see how the music affected their performance. They were told they could correct any errors but that they should only use the keyboard and not the mouse and that they should type as naturally as possible. The participants completed a 30 second practice transcription typing task, followed by 5 transcription typing tasks (4 with music, and 1 without music). The experimenter started timing the task when the participant began typing and stopped the participant after 4.5 minutes. Once the participants had finished all 5 typing tasks, they completed demographic questionnaires on paper.

2.2. Results

2.2.1. Typing Speed

Figure 1 shows a histogram of typing speed for all conditions. Visual inspection suggests the distribution may be bimodal rather than normal with a crossing at around 340 CPM as shown by the overlay which is an approximation of two normal distributions. There may also be 5 outlying data points above 520 CPM.

Scatterplots were created for the Alt Rock and Pop Rock conditions (Figure 2) to compare the speeds achieved by each participant within a single piece of music and establish an appropriate threshold level for separation into 'slow' and 'fast' typist groups. The histogram in 1 suggested a crossing at about 340 CPM. Inspection of the scatterplots led to a refinement of this threshold

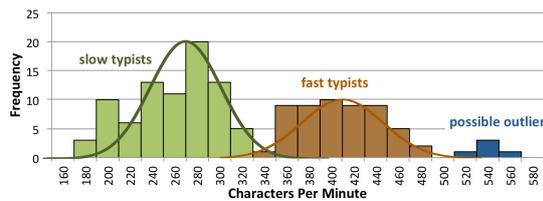


Figure 1: Histogram of CPM showing typist ability grouping with normal distribution overlays and a possible outlier.

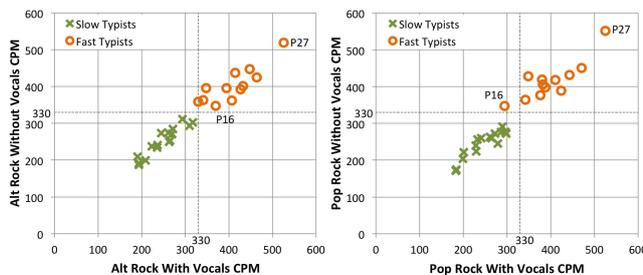


Figure 2: CPM scatterplots for both music styles, showing typist ability classification with the 330 CPM threshold applied. Speeds achieved by participants 16 and 27 are highlighted.

value to 330 CPM, resulting in 16 participants classified as slow typists and 12 as fast typists. The only participant that did not easily fit within this classification structure was participant 16 who achieved less than 330 CPM in the Pop Rock Without Vocals condition, but higher than 350 CPM in all other conditions. As 80% of tasks were completed high above the 330 CPM threshold, participant 16 was classified as a fast typist. The scatterplots also verify that a single participant (P27) achieved over 500 CPM in all tasks.

Repeated-measures ANOVAs were performed on the data using both the 1 by 5 and 2 by 2 experimental design paradigms. A second analysis using mixed designs ANOVAs allowed inclusion of typing ability as a between-participants factor. An alpha level of 0.05 was used in all statistical tests.

All of the underlying assumptions for Repeated Measures ANOVAs were met when the data was treated as a single distribution. The assumption of normality of distribution was assessed for each condition separately using Shapiro-Wilk's test. Despite the bimodal characteristic observed through visual inspection, all of the distributions were found not to deviate significantly from normality. When treated as a single distribution, the Repeated Measures ANOVA resulted in no significant omnibus effects or interactions in the 1 by 5 or 2 by 2 design paradigms.

When typing ability was included in the analysis as a between groups factor a statistical outlier was introduced. The analysis was performed both with and without the outlier included in the dataset. This outlier had no effect on the outcomes using either the 1 by 5 or 2 by 2 analysis paradigms. All other assumptions for Mixed Design ANOVAs were met.

With the 1 by 5 experimental design paradigm the effect of the music was not significant, $F(4,104)=1.20$, n.s., and neither was the interaction between music and typing ability, $F(4,104)=1.99$, n.s. Typist ability was a significant between groups factor, $F(1,26)=90.42$, $p<0.001$, $\eta_p^2=0.78$, with higher speeds achieved by the fast group ($M=408.17$, $SD=53.07$) than the slow group

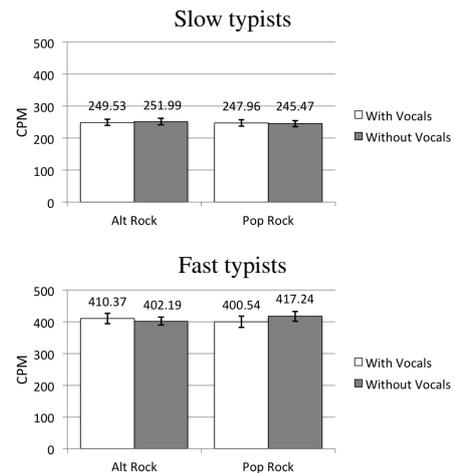


Figure 3: Significant 3-way interaction between music style and typing ability. Error bars show the standard error.

($M=249.80$, $SD=39.31$). This result is expected given the typist ability classification was applied post hoc based on the speeds achieved by each participant.

Using the 2 by 2 analysis paradigm, there were no omnibus effects identified for the vocals condition, $F(1,26)=9.32$, n.s., nor music style, $F(1,26)=0.06$, n.s. A trend towards a significant interaction between music style and vocals was identified, $F(1,26)=3.83$, $p=0.06$, $\eta_p^2=0.13$, which suggests any effect of vocals on speed may differ depending on the style of the music, but there is insufficient evidence from this experiment to be sure.

The 3-way interaction between music style, vocals condition and typing ability was significant, $F(1, 26)=8.57$, $p=0.007$, $\eta_p^2=0.25$ (Figure 3). A simple 2-way interaction between music style and vocals was significant for fast typists, $F(1,11)=7.50$, $p=0.02$, $\eta_p^2=0.41$, but not for the slow typists, $F(1,15)=0.77$, n.s. When accompanied by Alt Rock music the fast typists achieved a mean typing speed that was 8 CPM higher with vocals than without vocals (Alt Rock With Vocals: $M=410.37$, $SD=56.93$, Alt Rock Without Vocals: $M=402.19$, $SD=42.63$). For Pop Rock music, a 17 CPM difference in typing speeds was identified between the with and without vocals conditions with higher speeds achieved without vocals (Pop Rock With Vocals: $M=400.54$, $SD=61.86$; Pop Rock Without Vocals: $M=417.24$, $SD=52.96$). However, the simple main effect for vocals was not significant for either music style (Alt Rock: $F(1,11)=0.83$, n.s., Pop Rock: $F(1,11)=3.27$).

Typist ability was a significant between groups factor, $F(1,26)=90.28$, $p<0.001$, $\eta_p^2=0.78$, with a lower mean speed achieved by the slow typists than the fast typists (Slow: $M=248.74$, $SD=39.04$, Fast: $M=407.48$, $SD=53.60$).

2.2.2. Typing Accuracy

The error rate data had a negative skew so a square root transformation was applied. Before transformation 4 of 5 conditions (80%) were strongly non-normal by the Shapiro-Wilk's test ($p<.002$), with the 5th condition resulting in moderate non-normality ($p=0.03$). After the square root transformation was applied to the data none of the distributions deviated significantly from normality. Figure 4 show a histogram of the transformed

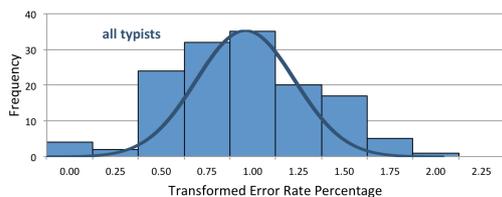


Figure 4: Histogram of Transformed Error Rate Percentage.

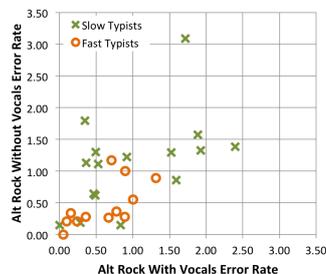


Figure 5: Error Rate scatterplot for Alt Rock music conditions, showing typist ability classification with 330 CPM threshold.

error rate percentage which does not contain the bimodal characteristic observed in the typing speed distribution.

A scatterplot of the two Alt Rock conditions is included in Figure 5 which shows the pattern of error rates achieved by each participant with the slow or fast typist classification. The Pop Rock scatterplot contained similar characteristics. The scatterplot shows that the range of error rates for the slow typists is greater than for the fast typists group. But, there is a clear mix with some slow typists making few errors, but some fast typists making a comparatively large number.

Treating the transformed error rate percentage data as a single distribution resulted in no significant omnibus effects or interactions when analysed using a repeated measures ANOVA for either paradigm. There were no assumption violations when the dataset was considered as a whole in either the 1 by 5 or 2 by 2 paradigms.

With the typist ability classification applied, two outlying data points are introduced. These two outliers had no impact on the outcomes of the analysis in either paradigm. There were no other violations of assumptions for Mixed Designs ANOVA.

Inclusion of the typist ability classification in the 1 by 5 analysis paradigm resulted in a non-significant omnibus effect for music, $F(4,104)=0.19$, n.s., and a non-significant interaction between music and typist ability, $F(4,104)=1.22$, n.s. Typist ability group was a significant between groups factor, $F(1,26)=6.15$, $p=0.02$, $\eta_p^2=0.19$. The mean transformed error rate for the slow typists was higher than for the fast typists (Slow: $M=0.95\%$, $SD=0.39\%$, Fast: $M=0.67\%$, $SD=0.31\%$) meaning that the fast typists made significantly fewer errors.

Performing the analysis using the 2 by 2 design paradigm with typist ability classification resulted in no significant omnibus effects. There was a significant 2-way interaction between vocals and typist ability group, $F(1,26)=7.24$, $p=0.01$, $\eta_p^2=0.22$. Figure 6 shows that the slow typists made more errors when hearing music that did not contain vocals. In the fast typists group, the magnitude of the difference in transformed error rate percentage between

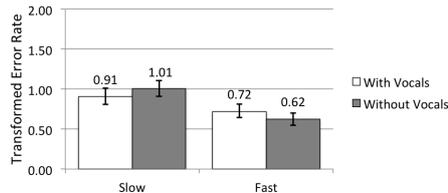


Figure 6: Significant interaction between vocals and typing ability. Error bars show the standard error.

with and without vocals condition is the same as for the slow typists, but in the opposite direction with fewer errors without vocals. A simple main effects analysis for vocals was significant for fast typists, $F(1,11)=6.92$, $p=0.02$, $\eta_p^2=0.39$) but not for slow typists, $F(1,15)=0.22$, n.s.

2.3. Discussion

The initial aims of this experiment were taken with the assumption that the data collected would be treated as a single distribution. However, when the dataset was analysed as a single distribution for both typing speed and accuracy measures, there were no significant omnibus effects or interactions. Visual inspection of the typing speed distribution suggested a bimodal distribution, leading to a means of classifying participants as either slow or fast and analysing the data to consider whether typist ability matters. The analysis with typist ability classification included led to a number of significant interactions involving this between groups factor. When typist ability classification is applied as a between groups factor the degrees of freedom for the error term are reduced by 1. This reduction leads to a more appropriate partitioning of the error variance and as the model fits the distribution better the statistical power is increased.

Typing ability grouping was a significant between groups factor for both the typing speed and accuracy measures. The former is expected given that typing speed determined participant allocation to a group. However, although the error rate scatterplots showed a mix of error rates for the slow and fast typists, there was still a significant difference in transformed error rate between the groups with the fast group making fewer errors than the slow group overall. This result is likely to be due to the differences in the spread of error rates by achieved by each group, with a smaller spread for the fast typists than for the slow typists. This outcome indicates that the fast typists were better in both the performance measures, suggesting that they are more skilled typists and not simply faster.

The fast typists made fewer errors when hearing music without vocals. In contrast, the slow typists made more errors without vocals, though the post hoc simple effect analysis was non-significant for the slow group. The non-significant simple effect may be due to the relatively small sample size when the ability classification is applied as a between groups factor. It is possible that the slow typists had to concentrate harder when hearing music that contained vocals, and as a result noticed and corrected their errors more frequently. The differences in effect of vocals on accuracy rates for the slow and fast typists groups warrants further investigation.

Typing ability was not intended to be included as a factor in this first experiment. However, the analysis has shown that this post hoc classification is a very important factor that must be considered in future experiments.

3. VOCALS, VOLUME AND TASK DIFFICULTY

The outcomes from the first experiment showed that the impact of the presence of vocals in a music accompaniment on transcription typing performance is worth pursuing. But, separating the sample into slow and fast ability categories resulted in small group sizes which potentially affected the outcomes. A larger sample of participants is needed to accommodate the ability classification. Sessions in first experiment using a laboratory-approach took 1 hour per participant. To maximise efficiency for the second experiment a classroom-based methodology was used. This approach allowed us to perform an experiment with at least 50 participants in 4 hours.

An online methodology could have been used to get high participation rates in less time with the experiment hosted online and participants completing tasks remotely. This approach would allow us to reach large numbers of participants without the experimenter needing to be present. But, conducting a remote experiment also removes the experimenter's ability to carefully regulate the environment. For example, the experimenter would be unable to tightly control the volume of playback of the music which has previously been shown to affect people (e.g. in [9]). The impact of volume needs to be investigated in a controlled environment prior to any online experiment. So, this experiment includes volume as a between groups IV. If volume does not affect typing performance in a controlled environment then this provides evidence that an online methodology would be appropriate. However, if volume is shown to be a factor that affects performance then careful consideration of how to accommodate the lack of control of volume within the experiment will be needed before any future online approach can take advantage of the large pool of potential participants.

The first experiment also indicated that any effect of the accompanying music differs according to participant skill level. Therefore, it is interesting to investigate whether the effect of accompanying music is consistent across tasks of varying difficulties. When transcription typing in one's own language, performance levels have been shown to be maintained even when words in the presented text are randomly ordered so that the text passage does not have semantic meaning [8]. As such, a simple reordering of the words is unlikely to make the task more difficult and another approach is needed. In this experiment the difficulty of the task is manipulated by including typing tasks in Dutch, which is expected to be a language that than is unknown to the participants. We expect that in harder transcription tasks, the extent of the effect due to the music will be reduced.

3.1. Method

The primary aim of this experiment was to investigate whether the presence of vocals in the music affects typing performance and how typist ability interacts with vocals. This experiment also considers if the volume of music affects performance and whether there is any interaction between the volume and presence of vocals. Finally, this experiment looks at whether an increase in task difficulty, manipulated by including a Dutch transcription typing task, reduces the affect of music on typing performance.

3.1.1. Experimental Design and Hypotheses

This experiment used a 2 by 2 by 2 mixed design. The between groups IV was music volume (low or high), the within groups IVs were vocals (2 levels – with and without vocals) and text (2 levels – English and Dutch language). It was expected that typing ability

would be included as a between groups factor but this would be incorporated post hoc based on the distributions of speed. The DVs for typing performance were speed measured in CPM, calculated using (1) and error rate calculated using (2).

3.1.2. Participants and Experimenters

Both the participants and experimenters were first year undergraduates studying Computer Science. The students were taking a Human Computer Interaction module where they learn how to perform experiments. Within one practical class the students all get to experience being a participant and being an experimenter. Fifty-five students (8 female, 47 male) were participants in this experiment, while another 55 acted as experimenter. All of the participants were aged between 18 and 24. Nine participants were non-native speakers of English, but all had demonstrated competency in English by achieving International English Language Test (IELT) scores in excess of 6.5. None of the participants in this experiment were familiar with the Dutch language. The participants were not asked to report whether they had a hearing disability or dyslexia as this would have been inappropriate given the classroom context. After the experiment, 5 participants were randomly selected to win £10 Amazon gift vouchers. The 3 best typists received £30, £20 and £10 Amazon gift vouchers.

3.1.3. Materials

The with and without vocals versions of the Alt Rock style music were used in this experiment. The Alt Rock style was chosen over the Pop Rock because the outcomes from the first experiment suggested that vocals might have had a significant effect on performance in the Pop Rock music condition. We did not want to constrain the generalisations of the outcomes from these two experiments to a single style and as there was no clear evidence that vocals had a negative effect with the Alt Rock music, it is more interesting to use this piece as the stimulus.

Four text passages were used for the typing tasks, two in English and two in Dutch. The English text passages again came from *The Outlaw of Torn* by Edgar Rice Burroughs while the Dutch text passages were taken from *Op Eigen Wieken*, a Dutch translation of Louisa May Alcott's *Good Wives*. The Dutch language was chosen for the difficult text condition because the Dutch alphabet is similar to English. Accents were removed from all characters.

Again bespoke webpages hosted the typing tasks. These webpages controlled the music, which began playback when the participant made their first keypress. After 4.5 minutes the webpage generated an alert box to end the task. Volume of playback was also set by the webpage. Half of the participants used webpages which played the music at 100% volume, while playback was set to 50% volume for the other participants. The suitability of playback volume (i.e. not too quiet or loud) had previously been verified in a pilot test.

The text passages were counterbalanced so that there were no pairings between the pieces of music and the different texts. The order of experimental conditions was systematically varied to avoid fatigue and practice effects.

All participants completed the typing tasks on personal computers running Windows 7 in the department's software laboratory. The experimenter was instructed to use the Firefox browser and disable spellcheck. Participants were given sets of inexpensive headphones (Astro Tools ATA 1144) to use.

3.1.4. Procedure

The experiment took place during two practical classes for a first year undergraduate module. The students worked in pairs, with one acting as experimenter and the other as the participant. Step-by-step instructions for running the experiment were given to the student acting as experimenter to ensure the procedure was followed correctly, including a script of what to say to the participant. The instructions and process had been thoroughly piloted with pairs of students, including non-native English speakers.

Each experimenter began by explaining the premise of the experiment to their participant. The participants were told to type as naturally as possible without prioritising speed or accuracy. The experimenter then set the computer's volume to maximum and informed consent was taken.

The first typing task allowed both students to practice the process. The music began playing on the participant's first keypress. After 30 seconds an alert message ended the practice task. The first experimental typing task was then loaded into the web browser.

Each of the experimental typing tasks lasted 4.5 minutes. After finishing all 4 typing tasks, participants completed demographic questionnaires. When all the participants had finished the experiment, the first author debriefed the class.

3.2. Results

3.2.1. Typing Speed

Figure 7 shows a histogram of the CPM data for the high and low volume conditions combined. The crossing point for the two distributions is again at approximately 330 CPM. Figure 8 is a scatterplot of CPM values for all participants in both of the English text conditions with classification of typists into fast and slow ability groups using a threshold value of 330 CPM. Visual inspection of the scatterplot verifies a clear gap between the slow and fast typists. The proportion of fast typists in this sample was smaller than in the first experiment, with just 11 out of 55 participants classified as fast typists. Of these 11, 6 were in the loud volume condition. Of the 42 slow participants, 22 heard loud music. Participants 14 and 28, both non-native English speakers, were the slowest typists. These participants were removed from the analysis as their Dutch and English typing speeds were similar, implying that both languages were unfamiliar. The other non-native speakers of English achieved lower speeds in Dutch than English, making them suitable for inclusion in the analysis. Participant 54, a non-native English speaker, was the fastest typist in this experiment.

With all the participants considered as a single dataset, 3 of the 8 distributions (37.5%) were strongly non-normal by a Shapiro-Wilk's test ($p < .015$). But, with participants classified as a fast or slow typist only 1 of the 16 conditions (6.25%) had a strong non-normal distribution ($p = 0.015$). This confirms that, the typing ability classification improves the normality of the distributions. Given the importance of the typing ability classification shown in the first experiment a decision was taken to classify the participants as either fast or slow typists according to their achieved typing speeds in English for the inferential analysis. The data was not analysed as a single distribution in this experiment.

A Mixed Design ANOVA was performed on the CPM data. Volume level (low and high) and typist ability (slow and fast) were between-participants IVs. Presence of vocals (with and without) and language of presented text (English and Dutch) were the within-participant IVs. The assumption of homogeneity of vari-

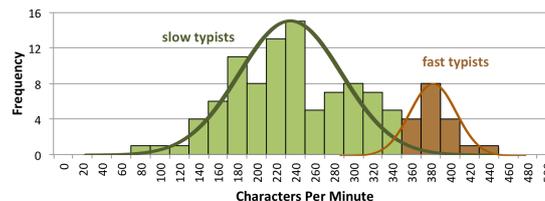


Figure 7: Histogram of CPM showing typist ability grouping with normal distribution overlays.

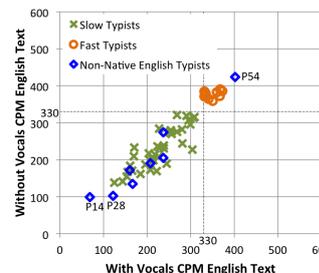


Figure 8: CPM scatterplot of English text condition showing typist ability classification. Non-native English speakers are highlighted, with attention drawn to the speeds achieved by participants 54, 14 and 28.

ance was not met, with violations in both of the English text conditions, (with vocals: $F(3,49)=3.43$, $p=0.02$, without vocals: $F(3,49)=4.93$, $p=0.005$). As the group sizes are unequal, this assumption violation requires further investigation. When sample sizes are different, heterogeneity of error variance is problematic if the larger variance is associated with the smaller group as the resulting F statistic is inflated leading to a higher chance of a type I error [11]. If the larger variance occurs in the larger group, the F statistic is conservative, with risk of a type II error. Inspection of box plots of each condition showed that smaller variances were associated with the fast typist group which had smaller numbers of participants. As such, the F statistic is at risk of being conservative rather than inflated so the analysis can proceed. Six participants were statistical outliers. The analysis was performed both with and without these outliers included in the dataset and the outcomes were not affected by their inclusion. All other assumptions for Mixed Design ANOVAs were met.

Table 1 presents the outcomes from this analysis. The language of the presented text had a significant omnibus effect with faster speeds achieved when typing in English ($M=254.16$, $SD=74.95$) over Dutch ($M=162.60$, $SD=47.42$). There was also a significant omnibus effect for vocals, with higher speeds achieved without vocals ($M=209.04$, $SD=63.50$) than with vocals ($M=206.82$, $SD=58.86$). Volume was not identified as an overall significant between groups factor, $F(1,51)=0.32$, n.s., though unsurprisingly the overall effect of typing ability was significant, $F(1,51)=81.32$, $p < 0.001$, $\eta_p^2=0.62$, with higher speeds achieved by those allocated to the fast typists group. The significant two-way interactions are not discussed in detail in this paper as they are subsumed by significant three-way interactions.

The three-way interaction between vocals, volume and typing ability was significant meaning that the two-way interaction between vocals and volume is different across levels of ability. This

Effect	Test	p	η_p^2
Vocals	F(1,49)=7.746	0.008	0.137
Vocals x Volume*	F(1,49)=4.296	0.043	0.081
Vocals x Ability*	F(1,49)=8.467	0.005	0.147
Vocals x Volume* x Ability*	F(1,49)=6.410	0.015	0.116
Text	F(1,49)=306.715	<0.001	0.862
Text x Ability*	F(1,49)=24.797	<0.001	0.336
Text x Volume*	F(1,49)=8.558	0.005	0.149
Text x Vocals	F(1,49)=3.630	n.s.	
Text x Volume* x Ability*	F(1,49)=7.799	0.007	0.137
Text x Vocals x Ability*	F(1,49)=2.516	n.s.	
Text x Vocals x Volume*	F(1,49)=0.093	n.s.	
Text x Vocals x Volume* x Ability*	F(1,49)=0.871	n.s.	
Volume*	F(1,49)=0.905	n.s.	
Ability*	F(1,49)=90.805	<0.001	0.650
Volume* x Ability*	F(1,49)=3.395	n.s.	

Table 1: Summary of Outcomes from 2 by 2 by 2 ANOVA. The * indicates a between groups factor.

three-way interaction is shown in Figure 10. The simple two-way interaction between vocals and volume was significant for fast typists, $F(1,9)=11.93$, $p = 0.007$, $\eta_p^2=0.57$, but not for slow typists, $F(1,40)=0.23$, n.s. For the fast typists, the simple main effect of vocals was significant with high volume music, $F(1,5)=26.68$, $p=0.004$, $\eta_p^2=0.84$, but not with low volume music, $F(1,4)=0.43$, n.s. When hearing high volume music, the fast typists were slower with music that contained vocals ($M=266.96$, $SD=96.31$) than without vocals ($M=294.50$, $SD=101.66$). The simple main effect of volume was significant without vocals, $F(1,9)=14.47$, $p=0.004$, $\eta_p^2=0.617$, but not with vocals, $F(1,9)=2.752$, n.s. When hearing music that contained vocals, the fast typists achieved higher speeds in the low volume condition ($M=311.98$, $SD=65.62$) than in the high volume condition ($M=266.96$, $SD=96.31$). These results suggest that for best performance faster typists should avoid listening to music at a loud volume if it contains vocals, but if the volume of the music is lower, the presence of vocals does not have a noticeable effect.

The three-way interaction between text, vocals and typing ability was also significant, indicating that the interaction between text and vocals is different across levels of ability. This three-way interaction is shown in Figure 9. The simple two way-interaction between text and volume was significant for the fast typists, $F(1,9)=5.64$, $p=0.04$, $\eta_p^2=0.39$, but not for the slow typists, $F(1,40)=0.03$, n.s. The simple main effect of text was significant for both the low, $F(1,4)=168.64$, $p<0.001$, $\eta_p^2=0.98$, and high volume groups, $F(1,5)=44.59$, $p=0.001$, $\eta_p^2=0.90$. The fast typists achieved higher speeds at both volume levels when typing in English (Low volume, $M=365.64$, $SD=15.14$; High volume, $M=368.54$, $SD=29.71$) over typing in Dutch (Low volume, $M=261.22$, $SD=26.65$; High volume, $M=192.93$, $SD=48.17$). The simple main effect of volume was significant when typing in Dutch, $F(1,9)=8.06$, $p=0.02$, $\eta_p^2=0.47$, but not in English,

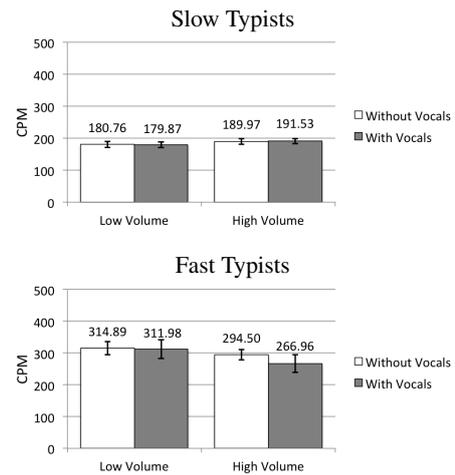


Figure 9: Significant 3-way interaction between volume, vocals and typing ability. Error bars show the standard error.

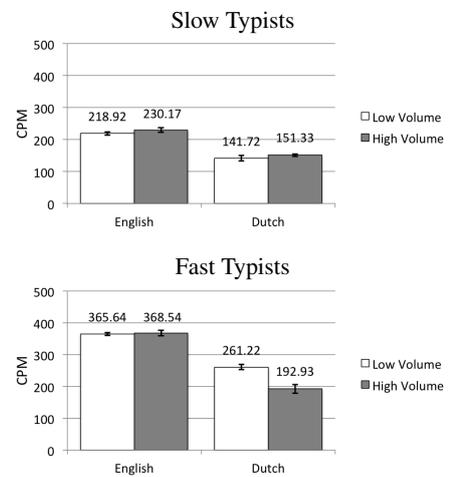


Figure 10: Significant 3-way interaction between text, vocals and typing ability. Error bars show the standard error.

$F(1,9)=0.06$, n.s. When typing in Dutch, the fast typists achieved higher speeds with low volume music ($M=261.22$, $SD=26.65$) and slower speeds with high volume music ($M=192.93$, $SD=48.17$). When typing in a familiar language, the volume of the music did not have an impact on the speeds achieved by the fast typists. However, when typing in an unknown language, the high volume music resulted in lower speeds.

3.2.2. Typing Accuracy

As in the first experiment, the error rate data had a strong negative skew. A logarithmic transformation was applied as a square root transformation did not improve the normality of the distributions sufficiently. Before transformation, 12 of 16 conditions (75%) were strongly non-normal ($p<.002$) by Shapiro-Wilk's. After transformation this reduced to just 3 of 16 conditions (18.75%) with moderate non-normal distributions ($p>.03$).

Analysing using a mixed ANOVA on the transformed data as

a single distribution resulted in a significant omnibus effect for text, $F(1,51)=6.00$, $p=0.02$, $\eta_p^2=0.11$. The transformed error rate was higher with Dutch ($M=0.23\%$, $SD=0.59\%$) than with English text ($M=0.05\%$, $SD=0.64\%$). There was a significant omnibus effect for vocals, $F(1,51)=5.64$, $p=0.02$, $\eta_p^2=0.10$, with higher error rates with vocals ($M=0.24\%$, $SD=0.67\%$) than without vocals ($M=0.07\%$, $SD=0.60\%$). Volume was not a significant factor, $F(1,51)=0.36$, n.s., and there were no significant interactions.

3.3. Discussion

In this experiment, when transcription typing while listening to music that contained vocals the participants typed significantly slower with higher error rates than when accompanied by instrumental music. This higher performance without vocals indicates that the typing task was easier with instrumental music and suggests that if task performance is important, people should choose to listen to instrumental music over music containing vocals.

Although there was a significant omnibus effect for vocals, the interaction between vocals, volume and typist ability is particularly interesting as for the slow typists the omnibus effect of vocals is not evident. Even with comparatively small numbers of fast typists in the experiment, for whom the effect of vocals is clear with the high volume music, the omnibus effect for vocals is achieved which demonstrates the strength of the effect for the fast typists. These experiments are quite different in nature and objective to Shaffer's [7] work, but the outcomes do seem to contradict the assertion that skilled typists are not affected by concurrent verbal material. This could be due to the differences between hearing spoken words and listening to music containing vocals, but further investigation is needed.

The other significant three-way interaction between volume, text and typing ability is also interesting as it clearly shows the effect of the background music on typing speed is connected to the difficulty of the task, but not in the way that we initially proposed. When typing in English, a task that was comparatively easy for all participants, the volume of the music had no obvious effect. However, when typing in Dutch, the fast typists group achieved significantly lower speeds when accompanied by louder music. This suggests that when the skilled participants had to concentrate more on the task, the louder music was more distracting and reduced their performance. For simple tasks, we might conclude that volume does not have a big effect but in more complex situations, the volume of the music should be carefully considered.

In this experiment, the participants heard music in all of the experimental conditions. There was no condition where the participants typed without music and so their base typing performance level was not included within the analysis. This limitation was caused primarily by the setting of the experiment, i.e. as part of a practical class rather than using a laboratory. The classroom is a busy environment, which is typically quiet, but not silent. It would have been hard to control the room sufficiently to eliminate confounds for a no-music IV so one was not included.

The modest number of fast typists in the experiment is a further limitation. For the accuracy DV, speed classification did not factor in the analysis so it is not an issue. But, for the speed analysis only 11 of 53 typists were classified as fast, limiting the generalisability of the outcomes. Due to the setting of the experiment, it was not possible to select participants with a range of abilities and no further participants could be added without introducing confounds. Experiments with a large group of fast typists are needed

to validate the typing speed outcomes.

4. CONCLUSIONS

This paper has highlighted the importance of the differences between skilled and novice typists. The typing literature, much of which was written in the 1980s, has typically focused on working with typists that had been trained. But, these two experiments have shown that the impact of different dimensions of background music on typing performance is dependent on typing skill level.

Both experiments in this paper have also shown that vocals can have a negative affect on transcription typing performance. Although this outcome may seem obvious, these experiments provide empirical evidence that was previously missing from the literature. The second experiment demonstrated that high volume music can have a negative affect on typing speed, so we recommend that careful consideration of the researchers' inability to control volume be taken in any future experiments performed online. This result also indicates that different dimensions of music have different effects, and suggests that we might be able to exploit loud, instrumental music to improve performance when working at a computer, especially for tasks where the user's level of skill is high. But, for difficult tasks, skilled user's should choose low volume music to maintain high levels of performance.

5. REFERENCES

- [1] Heather McElrea and Lionel Standing. Fast Music Causes Fast Drinking. *Perceptual and Motor Skills* 75, 1992, 362 – 362. Issue 2.
- [2] Ronald E. Milliman. Using Background Music to Affect the Behavior of Supermarket Shoppers. *Journal of Marketing* 46, 3, 1982, 86 – 91.
- [3] Teresa Lesiuk. The effect of music listening on work performance. *Psychology of Music* 33, 2, 2005, 173 – 191.
- [4] Kevin Taylor. 2007. An analysis of computer use across 95 organisations in Europe, North America and Australasia. Technical Report. Wellnomics Ltd.
- [5] K. Kallinen. Reading news from a pocket computer in a distracting environment: Effects of tempo of background music. *Computers in Human Behavior*, 18, 2002, 537 – 551.
- [6] M. B. Jensen. The influence of jazz and dirge music upon speed and accuracy of typing. *Journal of Educational Psychology*. 22, 1937, 458 – 462. Issue 6.
- [7] L. H. Shaffer. 1975. Multiple attention in continuous verbal tasks. In *Attention and Performance V*, P. M. A. rabbit and S. Dornic (Eds.). New York: Academic Press, 157 – 167.
- [8] Timothy A. Salthouse. Perceptual, Cognitive, and Motoric Aspects of Transcription Typing. *Psychological Bulletin* 99, 1986, 303 – 319. Issue 3.
- [9] J. Edworthy and H Waring. The effects of music tempo and loudness level on treadmill exercise. *Ergonomics* 49, 15, 2006, 1597 – 1610.
- [10] L. J. West and Y. Sabban. Hierarchy of stroking habits at the typewriter. *Journal of Applied Psychology* 67, 1982, 370 – 376.
- [11] Barbara G Tabachnick and Linda S Fidell. 2007. *Experimental Designs Using ANOVA*. Thomson/Brooks/Cole.

TUNING INTO THE TASK: SONIC ENVIRONMENTAL CUES AND MENTAL TASK SWITCHING

Toby Gifford

Griffith University
Queensland Conservatorium of Music
140 Grey St. Southbank QLD 4169
t.gifford@griffith.edu.au

ABSTRACT

This position paper suggests a novel approach to enhancing productivity for professionals whose core business is deep thinking, by manipulation of the sonic environment. Approaching the issue from the perspective of sound-design, it proposes the composition and algorithmic generation of background soundscapes that promote a psychological state of *flow* [1], and can become mentally associated with particular tasks through exposure, so as to facilitate task switching by switching soundscapes.

These background soundscapes are intended to mask distracting clatter, oppressive quiet, and other suboptimal sonic environments frequently encountered in office workplaces. Consequently, I call them *active-silences*—soundscapes designed to be *not heard*, although they may be relatively loud. The most commonly used active-silence is white noise, though there are surprisingly diverse other approaches to crafting active-silence. This variety suggests the possibility of training associations that pair distinct active-silences with distinct mental tasks.

1. INTRODUCTION

Cordoning off blocks of uninterrupted time is a serious challenge for many professionals. In today's 'on-line' society multitasking is the norm and extended periods of time to concentrate on a single task are scarce. Yet deep thinking is core business for many professions. Moreover, some professions require several distinct types of deep thinking, and switching between them can be difficult. For example, one of the most challenging aspects of academic life is juggling the responsibilities of teaching and research. I usually need at least an hour to shift mental gears between these categories, particularly switching into the deep thinking required for research, writing, computer programming, or creative development.

Psychologists refer to this as *task switching* [2]. Many authors have highlighted the need for ways to minimise switching time—the time lost in switching to a new task—in order to improve productivity and decrease stress in the workplace [3]. Environmental cues can aid resumption of suspended mental tasks, particularly by conditioning through association [4] and productivity tools that deliberately manipulate the visual environment to facilitate task switching are emerging [5]. However, the sonic environment does not appear to have been considered for this purpose. I suggest

manipulation of the sonic environment may facilitate switching between different deep thinking tasks.

2. HYPOTHESIS

The hypothesis of this paper is that *sonic conditioning* through the use of *background soundscapes* may facilitate 'getting into the zone' when switching between different categories of deep thinking. I propose a compositional approach, using various forms of active-silence to create distinct sonic environments to associate with distinct mental activities.

The use of sound to condition automatic physiological responses has a long history, going at least back to Pavlov [6] and his dogs—by training dogs to associate sonic cues with subsequent feeding, Pavlov observed the dogs become conditioned to salivate in response to these sonic cues. It seems plausible that an analogous mental conditioning may be achieved through crafting background soundscapes that become associated with a particular mental activity.

There does not appear to have been any research conducted on whether association through exposure to background sound can assist in re-entering a particular mental state.

3. ACTIVE SILENCE

Whilst most psychological studies on environmental noise concentrate on its distracting effect [7],[8],[9], there are a number of studies that find cognitive benefits to certain types of background noise. For example studies have found that background white noise gives cognitive benefits to monkeys [10], geriatrics [11] and ADHD children [12], and promotes sleep in infants [13]. Di et. al [14] find that pink noise and certain FM tones can alleviate annoyance caused by low frequency environmental sound.

A rarely discussed issue with studies on the enhancing or distracting effect of background sound is the ambiguity of the control group. Typically such studies will compare the effect of various types of background sound to 'silence'. However, silence is a problematic notion. Loudness is relative; in the right context a pin drop can be piercing. In the absence of incident vibrational energy the ear will provide its own, and in an anechoic chamber purpose-built to be devoid of sound, the noise of one's heart beating and blood circulating becomes loud [15]. Occupational health & safety regulations typically mandate maximum occupancy times of around an hour in these deafeningly silent chambers. Rather than an absence of sound, silence is better described as a culturally and contextually determined soundscape of familiarity.

I propose the term *active-silence* for a soundscape artificially generated (through technology) in order to create



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

a silent quality. Active-silence is designed to be *not heard*—to only be perceptible when it ceases. Various types of active-silence are becoming increasingly popular. White noise generators for the bedroom can now be found on Amazon and eBay, and there is a proliferation of noise-streaming websites, offering a selection of noise colours (white, brown, pink) along with various other ‘calming’ soundscapes, typically recordings of rain, the ocean, or other natural environments. For example, whilst writing this paper I have been listening to various soundscapes from *soundrown* (<http://soundrown.com>). Curiously, I have found that for me the ‘cafe’ setting seems most conducive to writing. Try it yourself and see what works best for you! It’s interesting since the cafe setting is quite noisy. The sound of the barista banging the group-head to empty the used coffee is ostensibly quite loud and intrusive.

3.1. Sound-masking

Since the 1970s there has been a commercial industry of active-silence generation for workplaces, often called sound-masking. The idea of sound-masking is to raise the noise floor with a relatively spectro-temporally homogeneous sound, such as white noise, so that other unwanted sounds are less noticeable. White noise *per se* is not typical for commercial office installations. For example, the following is taken from marketing material for a commercial sound-masking system manufacturer:

“Q: Isn’t a sound masking system the same as white noise?”

A: The term white noise describes a specific type of used in early masking systems in the 1970s. These systems were inflexible and the hissing quality of their sound prevented widespread acceptance ... The Logison Acoustic Network makes an engineered sound comparable to that of a soft airflow.” [16]

Pink noise, by which is meant noise with a spectrum whose power is inversely related to frequency, is more commonly used in practice, whilst Logison (for example) markets its product as ‘green noise’, presumably alluding to its naturalesque qualities – though they do not give any details on the synthesis process. Pink noise, also called $1/f$ noise, has been written about at length as a sort of spookily ubiquitous phenomenon, akin to the fibonacci series or the golden ratio, that pops up repeatedly in nature [17] and in music [18]. Treasure [19] similarly recommends naturalesque sounds for sound-masking, particularly “water, wind or waves”.

3.2. Natural Soundscapes

A number of studies have found evidence of positive and restorative effects of natural soundscapes [20], [21], [22], [23], [24]. Moreover, natural soundscapes that induce feelings of tranquility and pleasantness exhibit dramatic heterogeneity in terms of quantitative acoustical descriptors [25], [20], [26], and even in terms of other subjective descriptors [27].

However, playing back recordings of natural soundscapes in an office environment is problematic. It seems to be important for soundscapes to not be overtly incongruous with the visual surrounds [23], [26], [28]. Many natural soundscapes contain particular sounds that are readily recognisable, such as bird-calls, that seem odd in an interior setting.

Yet the abstract sonic qualities of natural soundscapes suggest a point of departure for the creation of active-silences, by mimicing these qualities without the particular and recognisable sounds of a wilderness field recordings: properties such as spatial engulfment [29], acoustic richness [30], acoustic complexity [25] and acoustic diversity [31].

For the purposes of this paper, the great diversity of natural soundscapes that promote tranquility adds confidence to the notion that a variety of active-silences may be created, different enough to condition distinct tasks with distinct soundscapes, whilst remaining undistracting.

3.3. Background Music

The notion that intentional background sound can enhance workplace productivity is not new: the Muzak corporation claimed exactly this from inception in the 1930s all the way through to bankruptcy in 2009, though the scientific rigour of their self-funded studies is disputed [32]. What is not disputed is Muzak’s lack of artistic merit: “people began to use the company’s name as a generic term for anything bland, soulless, and uninspired — so much that today many don’t realize that the word has a non-perjorative application” [33].

The Muzak corporation’s own research suggested that *continuous* Muzak was counterproductive (and unpalatable – workplaces would simply switch it off after a while) and instead opted for alternating periods of 15 minutes of Muzak followed by 15 minutes of “silence” [33].

The intentions of this paper are quite different to the goals of Muzak. Muzak was conceived in a modernist frame at the height of the Scientific Managerialism of Frederik Taylor [34], where workers were considered as “cog[s] in the machinery” [36] inspired by the production lines of Ford Motors. Like much of the modernist program of control-over-nature a side-effect was greater homogenisation of our lived experience. Modernist architecture similarly sought to erode acoustic idiosyncrasy: “reverberation equations, sound meters, microphones, and acoustical tiles were deployed in ... office skyscrapers. The control provided by these technologies, however, was applied in ways that denied the particularity of place, and the diverse spaces of modern America began to sound alike as a universal new sound predominated.” [37].

Critical responses to this modernist acoustic homogenisation included John Cage’s never realised Muzak-Plus [33], and Brian Eno’s Ambient Music: “whereas the extant canned music companies proceed from the basis of regularizing environments and blanketing their acoustic and atmospheric ideosyncracies, Ambient Music is intended to enhance these” [37]. Yet, despite the proliferation of ambient music as a genre tag, “Eno’s mini-manifesto about ambient music became less interesting to him with time, and [ambient music] became something of a moving target, meant to accomplish very different goals, some of which surely had much to do with marketing.” [38].

Other approaches to understanding everyday environmental sound emerged in the 70s with Murray Schafer, Barry Truax and others initiating the World Soundscape Project and the field of Acoustic Ecology [39]. Schafer argued that the invention of the window marked a significant turning point in human phenomenal experience where the soundscape and the visual scene became decoupled. In parallel, a number of theorists across sound studies, urban planning, and cultural theory discussed the contradictions

between treatment of sound in planning & law compared with sonic experience[40].

Yet despite these diverse strands of criticism and resistance to modernist soundscape approaches, practical application of these ideas indoors remain mostly confined to concerts or installations. The post-modern vision of interior acoustic environmental design is yet to be realised.

More recently, since Rauscher et al. [41] promoted the Mozart Effect—claiming that playing Mozart in the background during study enhances learning outcomes—there has been a hive of research activity investigating the influence of background music & sound on concentration, productivity, and learning. This area has been the subject of quite some controversy, of which more below, but it does focus on techniques for enhancing concentration over our timescales of interest, i.e. periods of several hours. None of these studies on background music & sound appear to explicitly consider the effect on task switching, though one might hypothesise that a concentration-enhancing technique is likely to decrease switching time. On the other hand it may be that deeper concentration comes at the cost of cognitive flexibility.

Controversy over the Mozart Effect has centred on the authors' claims that something about the genius of Mozart enhances cognitive ability whilst listening. Thompson et al. [42] argue that there is nothing inherently superior about Mozart's music in producing this effect; rather it is simply the case that people learn better when listening to music that they *liked* and are *familiar with*—the original researchers just happened to choose subjects who knew and liked Mozart.

4. GETTING INTO THE ZONE

A key consideration for productivity in an office environment is the time it takes to 'get in the zone', particularly given the frequency with which interruptions can be typically expected in many workplaces. Whilst there is a wealth of psychological literature relating to attention, concentration and distraction in multi-tasking environments, much of this research considers either very short timescales, or is primarily concerned with fatigue. For assessing the kinds of mental states of interest here, which may take around an hour to achieve, and last for several hours, I suggest *flow* theory.

4.1. Flow

The notion of the psychological state of *flow* was developed by Csikszentmihalyi (1990 / 1975) in his studies of peak experience. A flow state is one of effortless concentration, accompanied by high intrinsic motivation and positive emotion, sometimes described as "the zone" (Lafont 2015). The theory's key finding is that a *balance between challenge and skill* is conducive to entering flow. Though originally developed to describe the sorts of peak experiences often reported by athletes, musicians, and other high achieving individuals, "it is important to note that flow is ... not usually regarded as an all-or-nothing peak experience; rather, degree of flow is a continuous variable that can be used to characterize the experiential quality of any everyday activity" (Csikszentmihalyi & Csikszentmihalyi, 1992), and has been applied to workplace experience in particular (Eisenberger et al. 2005).

4.2. Mental Context Switching

A number of studies, mostly in cognitive psychology, have examined switching time mitigation techniques in various contexts, for example driving whilst talking on the phone. The tasks typically studied are moment-to-moment, rather than requiring cohesive cognition over larger timescales, and the switching times are very short. Wickens et al. comment:

"Numerous models of sequential operations in multi-task performance can be found, and these can be positioned along a time-scale continuum ... the majority of such models appear to lie toward the 'micro' end of the continuum, modeling task switching times in the order of milliseconds ... Often, their focus is exclusively on time, and on accounting for variance in multi-task performance time required to carry out relatively simple cognitive activities" [43]

Studies that investigate longer timescales—on the order of hours, and complex cognitive tasks such as involved in research and writing, are few.

Environmental cues can stimulate recall of mental states formed in the environment from which the cues are taken. Smith & Vela suggest "the reinstatement of context cues ... should benefit memory for information learned in the reinstated environment" [44] although the kinds of experimental manipulations they surveyed typically involved moving subjects between physically different environments (i.e. different rooms). The challenge this position paper explores is how to artificially modify an environment that is physically fairly static – sitting in the same room on the same chair looking at the same computer screen – to allow these artificially created cues to assist in switching mental states.

5. PROPOSED APPROACH

This position argued in this paper is that creating background soundscapes for everyday environments should be approached as *applied composition/sound-design* — an exploration, parameterisation, and evaluation of *active-silences*. The aim of these active-silence compositions would be to assist in 'getting into the zone' in two respects. First, the soundscapes should encourage flow. Second, the soundscapes should have some unique character to enable them to condition switching into a particular mental activity.

I expect the type of soundscapes that encourage flow will vary across people, locations and tasks. To provide depth and breadth, and enhance generalisability whilst retaining ecological validity, a multi-method investigation would be beneficial, including:

- Reflective practice by artist-researchers: longitudinal and iterative practice of composing active-silence, and evaluation of the success of sonic conditioning on the researcher's professional work — particularly the challenge of switching between research, writing, and creative development.
- User testing: development of software that can generate a variety of types of active-silence, allowing the user to select the appropriate silence for the task at hand — and evaluation of the success of the sonic conditioning for 'getting into the zone'.

One approach to evaluating the success of sonic conditioning could be comparison of the efficacy of different soundscapes in promoting *flow*.

5.1. Measuring Flow

There are several reasons for using flow theory in this context, rather than other measures such as concentration or annoyance that are frequently used in studies of noise abatement (e.g. [9], [5]):

(i) Flow theory studies psychology in context. Nakumara & Csikszentmihalyi emphasise that “a key characteristic [of] the flow model ... is *interactionism* ... Rather than focusing on the person, abstracted from context, flow research has emphasised the dynamic system composed of the person and the environment” [45]. Given the primary goal of this proposal is to help the mind perceive a new context despite substantial environmental similarity, through specific environmental modification, this seems apposite.

(ii) Flow theory has mostly been developed to account for cognitive experiences on the timescales of interest here—on the order of hours—whereas measures such as concentration and annoyance are shorter timescale, at least in the ways they have typically been operationalised in these studies.

I propose a novel evaluation tool for measuring flow, tentatively termed ‘Enhanced Experience Sampling Method (EESM)’. The EESM would combine several existing research instruments and technologies, leveraging the recent development of portable ubiquitous computing and biomonitoring devices. In particular it would combine the traditional *Experience Sampling Method* for measuring flow with *physiological markers* and *affective computing* techniques. These components and their possible synthesis in the EESM are elaborated in the next few sections.

5.2. Experience Sampling Method

A standard approach to measuring flow is the *experience sampling method* [46] which is a regimented form of self-reporting on subjective qualitative experience. Typically the sampling is prompted by a some form of signalling device (originally Csikszentmihalyi used electronic paging devices). These authors argue that experience sampling offers the “unique advantage of ... its ability to capture daily life as it is directly perceived from one moment to the next, affording an opportunity to examine fluctuations in the stream of consciousness and the links between the external context and the contents of the mind” [46].

5.3. Physiological Markers

Experience sampling has some methodological disadvantages: it relies on the subjective opinions of the participants, and can be obtrusive—frequent interruptions may hinder flow. To address this various authors have sought to probe flow through observable physiological measures. De Manzano et al. found “a significant relation ... between flow and heart period, blood pressure, heart rate variability, activity of the zygomaticus major muscle, and respiratory depth” [47] and Keller et al. [48] find a challenge-skill-balance measure to be a common cause for both self-reported flow measurements and the psychophysiological markers *heart-rate variability* and *salivary cortisol*.

5.4. Affective Computing

Affective Computing is another body of research that seeks to link observable physical measures with underlying mental states. Having arisen from the field of Human-Computer

Interaction, it concentrates on the computational detection of human emotional states whilst interacting with a computer (originally in the context of artificial intelligence, but now more broadly applied). The primary methods used are rooted in computer vision, for example automated facial expression recognition [49]. As the kinds of tasks this paper is discussing are primarily computer-based, affective computing techniques (using for example the computer’s webcam) could provide another stream of data for interrogating flow.

5.5. Enhanced Experience Sampling Method (EESM)

To evaluate flow in the workplace, an enhanced experience sampling method combining signal-prompted subjective qualitative self-reporting with objective quantitative psychophysiological measures including heart-rate variability and galvanic skin response, and affective computing measures, could be developed. An important design element for implementing an experience sampling instrument is the sampling-schedule, which may be regular, random, or event-based according to the needs of the project [46]. The availability of portable computing devices now means that it is feasible to construct an adaptive sampling-schedule, based on real-time feedback from physiological measures. In the early stages of exploring the efficacy of various background soundscapes in promoting flow, adaptive scheduling could be used to increase the power of tests relating flow states to physiological measures. Later, as these relationships become better established, adaptive scheduling could be used to strategically avoid interrupting moments of flow.

6. CONCLUSION

There is much to suggest that sonic conditioning of mental contexts is possible. It is established that both music and sound can condition emotion [50], enhance concentration [10], [12], [51], and stimulate recall of memories [52]. What has been less well understood is the broad variety of background sounds that can be used without causing distraction. It seems a reasonable extrapolation that training the mind to associate particular soundscapes with particular tasks may help switching between tasks by switching to the associated soundscape.

Other environmental variables are thought to assist in conditioning mental contexts. For example, amongst a growing population of telecommuters and work-from-home employees, freelancers or consultants, common strategies to address “work/home boundary permeability” [53] include deliberate environmental modifications such as setting up a physically separated ‘home-office’ and changing into work-clothes. Furthermore, the ability to control one’s sonic environment is often cited as a benefit of working from home [54]. It may be that the relatively sparse attention paid to workplace sound design reflects a general trend for sound to be of secondary concern in architecture [55].

On a task-based level, evidence from the Human-Computer Interaction field supports the notion of environmental conditioning (or priming) in facilitating mental contexts. For example, Altman & Trafton’s *Memory for Goals* theory [56] examines task resumption after an interruption in the workplace. Andrews et al. explain:

“over the course of an interruption, the activation level of the suspended primary task goal will decay and it will be more difficult to retrieve this goal upon

resumption of the primary task. The activation level of the goal is dependent on two constraints: strengthening and priming ... the priming constraint suggests that cues in the physical or mental environment influence activation by providing associative activation. Critically, an association between the cue and the suspended primary task goal must be established prior to the interruption.” [57]

Kersten & Murphy [58] provide empirical evidence that visual context priming can reduce task switching time, and boost productivity in the context of computer programming, and a growing number of software programs are available that attempt to organise computer monitor display according to task contexts, such as TaskTracer [59], User Monitoring Environment for Activities [59], and TaskTop [61]. None of these programs, however, consider the sonic environment.

The benefits of faster switching times, greater levels of workplace flow, and enhanced multitasking capacity, would be widespread. For most people who work in an office environment, multitasking has become the norm [62]. The pace of multitasking reduces productivity [63] and increases stress [64]. The issue is particularly problematic in academia [65].

This proposal is timely in that recent developments in personal physiological tracking and affective computing have made devices for these functions readily available to consumers, driven by a growing movement of personal bio-monitoring, sometimes referred to as “The Quantified Self” [66]. These devices, such as the Samsung Simband and Fitbit surge smart-watches, and the Intel RealSense camera, provide an unprecedented level of detail whilst remaining fairly unobtrusive.

7. REFERENCES

- [1] Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper & Row.
- [2] Soveri A et al. (2013). Set Shifting Training with Categorization Tasks. Edited by Bart Rypma. *PLoS ONE* 8 (12): e81693.
- [3] Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4), 763-797.
- [4] Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, 8(2), 203-220.
- [5] Kersten, Mik, and Gail C. Murphy. 2015. “Reducing Friction for Knowledge Workers with Task Context.” *AI Magazine* 36 (2).
- [6] Pavlov, I. P. (1902). *The work of the digestive glands*. London: Griffin.
- [7] Edwards, B., Hafter, E., Kalluri, S., & Sarampalis, A. (2009). Objective measures of listening effort: effects of background noise and noise reduction. *Journal of Speech, Language and Hearing Research*, 52(5), 1230+.
- [8] Landström, U., Löfstedt, P., Akerlund, E., Kjellberg, A., & Wide, P. (1990). Noise and annoyance in working environments. *Environment International*, 16(4), 555-559.
- [9] Errett, J., Bowden, E. E., Choiniere, M., & Wang, L. M. (2006). Effects of noise on productivity: does performance decrease over time. *Architectural Engineering Institute*. 190(18).
- [10] Carlson et al. (1997). “Effects of Music and White Noise on Working Memory Performance in Monkeys.” *NeuroReport* 8: 2853-56.
- [11] Burgio, L et al. (1996). Environmental "white noise": An intervention for verbally agitated nursing home residents: *Journals of Gerontology: Series B: Psychological Sciences and Social Sciences* Vol 51B(6) Nov 1996, P364-P373.
- [12] Söderlund, Göran BW, Sverker Sikström, Jan M. Loftesnes, and Edmund J. Sonuga-Barke. 2010. “The Effects of Background White Noise on Memory Performance in Inattentive School Children.” *Behavioral and Brain Functions* 6 (1): 55.
- [13] Forquer, LeAnne M., and C. Merle Johnson. 2005. “Continuous White Noise to Reduce Resistance Going to Sleep and Night Wakings in Toddlers.” *Child & Family Behavior Therapy* 27 (2).
- [14] Di, G., Li, Z., Zhang, B., & Shi, Y. (2011). Adjustment on subjective annoyance of low frequency noise by adding additional sound. *Journal of Sound and Vibration*, 330(23), 5707-5715.
- [15] Cage, John. (1973 / 1937). *Silence: Lectures and Writings* by John Cage. Middletown, Connecticut: Wesleyann University Press.
- [16] Logison (a). Understanding Sound Masking. Retrieved from http://www.logison.com/site_Files/Content/PDF/Logison_Understanding_Sound_Masking.pdf
- [17] Handel PH & Chung AL (1993) *Noise in Physical Systems and 1/f Fluctuations*. New York: American Institute of Physics.
- [18] Voss, R. F., & Clarke, J. (1978). "1/f noise" in music: Music from 1/f noise. *Journal of the Acoustical Society of America* 63: 258-263.
- [19] Treasure, J. (2011). *Sound business*. Management Books 2000 Limited.
- [20] De Coensel, B., & Botteldooren, D. (2006). The quiet rural soundscape and how to characterize it. *Acta Acustica United with Acustica*, 92(6), 887-897.
- [21] Lam, K.-C., Brown, A., Marafa, L., & Chau, K.-C. (2010). Human Preference for Countryside Soundscapes. *Acta Acustica United with Acustica*, 96, 463-471.
- [22] Yang, W., & Kang, J. (2005). Acoustic comfort evaluation in urban open public spaces. *Applied Acoustics*, 66(2), 211-229.
- [23] Brown, A. L. (2012). A review of progress in soundscapes and an approach to soundscape planning. *Int. J. Acoust. Vib*, 17(2), 73-81.
- [24] Davies, W. J., Adams, M. D., Bruce, N. S., Cain, R., Carlyle, A., Cusack, P., ... Poxon, J. (2013). Perception of soundscapes: An interdisciplinary approach. *Applied Acoustics*, 74(2), 224-231.
- [25] Farina, F., Bogaert, J., Schipani, J., 2005. Cognitive landscape and information: new perspectives to investigate the ecological complexity. *BioSystems* 79, 235-240.
- [26] Pheasant, R., Horoshenkov, K., Watts, G., & Barrett, B. (2008). The acoustic and visual factors influencing the construction of tranquil space in urban and rural environments tranquil spaces-quiet places? *The Journal of the Acoustical Society of America*, 123(3), 1446.
- [27] Axelsson, Ö., Nilsson, M. E., & Berglund, B. (2010). A principal components model of soundscape perception. *The Journal of the Acoustical Society of America*, 128(5), 2836.

- [28] Cerwén, G. (2016). Urban soundscapes: a quasi-experiment in landscape architecture. *Landscape Research*, 1–14.
- [29] Paine, G., Szadov, R., & Stevens, K. (2007). Perceptual Investigation into Envelopment, Spatial Clarity, and Engulfment in Reproduced Multi-Channel Audio. In *Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio*.
- [30] Towsey, M., Wimmer, J., Williamson, I., & Roe, P. (2014). The use of acoustic indices to determine avian species richness in audio-recordings of the environment. *Ecological Informatics*, 21, 110-119.
- [31] Sueur, J., Gasc, A., Grandcolas, P., Pavoine, S., 2012. Global estimation of animal diversity using automatic acoustic sensors. In: *Sensors for Ecology: Towards Integrated Knowledge of Ecosystems*. CNRS Editions, pp. 101–119.
- [32] Vanel, H. (2013). *Triple Entendre: Furniture Music, Muzak, Muzak-Plus*. University of Illinois Press.
- [33] Owen, D. (2006, April 10). The Soundtrack of your Life: Muzak in the Realm of Retail Theatre. *The New Yorker*, 66–71, p. 69.
- [34] Jones, S., & Shumacher, T. (1992). Muzak: on functional music and power. *Critical Studies in Mass Communication*, 9(2), 156–169.
- [35] Rosen, E. (1993). *Improving Public Sector Productivity: Concepts and Practice*. Thousand Oaks, CA: Sage Publications.
- [36] Thompson, E. (2004). *The Soundscape of Modernity: architectural acoustics and the culture of listening in America*. Cambridge, MA: MIT Press.
- [37] Eno, B. (1978/2004). Ambient music. *Audio Culture. Readings in Modern Music*, 94–97.
- [38] Richardson (2002). As Ignorable as it is Interesting: the Ambient Music of Brian Eno. Pitchfork.
- [39] Murray Schafer, Raymond. "The soundscape: Our sonic environment and the tuning of the world." *Vancouver: Destiny Books* (1977).
- [40] LaBelle, B. (2010). *Acoustic territories: sound culture and everyday life*. A&C Black.
- [41] Rauscher, Frances H., Gordon L. Shaw, and Katherine N. Ky. "Listening to Mozart enhances spatial-temporal reasoning: towards a neurophysiological basis." *Neuroscience letters* 185, no. 1 (1995): 44-47.
- [42] Thompson W et al. (2001). Arousal, Mood, and The Mozart Effect. *Psychological Science* 12 (3): 248–51.
- [43] Wickens, Christopher D., Robert S. Gutzwiller, and Amy Santamaria. 2015. "Discrete Task Switching in Overload: A Meta-Analyses and a Model." *International Journal of Human-Computer Studies* 79 (July): 79–84.
- [44] Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, 8(2), 203–220.
- [45] Nakamura, J., & Csikszentmihalyi, M. (2002). The concept of flow. *Handbook of positive psychology*, 89-105.
- [46] Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Sage.
- [47] de Manzano, Ö., Theorell, T., Harmat, L., & Ullén, F. (2010). The psychophysiology of flow during piano playing. *Emotion*, 10(3), 301–311.
- [48] Keller, J., Bless, H., Blomann, F., & Kleinböhl, D. (2011). Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology*, 47(4), 849-852.
- [49] Zeng, Z., Pantic, M., Roisman, G., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1), 39-58.
- [50] Eifert, G. H., Craill, L., Carey, E., & O'Connor, C. (1988). Affect modification through evaluative conditioning with music. *Behaviour Research and Therapy*, 26(4), 321-330.
- [51] Haake AB (2011). Individual Music Listening in Workplace Settings: An Exploratory Survey of Offices in the UK. *Musicae Scientiae* 15(1).
- [52] Sacks, Oliver. 2007. *Musophilia: Tales of Music and the Brain*. Picador.
- [53] Gajendran, Ravi S., and David A. Harrison. 2007. "The Good, the Bad, and the Unknown about Telecommuting: s." *Journal of Applied Psychology* 92 (6): 1524–41.
- [54] Pearlson, K. E., & Saunders, C. S. (2001). There's no place like home: Managing telecommuting paradoxes. *The Academy of Management Executive*, 15(2), 117–128.
- [55] Blesser, B & Salter L. (2009). *Spaces Speak, Are You Listening?: Experiencing Aural Architecture*. MIT press.
- [56] Altmann E & Trafton JG (2002). Memory for goals: An activation-based model. *Cognitive Science*. 26(1), 39-83.
- [57] Andrews A et al. (2009). "Recovering From Interruptions: Does Alert Type Matter?" *Human Factors and Ergonomics* 53 (4): 409–13.
- [58] Kersten, M., & Murphy, G. C. (2012). Task context for knowledge workers. In *Proc. AAAI 2012 Activity Context Representation workshop*.
- [59] Dragunov, et al. 2005. "TaskTracer: A Desktop Environment to Support Multi-Tasking Knowledge Workers." In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, 75–82. ACM.
- [60] Kaptelinin, V. 2003. UMEA: Translating Interaction Histories into Project Contexts. In *Proceedings of the 2003 ACM SIGCHI Conference on Human Factors in Computing Systems*, 353–360. New York: Association for Computing Machinery.
- [61] Kersten, M., & Murphy, G. C. (2015). Reducing Friction for Knowledge Workers with Task Context. *AI Magazine*, 36(2).
- [62] González, Victor M., and Gloria Mark. 2004. "Constant, Constant, Multi-Tasking Crazy: Managing Multiple Working Spheres." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 113–20. ACM.
- [63] Leshed, Gilly, and Phoebe Sengers. 2011. "I Lie to Myself That I Have Freedom in My Own Schedule: Productivity Tools and Experiences of Busyness." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 905–14. ACM.
- [64] Mark, Gloria, Daniela Gudith, and Ulrich Klocke. 2008. "The Cost of Interrupted Work: More Speed and Stress." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 107–10. ACM.
- [65] Levy, David M. 2007. "No Time to Think: Reflections on Information Technology and Contemplative Scholarship." *Ethics and Information Technology* 9 (4): 237–49.
- [66] *Technology Quarterly*. (2012). "The Quantified Self: Counting Every Moment." *The Economist*.

VIRTUAL-AUDIO AIDED VISUAL SEARCH ON A DESKTOP DISPLAY

Clayton Rothwell

Infoscitex Corporation
4027 Colonel Glenn Hwy, Suite 210
Dayton, OH 45431, USA
crothwell@infoscitex.com

Griffin Romigh and Brian Simpson

Air Force Research Laboratory
2610 Seventh Street, Area B, Bldg 441
Wright-Patterson Air Force Base, USA
griffin.romigh@us.af.mil
brian.simpson.4@us.af.mil

ABSTRACT

As visual display complexity grows, visual cues and alerts may become less salient and therefore less effective. Although the auditory system's resolution is rather coarse relative to the visual system, there is some evidence for virtual spatialized audio to benefit visual search on a small frontal region, such as a desktop monitor. Two experiments examined if search times could be reduced compared to visual-only search through spatial auditory cues rendered using one of two methods: individualized or generic head-related transfer functions. Results showed the cue type interacted with display complexity, with larger reductions compared to visual-only search as set size increased. For larger set sizes, individualized cues were significantly better than generic cues overall. Across all set sizes, individualized cues were better than generic cues for cueing eccentric elevations ($> \pm 8^\circ$). Where performance must be maximized, designers should use individualized virtual audio if at all possible, even in small frontal region within the field of view.

1. INTRODUCTION

The complexity and clutter of visual displays, such as dynamic interactive map displays, has increased over the last decade. Visual alerts on maps, for instance, have to compete with: the map symbols, colors, contrast and motion that are represented in the map. In other words, a designer is challenged to create a visual pop-out effect in an already colorful and moving scene. Increases in visual complexity may reduce the effectiveness of the visual alerts that have previously been effective in simpler, less cluttered maps. Spatialized auditory alerts can point to a location in space, such as the location of a particular visual object on a map display, as an act of deixis [1]. Yet the visual modality has better spatial resolution than the auditory modality, so it has often been the case the visual alerts alone have been used to alert different spatial locations on the monitor. This research investigated if auditory spatial alerts can aid visual search in a small frontal spatial region and what their utility is as a function of the complexity of the visual display (here in terms of the number of visual distractors).

Auditory spatial acuity is relatively worse than visual acuity. Visual vernier acuity averages around 5 arc seconds (or 0.0014° ; [2]). Auditory acuity has been estimated in a variety of ways.

Measurements of minimum audible angle (MAA; [3, 4]) suggest resolution around 1° . Measurements of localization error in the free field find $\sim 11^\circ$ absolute angular error (after removing front-back confusions) and many studies suggest that localization errors increase for virtual audio (See [5] for a discussion). Within virtual audio there are differences in accuracy as well; individualized spatial audio can be near free-field performance whereas generic (i.e., non-individualized) spatial audio is worse [6]. Still, auditory cues have been shown to benefit visual search tasks despite the auditory system's resolution, such as a pilot searching for nearby aircraft traffic on the ground [7] or in the air [8], or in the task paradigm of *aurally aided visual search*, a visual search in 360° space surrounding the searcher (e.g., [9, 10]). In the research of Perrott et al. [9] and Bolia et al. [10], a spherical search space comprised of 277 loudspeakers placed approximately 15° apart surrounded the participant. Each loudspeaker has a cluster of 4 LEDs that can be independently lit. A target was displayed along with varying numbers of distractors (i.e., different set sizes) and the target was present on every trial. The target was one of two possible configurations of LEDs and the participant's task was to find and identify the target configuration. Accuracy and response time were measured as a function of the availability of a cue and/or the type of cue and the set size.

The aurally aided visual search paradigm has been used to show the benefit (i.e., reduction in search times) of an audio cue compared to visual only search. Additionally, this research has been used to discriminate between the effectiveness of different types of auditory cues. For example, researchers have used free-field sounds played from the target location and compared those to non-individualized virtual sound sources for that location (e.g., [10]). They found that both free-field and non-individualized virtual cues provided a benefit, but non-individualized virtual cues did not provide as much of a benefit as free-field cues did. Additional research has manipulated free-field and virtual auditory cues further by changing cue reliability / precision, measured the impact of hearing protection devices on spatial hearing, and investigated potential for multi-sensory cues to facilitate search times [11, 12, 13]. The reduction in search times from spatial audio cues is unsurprising in part because of the discretization of the visual search area and the possibility for visual targets to appear outside of the field of view. For instance, an auditory cue that was localized within 11° of the target would orient a searcher within one or two visual stimuli from the target. Also, an auditory cue to a region outside of the current field of view would naturally improve search times. Neither of these circumstances hold true in a small spatial region represented by a computer monitor. It is unclear if being



This work is licensed under Creative Commons Attribution Non-Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

oriented within 11° of the intended location would reduce search times in a search area that may only subtend $50^\circ \times 30^\circ$, has many visual stimuli within that region, and is completely within the field of view. Yet other research has shown that free-field spatial auditory cues can speed target identification even in the frontal region and even in the absence of distractors (two frontal locations were measured: 0° and 15° ; [14]), suggesting perhaps virtual auditory cues could speed search times.

The experiments presented here tested the utility of auditory spatial cues in a visual search task in a small frontal region. All audio cues were virtual, yet two different cue types were tested: cues created with individualized head-related transfer functions (HRTFs) and cues created with generic HRTFs (measured on a Knowles Electronics Mannequin for Acoustics Research, or KEMAR). This manipulation was to test if the previously found differences in localization accuracy between individualized and generic virtual audio would matter in this small region [6], similar to the effects found in previous work investigating spatial precision of free-field auditory cues on visual search [15]. Moreover, the comparison between generic-HRTFs and individualized-HRTFs was motivated by a practical issue: if auditory cues were shown to reduce search times and generic HRTFs were no different from individualized cues, then displays with spatialized auditory alerts could be deployed with one set of generic HRTFs rather than needing to measure HRTFs for every user and switch the HRTFs being used. The effectiveness of auditory cues was measured for many different levels of visual complexity, i.e., the number of distractors in the visual scene. Two experiments investigated different ranges of set size.

2. EXPERIMENT 1

The first experiment measured search times when there was no audio cue (visual only), when there were virtual audio cues rendered with individualized HRTFs, and when there were virtual audio cues rendered with generic HRTFs (KEMAR). Also, visual scene complexity was varied by manipulating set sizes, defined as the number of visual stimuli on the screen (including the target). The set sizes tested in Experiment 1 were: 1 (target only), 6, 12, and 24.

2.1. Method

2.1.1. Participants

Nine participants (4 female) with audiometrically-normal hearing and normal or corrected-to-normal vision were paid for their participation. All participants had previous experience with psychoacoustic tasks, including free-field and virtual audio localization experiments. All participants provided informed consent under a protocol approved by the Air Force Research Laboratory, 711th HPW Institutional Review Board.

2.1.2. Stimuli

Visual and auditory stimuli creation and experiment control was done within MATLAB (MathWorks), using the Psychtoolbox [16]. The visual search task was to indicate which one of two possible targets was present. The targets were similar to a Landolt C; they were circular rings with a diameter of 1.24° that had an opening of 0.13° on either the right or left side. The thickness of the circle's line (i.e., the stroke width) was 0.10° . The distractors were

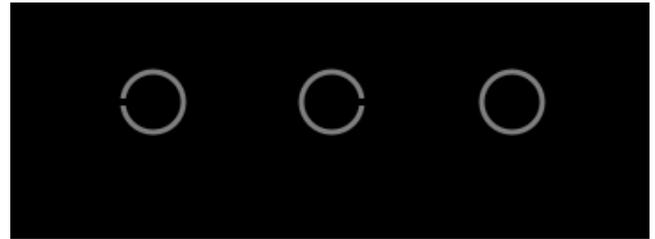


Figure 1: Examples of the visual stimuli that appeared in Experiments 1 and 2, not shown to scale. The leftmost stimulus is a target facing left, the middle stimulus is a target facing right, and the rightmost stimulus is a distractor.

circles of the same diameter and stroke width that had no opening. Examples of both possible targets and the distractor are shown in Figure 1. To maximize the sensitivity to differences in the two spatial auditory cues, the target opening was made small such that the target could not be identified with peripheral vision, but had to be foveated [17]. Visual stimuli were presented on a monitor that subtended $\pm 27^\circ$ azimuth and $\pm 16^\circ$ elevation. Visual stimuli were presented against a black background and contrast of the visual stimuli was the same for the target and distractors. Pilot studies using a higher contrast value had pronounced perceptual tracers that were distracting to searchers. For each trial, the target was randomly placed and distractors, when present, were randomly placed such that they never overlapped the target or other distractors. When stimuli were immediately adjacent to each other, there was 0.7° between them.

The auditory stimuli were 250-ms bursts of broadband noise (0.2-14.5 kHz) with a pink spectrum. Stimuli had 5-ms cosine ramps and were played at approximately 65 dB. These stimulus parameters were used in prior experiments on aurally aided visual search [9, 10] and were used for comparison. For each individual listener and a KEMAR acoustic mannequin, an HRTF was measured prior to the study according to the methods described in [18]. In short, subjects were outfitted with binaural microphones that blocked off, and sat flush with, the entrance of the ear canal while broadband signals (periodic chirps) were presented from 277 loudspeaker locations surrounding the listener and recorded binaurally. A similar process was used for the KEMAR mannequin, but utilized the built-in ear-canal microphones (GRAS 46AO). The resulting recordings were subsequently used to calculate a sample HRTF for each location in the form of 256 Discrete Fourier Transform magnitude coefficients for each ear and a corresponding ITD. ITDs were found by taking the difference in slope of the best-fit lines to the unwrapped low-frequency (300-1500 Hz) phase response of each ear. Headphone (Sennhiser HD-280) correction filters were also collected for each subject (and KEMAR) using a similar measurement technique (described in [18]). Final spatial filters were created for each measurement location by constructing a time domain filter using the headphone-corrected HRTF magnitude and a minimum phase assumption. ITDs were incorporated into the minimum phase filters by delaying the contralateral ear by the corresponding delay.

2.1.3. Procedure

Participants sat in a double-wall sound-isolated booth and used a chin rest. Their eyes were approximately 54 cm away from the

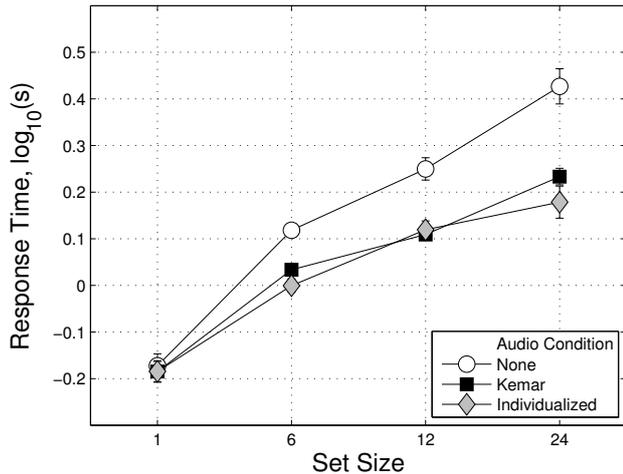


Figure 2: Log response times for Experiment 1, shown as a function of set size and audio cue type. Within-subject standard error bars are shown after Morey [19].

monitor and their eyes were approximately centered on the screen. Each trial began with a fixation point at the center of the screen that was present for 500-1000 ms then disappeared and the search display immediately appeared. Participants searched until they found the target and they used the keyboard arrow keys to indicate their response of left/right. Participants were instructed to find the target as fast as possible while maintaining accuracy. Block length was 50 trials, which varied in duration due to the variation in response times for different conditions. There were 2 blocks of each combination of cue type and set size (i.e., 24 blocks total) and the order of blocks was randomly determined.

2.2. Results

A repeated-measures analysis of variance (ANOVA) was conducted on log response times. Accuracy of target identification responses was quite high (98.9%) so all trials were included in the response time analysis. To assist in plotting data that are averaged across subjects, within-subject standard error bars are calculated after Morey [19] to visually represent the within-subject error term used in statistical tests.

There was a significant main effect of cue type ($F(2, 16) = 47.14, p < .001$). Both cue types led to shorter response times compared to the visual only condition, and the HRTF cues were not different from each other. Consistent with past research, there was a significant effect of set size ($F(3, 24) = 184.70, p < .001$), larger set sizes led to longer response times. In addition, there was a significant interaction between cue type and set size (Figure 2, $F(6, 48) = 18.77, p < .001$). As set size increases, auditory cues provide a larger reduction in search times.

2.3. Discussion

We found that both the individualized-HRTF cues and the generic-HRTF cues reduced search times in comparison to the visual only condition and that the auditory cues provided a larger reduction as set size increased. The individualized-HRTF and

generic-HRTF cues were not different from each other, suggesting that the increased localization accuracy of individualized HRTFs does not affect this task for simple displays with few visual objects. However, it is possible that there would be differences between the two different auditory cues for set sizes larger than those tested here.

A common finding in the literature on virtual audio is that elevation error is larger than azimuth error, particularly with generic HRTFs [20]. This likely is due to the nature of the spectral monaural cues used for elevation, which vary more between individuals than the interaural time and level cues used for azimuth. Therefore, we tried to examine if there was a difference between individualized and generic HRTFs when the azimuth and elevation of the target was considered. We separated the trials where the target appeared at an eccentric azimuth or elevation and compared that to trials where targets appeared at a central azimuth or elevation. The screen subtended $\pm 27^\circ$ in azimuth and $\pm 16^\circ$ in elevation. Therefore, eccentric azimuth was defined as targets that appeared at an absolute azimuth greater than 13.5° and eccentric elevation was defined as targets that appeared at an absolute elevation greater than 8° . The left panel of Figure 3 shows eccentric azimuths for the three cue types. Eccentric azimuths were not slower for generic or individualized cues, though they were slower for visual only conditions ($p < .01$). Eccentric elevations, shown in the right panel of Figure 2, were slower compared to the central elevations ($p < .001$). In addition, there was an interaction between cue type and eccentricity ($p < .05$). The KEMAR cue was not as fast as the individualized cue for eccentric elevations, though they are not different for the central elevations

3. EXPERIMENT 2

Experiment 2 investigated a larger range of set sizes, while using the same cue types as Experiment 1. We hypothesized that larger set sizes would introduce an overall difference in response times between the individualized-HRTF cues and the generic-HRTF cues in addition to the differences for eccentric elevations found in Experiment 1.

3.1. Method

The same participants from Experiment 1 were used for Experiment 2. The same stimuli and equipment from Experiment 1 were used for Experiment 2, with the exception that the set sizes tested were: 24, 48, 96, and 1092 (filled screen). When the screen was filled, the stimuli comprised a grid. No stimuli overlapped and there was 0.7° between them.

3.2. Results

The same data analysis was conducted in Experiment 2 as had been conducted for Experiment 1 on the log response times for target identification. Again, target identification accuracy was quite high (98.8%), so all trials were included in the response time analysis. There was a significant main effect of cue type (Figure 4, $F(2, 16) = 67.87, p < .001$). Performance in the visual-only condition was slower than when KEMAR-HRTF cues were present, and search times with KEMAR-HRTF cues were slower than search times with individualized-HRTF cues. There was a significant main effect of set size ($F(3, 24) = 53.90, p < .001$). Response times increased as set size increased. There was a significant interaction

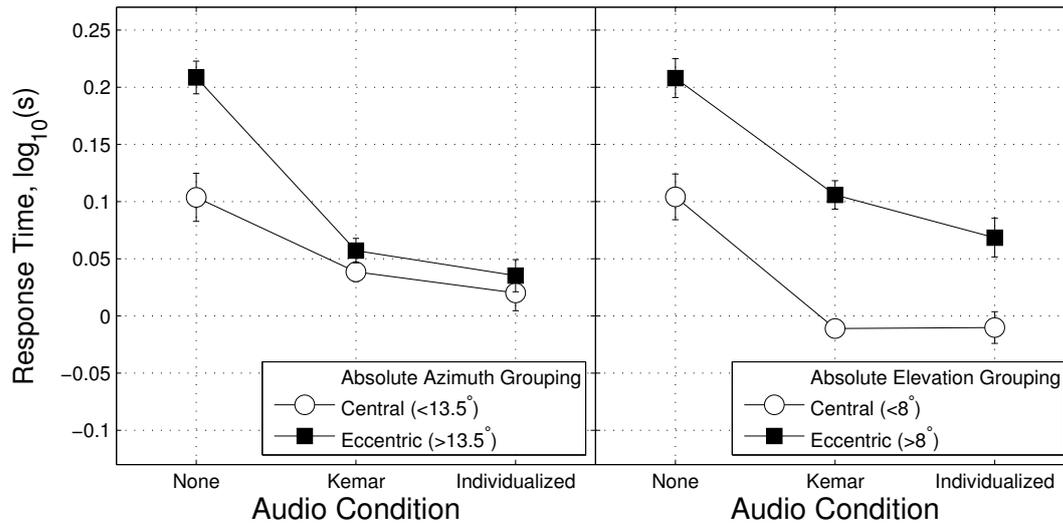


Figure 3: Log response times as a function of eccentricity in azimuth (left panel) and elevation (right panel) for Experiment 1. Within-subject standard error bars are shown after Morey [19].

between cue type and set size (Figure 5, $F(6, 48) = 3.0, p < .05$). As set size increased, the reduction in response times provided by individualized-HRTF cues in comparison to visual only search increased ($p < .001$). There was no interaction between KEMAR-HRTF cues and individualized-HRTF cues as a function of set size or between KEMAR-HRTF cues and visual-only search as a function of set size (both $p > .30$).

The trials were separated into eccentric azimuths or elevations and central azimuths or elevations using the same criteria as in Experiment 1 (Figure 6). For azimuth (left panel), target identification times were faster for individualized-HRTF cues than KEMAR-HRTF cues for both central and eccentric target azimuths. There was a significant interaction ($p < .01$), which was due to the slower responses for visual-only eccentric locations compared to the visual-only central locations. For elevation (right panel), there was no interaction between eccentricity and cue type ($p = .07$). Because of the particular hypothesis that eccentric locations might reveal differences between individualized-HRTF cues and KEMAR-HRTF cues, another ANOVA was conducted without the visual-only data. This test was significant ($p < .01$). For central elevations, KEMAR cues and individualized cues were not different but for eccentric elevations, KEMAR cues were slower than individualized cues.

3.3. Discussion

Experiment 2 showed that auditory cues led to faster response times compared to visual only search. In addition, it showed that individualized-HRTF cues were faster than KEMAR-HRTF cues. Further analyses showed that this enhancement was found in eccentric elevations, and surprisingly found also in azimuth (central and eccentric). The difference in azimuth may be attributed to the large main effect of cue type that appears in elevation.

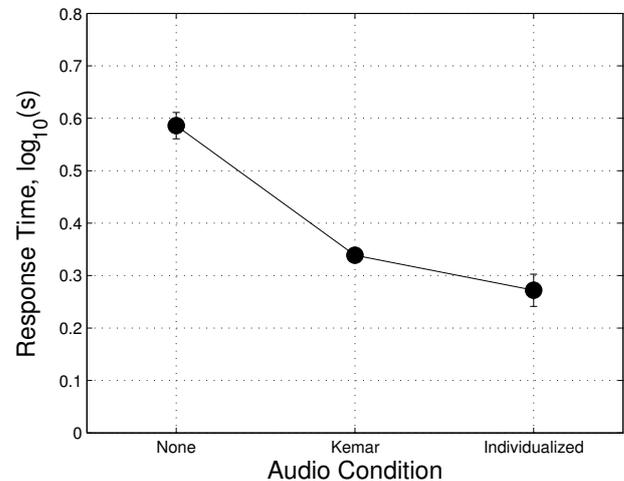


Figure 4: Log response times for Experiment 2, shown as a function of audio cue type. Within-subject standard error bars are shown after Morey [19].

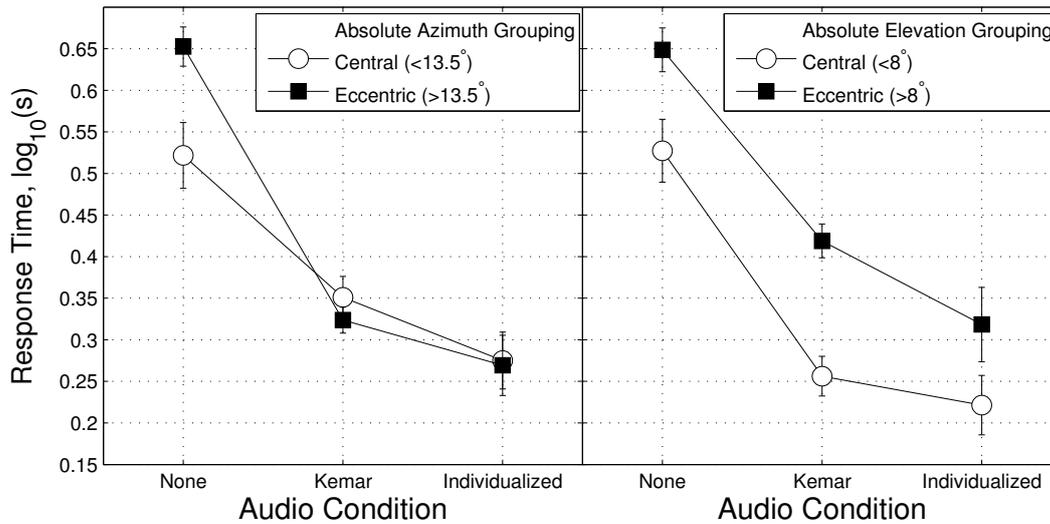


Figure 6: Log response times as a function of eccentricity in azimuth (left panel) and elevation (right panel) for Experiment 2. Within-subject standard error bars are shown after Morey [19].

4. GENERAL DISCUSSION

This study explored whether or not spatialized auditory cues could provide effective spatial cues in a visual search task in a small frontal region. Auditory cues were either absent, created using generic HRTFs or created using individually-measured HRTFs. Experiments 1 and 2 tested performance in two different ranges of visual display complexity. Experiment 1 used set sizes from 1 (target only) to 24, and Experiment 2 used set sizes from 24 to 1092 (filled screen).

In general, the presence of a spatial auditory cue reduced search times and interacted with visual scene complexity. In simple visual displays, there was no overall difference between search times with individualized cues and generic cues, though individualized cues to eccentric elevations were faster than generic cues. In Experiment 1, there was no difference between visual only search times and search times with an audio cue when the set size was 1 (no distractors present). This is in contrast to the research done by Perrot et al [14] that found that free-field audio reduced target identification times even on target-only trials. We did not find this, perhaps because virtual audio was used here or perhaps because the identifying features of their visual targets may have been more easily detected using peripheral vision than our stimuli were. Perrot et al’s target subtended a visual angle of 0.97° and its orientation was the identifying feature whereas we had a small feature by comparison (0.13°).

For complex displays, there were overall differences between the two types of virtual audio cues, with individualized cues providing faster response times than generic cues. The eccentricity effects were found in complex displays as well, suggesting that individualized cues would become more and more important with the use of larger displays. The findings for both Experiments agree

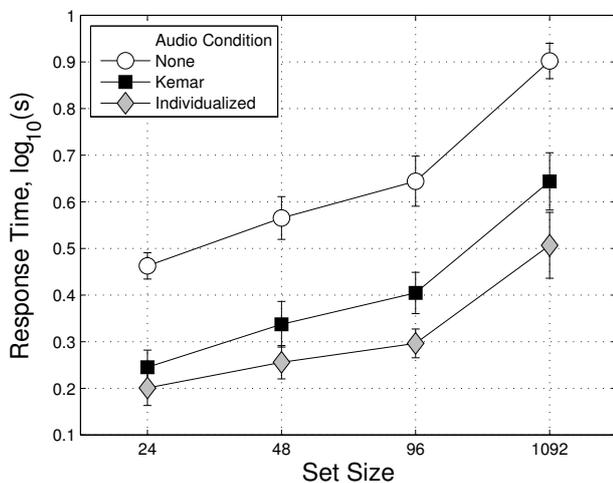


Figure 5: Log response times for Experiment 2, shown as a function of set size and audio cue type. Within-subject standard error bars are shown after Morey [19].

with a study by Vu et al. on free-field audio cues to a visual target that investigated cue displacement [15]. They found that non-displaced cues (cues at the target location) were better than displaced cues (cues off the target location in either horizontal or vertical dimension) which were better than a non-informative cue at reducing search times, and the magnitude of these effects varied with the number of distractors and the amount of displacement. In the present work, the virtual cues created using KEMAR HRTFs may have been displaced or more displaced than the cues created with individualized-HRTFs, and therefore may have been perceived at a spatial location that was not the visual target's location. The displacement may have been mostly in elevation as indicated by search times for eccentric elevations, and Vu et al. found that elevation displacement reduced search times more than horizontal displacement. However, the indication of displacement here is only indirect, no measure of localization was conducted for the virtual stimuli. Future work using virtual audio in visual search should measure localization of the stimuli, and perhaps an *in situ* measurement of localization could be accomplished through eye tracking.

The data from Experiments 1 and 2 suggest that the primary benefit of individualized-HRTF cues is found in the elevation dimension, consistent with previous localization research [20]. In both experiments, individualized-HRTF cues to targets at eccentric elevations resulted in faster searches than generic-HRTF cues. Cues to targets at central elevations resulted in search times that were not different between the two cue types. This finding suggests that other alternative auditory cues that do not indicate elevation, such as stereo panning or interaural level differences alone, would not perform as well as individualized-HRTFs. We did not test these alternative auditory cues but future work could investigate if they may reduce search times compared to visual-only search and if they provide comparable search times to generic-HRTFs or if generic-HRTFs still perform better perhaps due to providing some elevation information.

In conclusion, these data support the notion that spatial audio cues are useful spatial cues to visual displays. Furthermore, individualized spatial audio is functionally superior to generic spatial audio with eccentricity and display complexity.

5. REFERENCES

- [1] J. A. Ballas, "Delivery of information through sound," in *Santa Fe Institute Studies in the Sciences of Complextex- Proceedings Volume*, vol. 18. Addison-Wesley Publishing, 1994, pp. 79–79.
- [2] D. M. Levi, S. A. Klein, and A. Aitsebaomo, "Vernier acuity, crowding and cortical magnification," *Vision Research*, vol. 25, no. 7, pp. 963–977, 1985.
- [3] A. W. Mills, "On the minimum audible angle," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, 1958.
- [4] D. W. Grantham, "Detection and discrimination of simulated motion of auditory targets in the horizontal plane," *The Journal of the Acoustical Society of America*, vol. 79, no. 6, pp. 1939–1949, 1986.
- [5] G. D. Romigh, D. S. Brungart, and B. D. Simpson, "Free-field localization performance with a head-tracked virtual auditory display," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 9, no. 5, pp. 943–954, 2015.
- [6] 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.
- [7] D. R. Begault, "Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 35, no. 4, pp. 707–717, 1993.
- [8] B. D. Simpson, D. S. Brungart, R. H. Gilkey, J. L. Cowgill, R. C. Dallman, R. F. Green, K. L. Youngblood, and T. J. Moore, "3d audio cueing for target identification in a simulated flight task," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 48, no. 16. SAGE Publications, 2004, pp. 1836–1840.
- [9] D. R. Perrott, J. Cisneros, R. L. McKinley, and W. R. D'Angelo, "Aurally aided visual search under virtual and free-field listening conditions," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 38, no. 4, pp. 702–715, 1996.
- [10] R. S. Bolia, W. R. D'Angelo, and R. L. McKinley, "Aurally aided visual search in three-dimensional space," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 41, no. 4, pp. 664–669, 1999.
- [11] J. C. Mateo, B. D. Simpson, R. H. Gilkey, N. Iyer, and D. S. Brungart, "Spatial multisensory cueing to support visual target-acquisition performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, no. 1. SAGE Publications, 2012, pp. 1312–1316.
- [12] B. D. Simpson, R. S. Bolia, R. L. McKinley, and D. S. Brungart, "The impact of hearing protection on sound localization and orienting behavior," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 47, no. 1, pp. 188–198, 2005.
- [13] J. M. Haggit, "Cued visual search and multisensory enhancement," Master's thesis, Wright State University, 2014.
- [14] D. R. Perrott, T. Sadralodabai, K. Saberi, and T. Z. Strybel, "Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 33, no. 4, pp. 389–400, 1991.
- [15] K.-P. L. Vu, T. Z. Strybel, and R. W. Proctor, "Effects of displacement magnitude and direction of auditory cues on auditory spatial facilitation of visual search," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 48, no. 3, pp. 587–599, 2006.
- [16] D. H. Brainard, "The psychophysics toolbox," *Spatial Vision*, vol. 10, pp. 433–436, 1997.
- [17] J. P. McIntire, P. R. Havig, S. N. Watamaniuk, and R. H. Gilkey, "Visual search performance with 3-d auditory cues: Effects of motion, target location, and practice," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2010.
- [18] D. S. Brungart, G. Romigh, and B. D. Simpson, "Rapid collection of hrtfs and comparison to free-field listening," in *International Workshop on the Principles and Applications of Spatial Hearing*, 2009.

- [19] R. D. Morey, "Confidence intervals from normalized data: A correction to Cousineau (2005)," *Tutorials in Quantitative Methods for Psychology*, vol. 4, no. 2, pp. 61–64, 2008.
- [20] G. D. Romigh and B. D. Simpson, "Do you hear where i hear?: isolating the individualized sound localization cues," *Frontiers in Neuroscience*, vol. 8, 2014.

WORD SPOTTING IN A MULTICHANNEL VIRTUAL AUDITORY DISPLAY AT NORMAL AND ACCELERATED RATES OF SPEECH

Derek Brock, Christina Wasylshyn, and Brian McClimens

U.S. Naval Research Laboratory,
4555 Overlook Ave., S.W.
Washington, DC 20375 USA

[\[derek.brock\]](mailto:derek.brock@nrl.navy.mil) | [christina.wasylshyn](mailto:christina.wasylshyn@nrl.navy.mil) | [brian.mcclimens](mailto:brian.mcclimens@nrl.navy.mil) | [@nrl.navy.mil](mailto:nrl.navy.mil)

ABSTRACT

The demands of concurrent radio communications in Navy shipboard command centers contribute to the problem of operator information overload and impede personnel optimization goals for new platforms. Motivations for serializing this task and human performance research with virtual, multichannel, rate-accelerated speech in support of this idea are briefly reviewed, and the results of a recent listening study in which participants carried out a Navy-relevant word-spotting task in this context are reported.

1. INTRODUCTION

The broad operational range of radio circuits that require active attention in Navy command centers is a factor in operator information overload and an impediment to personnel optimization goals for new and existing platforms [1][2]. As part of an effort to address this and other performance issues related to multitasking in shipboard decision environments, our research group is exploring machine-mediated task serialization concepts.

Under one of our proposals, concurrent voice communications would be buffered and handled one at a time. To offset the extra time serialization would potentially impose, operators would monitor and/or interact with rate-accelerated speech as needed [3].

In a series of recent participant studies with news and story-based speech materials, measures of attention, comprehension, and effort were significantly improved by mediated serialization, in marked contrast to concurrent listening at normal speaking rates in the same span of time [4].

More recently, we have developed and vetted a corpus of simulated Navy voice communications based on a short set of fictitious tactical scenarios. In the present report, we describe the outcome of a preliminary listening study with these speech materials in which we simulated a mission-specific attentional concern for rate-accelerated voice communications in virtual auditory displays.

2. BACKGROUND

Navy watchstanders work in heavily loaded, multitask settings and must attend to and integrate a wide variety of auditory and visual information tasks. Specialists who have responsibility for particular tactical information domains,



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

such as air or (ocean) surface defense, sit before a visual representation of entities being tracked in the operational theater, monitor and initiate relevant voice communications, and maintain an up-to-the-moment assessment of the tactical situation and their ship's capacity to act. Recent growth in shipboard information capacities has profoundly increased the human performance challenges of this work. Training in the expertise and skills these positions require is done at Navy facilities where teams learn current operational practices on legacy systems and coordinate with each other in highly realistic tactical simulations. Coverage of all requisite voice communications is ensured via team augmentation and redundant monitoring. The communications workload is limited to two active channels (operationally referred to as circuits) per operator, and critical circuits are assigned to two or more individuals as needed [4].

Fleet modernization has brought with it an ongoing opportunity to study how watchstanding can be reshaped to move beyond the constraints of its present operational framework. Increased task loads are already being supported by tactical information systems that have far more functionality and three times the visual display space of previous consoles.

With new ship classes coming online, machine-aided techniques for conveying and managing the display of competing information tasks are being investigated. The intent is to reduce the metacognitive effort associated with unassisted multitasking and to structure the presentation and/or modulation of information in accord with the operator's perceptual strengths. Team augmentation and redundant listening, for example, are not optimal uses of personnel, but this strategy does succeed in minimizing the operational risk of missing critical information in the process of having to attend to two voices at once. Ideally, operators should be able to focus on one message at a time. Mediated serialization of the communications task would allow operators to do this and would also afford a number of optimizations. As competing messages are enqueued, they could be rendered to text, analyzed for priority and duration, and their rates of speech accelerated as needed. The effort of divided listening would be alleviated, and because of varying distributions of idle time on competing channels, trained operators could reasonably cover an array of more than two voice communications circuits in a virtual auditory display [4].

Important listener performance questions are raised by the use of serialization and rate-accelerated speech, especially if the latter is to be rendered as virtual sound. These concerns include listeners' abilities to attend to and encode rapid messaging, adapt to imposed aural attention switching, and maintain and resume an understanding of multiple, aurally modulated, information contexts. In a

sequence of studies with continuous speech materials developed from news commentaries and short narratives used in age-related cognition research, listeners in our lab were found to be markedly (and significantly) better at following and understanding serialized, rated-accelerated speech from a forward array of up to four spatially separate sources in a three-dimensional auditory display than they were when listening to two concurrent, opposing, and unaccelerated sources in the same virtual setting [5][6][4]. Attention and comprehension were measured, respectively, by the ability of participants to discriminate between actual and falsely sampled content while they were actively listening and, afterwards, to categorize queries derived from the spoken materials as being in agreement with, or not stated in, what they had heard. Listening performance was found to be equivalent for normal and accelerated listening up to an increase of at least 65% and then to exhibit an approximately linear decline that remains well above chance up to an increase of at least 125% [6][4]. Mixed costs were found for imposed attention switching in manipulations that compared serial listening at normal and 100% faster rates of speech. Attention and comprehension dropped significantly when the rate of speech was doubled but remained substantially above the same measures for current listening, whereas only attention exhibited an additional significant drop—albeit modest—when successive utterances in four serialized contexts were also randomly alternated to completion, as opposed to not being alternated [7].

These findings show that mediated serialization and rate-accelerated speech may in fact be a feasible technology for increasing performance capacities in Navy voice communications. Up to this point, however, only continuous speech has been evaluated, which is not representative of the radio communication patterns operators are actually exposed to. There are important reasons for having adopted this approach. Since continuous speech does not feature intermittent periods of idle time, it allows effectively equivalent performance comparisons to be made between serial and concurrent listening and between normal and accelerated listening. The next step is to explore these and other performance questions with more realistic speech corpora situated in the context of a Navy-relevant task.

As was noted earlier, watchstanders must attend to and integrate a wide variety of auditory and visual information. One of the voice communications skills they learn is the use of changing sets of code words, which are employed to disguise names and other references that would expose operational goals. In our previous work, we measured aural attention performance with ordered checklists of spoken phrases from each virtual source of speech, wherein listeners had to mark phrases they heard and pass over spurious phrases. In a more Navy-like setting involving several channels of speech communications, listeners would be expected to be aware of code words and this could also serve as a measure of aural attention. In the remainder of this paper, we outline and present the results of a limited study that was designed to explore the ramifications of this type of attentional measure.

3. EXPERIMENTAL DESIGN AND METHOD

As part of a related research effort studying the use of chat-based communications in watchstanding operations, we recently developed a corpus of interrelated voice communications with multiple talkers on four radio circuits that serve different operational functions. The scripted speech

materials cover four fictional naval scenarios that run for about seven or eight minutes each. Each scenario involves an ongoing tactical operation in which the listener, in the role of a watchstander, is expected to monitor the actions of several radar-tracked air and surface entities/objects that are verbally identified and visually depicted on a corresponding tactical situation display, known as a “TACSIT.” The scenarios are designed for use in a laboratory-based mock-watchstanding environment, and an experienced Navy reservist has vetted the visual and spoken content for operational realism and difficulty appropriate to non-specialists in an experimental setting.

The laboratory setup entails a multi-screen tactical information console and a head-tracked immersive auditory display with a fixed, real-world frame of reference corresponding to the console’s center screen. Speech and/or cueing and other forms of auditory information are virtually positioned in the listening space using non-individualized head-related transfer functions (HRTFs) and are binaurally rendered with a stereo headset. The center screen displays the TACSIT, and other tactical information tasks are shown separately, as needed, on adjacent screens to the left and right.

For the present study, we identified a different set of eight “code words” in each of the communications scenarios and modified our TACSIT software to show a given set as a list in an onscreen box, together with an overhead depiction of each voice circuit’s virtual location in the auditory display relative to the listener. A screen shot of the TACSIT and these additional elements is shown in Figure 1. Both the code word list and the four circuit positions, labeled “1,” “2,” “3,” and “4,” were implemented as interactive widgets. The list was only visible when its box was moused over with the computer’s pointer, and the circuit labels functioned as clickable buttons. Interactions with the widgets were programmed to be logged as time-stamped performance data.

The code word lists were then used as the basis of an active listening task. Five volunteer listeners from our laboratory (three men and two women, with a mean age of

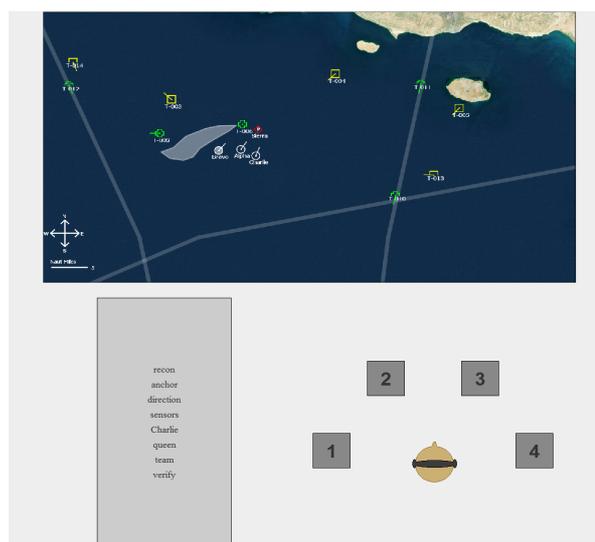


Figure 1. A screenshot of the visual display listeners used. The TACSIT showing a number of radar-tracked objects is positioned at the top. The list of code words can be seen in the tall box on the lower left, and the interactive depiction of each voice circuit’s virtual location in the auditory display relative to the listener is positioned to the right of the list.

34) were given a short time to study and commit one of the lists to memory. Next, they listened to the corresponding scenario and were told they could follow radar tracks and attendant behaviors on the TACSIT as they were mentioned on any of the radio circuits. At the same time, they were asked to spot any spoken instances of the listed code words and to indicate where each instance came from by clicking on the corresponding circuit label as quickly as possible. Approximately equal numbers of code words were spoken on each circuit during the exercise, with half of the eight words distributed across the four circuits and occurring only once and the others occurring up to four times. The four circuits were virtually positioned in the listener’s forward horizontal plane at 75° and 25° to the left and 25° and 75° to the right of the console’s centerline; the spread between these positions was exaggerated in the visual display to make it easier to click on the circuit labels (see Fig. 1). All of the voice communications were serially interleaved, meaning that the temporal order of utterances across all circuits was preserved in the manner of a first-in-first-out queue, but only one circuit was sounded at a time. To determine how many times listeners needed to look at the list for verification, as well as the amount of time they spent looking at the list, the code words were intentionally hidden during the listening exercise, but could be revealed, if needed, by mousing over the list box. Participants were urged to refer to the list as little as possible.

Each participant performed four word-spotting exercises, each based on a different scenario and each corresponding to a different manipulation of the speech materials. The voice communications were unaccelerated in one of the exercises and were uniformly 50%, 65%, and 100% faster in the other three, respectively. Speech in the faster manipulations was accelerated with a speech analysis/synthesis technique known as pitch-synchronous segmentation developed at our facility in the early 1990s that preserves pitch and facets of the speech waveform associated with intelligibility [8]. To ensure that shorter utterances corresponded to what was being shown on the TACSIT, which was not accelerated in the faster manipulations, each accelerated utterance was played at the same point in time it had originally been scripted to occur prior to being accelerated. To be clear, the visual part of each of the “faster” scenarios ran for its original length of time, and each accelerated utterance u_{accel} started at the same time t , relative to the start of the visual part of its scenario, as its source $u_{unaccel}$ did in the original,

Table 1. Distribution of code words spoken on each radio circuit in each manipulation. Four of the eight code words participants were asked to spot in each listening exercise occurred only once and were uniformly distributed; these occurrences are respectively indicated with the number “1” in each cell of the 4x4, manipulation-by-circuit matrix shown in the table. The remaining four code words in each exercise were spoken up to four times and were distributed so that the total number of code words spoken on each circuit was approximately equal; these occurrences are indicated with the parenthetical numbers in each cell. In the first cell, for example, three different code words were spoken, one being said three times, for a total of five occurrences.

	Circuit 1	Circuit 2	Circuit 3	Circuit 4
Normal speech	1+(3+1)=5	1+(3)=4	1+(3)=4	1+(1+2)=4
50% faster	1+(3+1)=5	1+(4)=5	1+(3)=4	1+(3)=4
65% faster	1+(3)=4	1+(2)=3	1+(3+1)=5	1+(1+1+1+2)=6
100% faster	1+(4)=5	1+(4)=5	1+(4)=5	1+(1+3)=5

“unaccelerated” version of the scenario. The unaccelerated exercise was given to all participants first, and the other three were given to each in a successively changed order. The distribution of spoken code words on each circuit in the four manipulations is described in Table 1.

In the following analysis, list visits, list look times, the timing of mouse clicks on circuit numbers in the visual display, the number of errors, and the proportions of correct responses were treated as dependent variables. Our expectations were a) that listeners would uncover the code lists two or three times during each exercise, b) that response times would be slower in the faster manipulations, c) that there would be few if any erroneous identifications of the circuit a given code word was spoken on, and d) that the mean proportion of correct responses across all manipulations would be above 50%, with the lowest scores occurring in the faster manipulations.

4. RESULTS

There were no significant differences in the number of list visits across the four speech rates, $F(3,9) = 0.434, p = 0.731$. However, the mean number of list visits per listening exercise was 10.35, which was much higher than anticipated. There were also no significant differences in the average amount of time participants looked at the list (and/or kept the list visible) across the four manipulations, $F(3,9) = 1.515, p = 0.276$. On average, participants spent roughly 5% of each exercise referring to the code words. In summary, increasing the speaking rate of the talkers on each of the radio circuits by up to 100% did not lead to meaningful changes in the number of times subjects looked at the list of code words nor in the amount of time the list was kept visible on the TACSIT.

In the response data, one listener’s clicks were lost in the 65% and 100% faster manipulations due to a technical problem. The remaining data for this participant was included in the following analyses. There were no significant differences in the time listeners took to click on a circuit after spotting a code word, $F(3,9) = 2.234, p = 0.154$. A mixed criterion was used for this measure: a 4000 ms cutoff was applied unless the code word was embedded in an utterance that took more than this amount of time to complete. Contrary to what was expected, the mean values for this performance metric ranged from 3240 ms for normal speech to 1747 ms in the 65% faster manipulation; The next slowest mean response time (2598 ms), however, occurred in the 100% faster scenario. There were no significant differences in the number of errors listeners made across the four listening exercises, $F(3,9) = 1.0, p = 0.436$. An average of 2.06 errors were made in each manipulation, an error being defined as clicking on the wrong circuit when a code word was spoken. More notably, the total number of clicks listeners made decreased significantly as the rate of speech was accelerated. Thus, the proportion of code words listener’s spotted and clicked the correct source of was

Table 2. Summary of mean performance measures in each manipulation. “*” indicates a main effect.

	Normal speech	50% faster	65% faster	100% faster
List visits	9.6	13.8	8.25	9.75
List look time (s)	15.3	27.5	17.5	23.4
Resp. time (ms)	3240	2133	1747	2598
Errors	1.5	3.25	1.75	1.75
Prop. correct*	0.4853	0.2639	0.3472	0.2000

significantly predicted by speech acceleration rate, $F(3,9) = 8.440$, $p = 0.006$, $\eta^2 = 0.738$. This proportion dropped from nearly 50% for unaccelerated speech to 20% when the rate of talker's speech was doubled. A summary of the performance measures discussed up to this point is given in Table 2.

To assess performance differences associated with the central and/or peripheral circuits, a 4x4 repeated measures ANOVA of the corresponding proportions of correct responses across the four rates of speech, showed there was no main effect of spatial position, $F(3,9) = 0.072$, $p = 0.974$. However, there was a significant circuit by speed interaction $F(9,27) = 5.542$, $p < 0.001$, $\eta_p^2 = 0.649$. These results are depicted in Figure 2.

To more closely examine the interaction, separate repeated measures ANOVAs of the word spotting responses were conducted for each rate of speech. As can be seen in Figure 2, listeners correctly responded to a greater proportion of code words spoken on the third circuit (0.80, right central position) in the normal speech manipulation (blue line) than in any other part of the study; the performance differences in this manipulation were marginally significant, $F(3,12) = 3.124$, $p = 0.066$, $\eta^2 = 0.439$. In addition, the comparatively high proportion of correct responses to code words on the fourth circuit (0.55, right peripheral position) was significant, relative to performance on the other circuits in the 100% faster manipulation (purple line), $F(3,9) = 10.500$, $p = 0.003$, $\eta^2 = 0.778$. There were no significant performance differences between the four circuit positions in the 50% and 65% faster manipulations (respectively, the red and green lines).

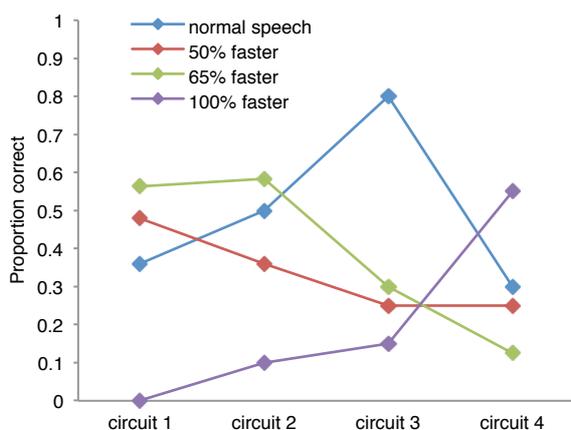


Figure 2. Correct identifications of circuits code words were spoken on, expressed as a proportion, in each of the four manipulations.

5. DISCUSSION

Several performance questions for mediated serialization of voice communications were explored in this preliminary study. In particular, the use of code words in Navy tactical communications was studied as a novel way to compare aural attention performance under normal and accelerated rates of speech as well as when rendered at different source positions in a virtual auditory display. While no test of scenario comprehension was included in the study and only a small number of participants were recruited, the findings showed that when listeners recognized code words, they could

generally identify its virtual source shortly after it was spoken regardless of the rate of speech. Surprisingly, instead of being slower to respond to accelerated speech as we conjectured, listeners responded progressively faster in the 50% and 65% manipulations (see “Resp. time (ms)” in Table 2). The reasoning behind our expectation was that because making sense of speech can occasionally require a momentary mental review of what has been variously called “echoic memory” [9], “precategory acoustic storage” [10], and “brief auditory storage” [11], it therefore seems plausible that listeners without exposure or practice might need to do this more often when processing rate-accelerated speech, and so, become progressively slower to respond. The decline in the mean proportion of correct responses shown in Table 2 as the rate of speech increases across manipulations—albeit, not fully linear and not significant until the rate of acceleration is 100%—is consistent with this processing conjecture in the sense that attentional difficulties tend to correlate with missed information. Moreover, listeners’ mean response time in the 100% faster condition was the second slowest in the study, suggesting that, as in our previous studies (e.g., [6]), at some point above an increase of 65% in the rate of normal speech, aural processing begins to require genuine attentional effort. Noting that “correct responses” required listeners to identify the circuit a code word was spoken on, rather than the word itself (thus, only implying that a specific code word was heard), the unexpected pattern of response times observed here may reflect a native attentional ability to adapt to the pace of auditory events up to a point, much as melodic and/or rhythmic information can generally be followed within a range of tempos and becomes difficult to process outside of this range (see, e.g., [12] and [13]).

That said, it is evident that the word-spotting part of the experimental task was not as easy as we had anticipated. While lower scores did occur in the faster manipulations as anticipated, the mean proportion of correct responses across all manipulations was below 50%, and despite the seemingly low numbers shown in Table 2, listeners made more errors than were expected. Put another way, listeners simply missed more than half of the code words in each of the manipulations, and although on the whole they were equally attentive to both the central and peripheral circuits (see Fig. 2: by circuit, the mean proportion correct responses ranges from .3125 for circuit 4 to .3858 for circuit 2), on average, listeners clicked on the wrong circuit 27% of the time (the error rate ranged from 15.4% for normal speech to 40.6% when the speech was 50% faster). Given the extent and pattern of missed code words across manipulations and the surprising rate of source identification errors, it is clear that listeners struggled to do well. Factors that may have contributed to their performance difficulties include distracted listening arising from looking at the code word list (note that these numbers in Table 2 are much higher than were expected), response completion errors arising from ongoing listening demands, attentional fatigue arising from varying distributions of code words in each manipulation and the relative sparsity of instances (just under 2.5 words per minute) and, to a lesser extent (because of the horizontal layout of source positions) poor display fidelity due to the use of non-individualized HRTFs. An aspect of the spoken material worth considering, that may have also influenced listeners’ performance, is the relatively innocuous nature of the code words that were used in each manipulation. The incorporation of a range of Navy-like code words—words that are generally colorful and somewhat salient relative to ordinary speech—was not considered in the scripting and

recording of the radio communications part of the four tactical scenarios, which was done several months before the study this conceived. Thus, words that could plausibly function as “code words” had to be identified in each script. A range of nouns (“knots,” “sensors,” “queen,” etc.) and, to a lesser extent, verbs (“proceed,” “verify”) and response words (“aye,” “copy”) were chosen within the fixed wording of the scripts so as to achieve a relatively uniform spread of words for listeners to spot across the four circuits in each scenario; the full selection of code words is given in Table 3. Additionally, in each manipulation, four of the eight code words were spoken only once and the remaining four were said multiple times in roughly equal numbers (see Table 1). Performance very nearly at or above 50% correct occurred in only six cells of the 4x4 manipulation-by-circuit response matrix (these measures are plotted in Fig. 2). Listeners only needed to spot two words, “report” and “whiskey,” in the best of these cells (circuit 3 under normal speech), and it could be argued that “whiskey” is a more readily spotted word than any or most of the more ordinary words participants were asked to listen for. In contrast, though, listeners also only had two words to spot in three of the four cells with the lowest performance (circuits 1, 2, and 3 under 100% faster speech and circuit 4 under 65% faster speech). Curiously, in the lowest of these cells (circuit 1, with no correct responses), one of the two words was “reconnaissance,” which was chosen for its potential salience. In spite of these somewhat contradictory performance patterns, similar to a point made above, the predominant occurrence of the poorest performance in the study under the fastest rate of speech is consistent with our earlier finding that listeners perform at parity with normal speech only up to a 65% increase in the rate of speech. In the other lowest performing cell (circuit 4 under 65% faster speech), a different factor may have interacted with the listening task: here, instead of only two words, there were five separate words to spot—more than in any other cell in the study—giving listeners more work to do. Even so, this observation is somewhat countered by the best performance in the 100% faster manipulation (circuit 4) wherein, unlike circuits 1, 2, and 3 in this manipulation, there were three words to spot. The possibility that many of the code words were not memorable is also supported by the much higher than expected numbers of list looks participants resorted to in each manipulation in spite of having been given ample time to study each list before each of the listening exercises.

In addition to examining a structured set of performance questions, the broader intent of this study was to gain a preliminary sense of how carefully listeners are likely to listen in a somewhat “realistic” serialized multimodal

Table 3. Lists of code words listeners were respectively asked to spot in each of the four manipulations. The first four in each list were only spoken once (each on a separate circuit) and the remaining four were said more than once in a given manipulation (see Table 1.)

Normal speech	50% faster	65% faster	100% faster
attention	inbound	supplies	reconnaissance
route	vessel	information	anchor
report	feet	neutral	direction
threat	mark	reflection	sensors
proceed	copy	knots	charlie
intentions	blockade	waters	queen
whiskey	values	aye	team
sea	status	stations	verify

framework wherein both auditory and visual information display components are referentially related. If mediated serialization of competing aural information tasks is to be adopted, it must be viable within a unifying operational context such as tactical situation monitoring, which was used here, or air traffic control. The counterpart of this study’s measure of auditory attention will be an evaluation of listeners’ attention to and knowledge of the tactical information content of the scenarios in a future experiment. Another aspect of managed task switching we have explored in the past and now plan to study in an integrated operational setting is the utility of virtual auditory cueing as a technique for guiding the operator’s attention from one task to the next [14][15][16]. Although auditory cues significantly improved task performance in a series of prior studies with a cockpit-like dual task involving rapid decision making and continuous tracking, a range of additional questions are raised by their use in mixed auditory information settings. Among these are the development of an empirically based set of organizing principles for the presentation of competing sounds and performance-based evaluations of different cueing designs for cross-modal task switching involving prioritization and modulated information displays such as rate accelerated speech and visual augmentation to guide the operator’s attentional focus.

6. ACKNOWLEDGMENT

This research was supported by the Office of Naval Research under work order N0001412WX20879.

7. REFERENCES

- [1] D. Wallace, C. Schlichting, 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, January, 2002.
- [2] C. T. Bush, J. R. Bost, P. S. Hamburger, and T. B. Malone, “Optimizing manning on DD21,” in *Proceedings of the Association of Scientists and Engineers (ASE) 36th Annual Technical Symposium*, April, 1999.
- [3] B. McClimens, D. Brock, and F. E. Mintz, “Minimizing information overload in a communications system utilizing temporal scaling and serialization,” in *Proceedings of the 12th International Conference on Auditory Display (ICAD)*, London, UK, June, 2006.
- [4] D. Brock, C. Wasylyshyn, B. McClimens, and D. Perzanowski, “Facilitating the watchstander’s voice communications task in future Navy operations,” in *Proc. of the 2011 IEEE Military Communications Conference (MILCOM)*, Baltimore, MD, 2011.
- [5] 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 *Proceedings of the 14th International Conference on Auditory Display (ICAD)*, Paris, France, June, 2008.
- [6] C. Wasylyshyn, B. McClimens, and D. Brock, “Comprehension of speech presented at synthetically accelerated rates: Evaluating training and practice effects,” in *Proceedings of the 16th International Conference on Auditory Display (ICAD)*, Washington, DC, USA, June, 2010.

- [7] D. Brock, S. Camille Peres, and B. McClimens, "Evaluating listeners' attention to and comprehension of serially interleaved, rate-accelerated speech," in *Proceedings of the 18th International Conference on Auditory Display (ICAD)*, Atlanta, GA, USA, June, 2012.
- [8] 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, November, 1994.
- [9] U. Neisser, *Cognitive Psychology*, New York: Appleton-Century-Crofts, 1967.
- [10] R.G. Crowder and J. Morton, "Precategorical acoustic storage," *Perception and Psychophysics*, vol. 5, pp. 365-373, 1969.
- [11] C.J. Darwin, M.T. Turvey, and R.G. Crowder, "An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage," *Cognitive Psychology*, vol. 3, pp. 255-267, 1972.
- [12] M.M. Baese-Berk, C.C. Heffner, L.C. Dille, M.A. Pitt, T.H. Morrill and J.D. McAuley, "Long-term temporal tracking of speech rate affects spoken-word recognition," *Psychological Science*, vol. 25, pp. 1546-1553, 2014.
- [13] J.D. McAuley, "Tempo and rhythm," in M.R. Jones, R.R., Fay, and A.N. Popper (Eds.), *Music Perception*, Springer Handbook of Auditory Research, vol. 36, pp. 165-199, 2010.
- [14] 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.
- [15] D. Brock, B. McClimens, and M. McCurry, "Virtual auditory cueing revisited," in *Proceedings of the 16th International Conference on Auditory Display*, Washington, DC, June 9-15, 2010.
- [16] D. Brock and B. McClimens, "To what extent do listeners use aural information when it is present?" in *Proceedings of the 17th International Conference on Auditory Display*, Budapest, Hungary June 20-24, 2011.

ORAL PAPERS

*3D Audio and Spatial
Sound*

A BONE CONDUCTION BASED SPATIAL AUDITORY DISPLAY AS PART OF A WEARABLE HYBRID INTERFACE

Amit Barde, Gun Lee

HIT Lab NZ,
University of Canterbury,
Private Bag 4800,
Christchurch 8140, New Zealand.
amit.barde@pg.canterbury.ac.nz,
gun.lee@canterbury.ac.nz

Matt Ward, William S. Helton

Department of Psychology,
University of Canterbury,
Private Bag 4800,
Christchurch 8140, New Zealand.
matt.ward@pg.canterbury.ac.nz,
deak.helton@canterbury.ac.nz

Mark Billinghurst

School of ITMS,
University of South Australia,
Mawson Lakes, SA 5095, Australia.
mark.billinghurst@unisa.edu.au

ABSTRACT

Attention redirection trials were carried out using a wearable interface incorporating auditory and visual cues. Visual cues were delivered via the screen on the Recon Jet – a wearable computer resembling a pair of glasses – while auditory cues were delivered over a bone conduction headset. Cueing conditions included the delivery of individual cues, both auditory and visual, and in combination with each other. Results indicate that the use of an auditory cue drastically decreases target acquisition times. This is true especially for targets that fall outside the visual field of view. While auditory cues showed no difference when paired with any of the visual cueing conditions for targets within the field of view of the user, for those outside the field of view a significant improvement in performance was observed. The static visual cue paired with the binaurally spatialised, dynamic auditory cue appeared to provide the best performance in comparison to any other cueing conditions. In the absence of a visual cue, the binaurally spatialised, dynamic auditory cue performed the best.

1. INTRODUCTION

One of the most common ways of interacting with mobile devices is through visual interfaces. Similarly, widely available wearable computers such as Google Glass [1] and the Recon Jet [2] also use visual interfaces as the primary means for information delivery. However, visual presentation can overwhelm a user due to an inordinate number of data streams vying for the same screen space, or the users' inability to divide attention between the presented information streams on the screen and the world around them. Either way, the information presented only via the visual faculty can overload the users' senses and cognitive ability [3].

Addressing this problem for wearable displays, with their severely limited screen space and constant demand on the user's attention, is particularly important. Attention critical tasks such as driving, search and rescue etc. may be adversely

affected by the use of such devices [4] [5] [6]. This presents us with a set of unique challenges; (1) information presentation without overloading the user and (2) unobtrusive information delivery requiring minimum attention from a user perspective.

As part of our research, we are interested in exploring the use of binaurally spatialised auditory cues delivered over a bone conduction headset (BCH) for wearable interfaces. By doing this, information can be presented using auditory and visual cues and hopefully reduce the problem of information overload. In the remainder of this paper we first review related work and then describe an experiment exploring the effectiveness of using several combinations of auditory and visual cues for information presentation in a wearable computer interface. We then summarise results from the experiment and conclude, providing suggested directions for future research.

2. BACKGROUND

Wearable spatial auditory displays have been the subject of research for well over two decades now [7] [8] [9] [10] [3]. Almost all of these systems incorporate the use of Head Related Transfer Functions (HRTF) [11], either individualised or non-individualised, to deliver a three dimensional synthetic rendering of the acoustic environment. However, the drawback with these systems is the use of headphones to deliver the required auditory cues or create an auditory environment. This can isolate a user from his/her acoustic environment [12] [13]. One solution to this problem is the implementation of techniques that allow the capture and reproduction of the ambient acoustic environment. Harma et al. [14] and Tikander et al. [15] have demonstrated the use of such an 'audio augmented reality' device. Devices such as the Here Active Listening system [16] developed by Doppler Labs are gaining popularity in the consumer market.

All these systems need to be inserted in to the ear canal, and rely on external microphone inputs to reproduce the ambient environment around the user. Their sizable form factor also occludes the pinna, meaning signals reaching the

microphone are not necessarily those that have been filtered as a result of the reflections resulting from the shape of user's pinna. From existing literature we know that the occlusion of the pinna can cause a greater number of front – back confusions and a diminished ability to localise in the vertical plane [17].

The problems described above are unacceptable in attention critical environments that demand high levels of awareness of auditory and visual cues. Leaving the ears open to the ambient acoustic environment while engaging in a visually demanding task is safer than having the ears plugged. To overcome this problem, we use a bone conduction headset (BCH) to deliver binaurally spatialised audio as part of a wearable interface. Previous studies exploring the use of the BCH as an auditory display device appear promising [18] [19] [20] [21] [22]. While the use of the BCH has been primarily restricted to its implementation as an auditory display for the visually challenged [18] [20], some studies have demonstrated its effectiveness even for sighted users [23] [24]. However, besides [23] [24] we are unaware of any studies that incorporate the use of BCH as part of a wearable interface. We hope to demonstrate the practical utility of a BCH, as part of wearable interface, to deliver binaurally spatialised cues.

In the following sections we describe our implementation of the BCH as a spatial auditory display device as part of a wearable hybrid interface. We then present a user study conducted to evaluate the use of audio-visual cues in a visual search task. The ability to reorient attention with the use of these cues is explored. Existing studies demonstrate that there is a significant increase in visual search task performance when auditory cues are used [25] [26] [27]. The amalgamation of the visual and auditory faculties and the ability to exploit their mechanisms of perception could make for a more efficient interface than one that relies completely on a single modality. Such an interface assumes great importance in attention critical fields such as driving and search and rescue. Being able to receive task specific information without having to divide attention between the primary task and information retrieval that may directly affect the outcome of the task is important. For example, receiving binaurally spatialised auditory beacon based navigation during a driving task is safer than having to constantly direct one's visual attention to a GPS device. While the navigation information is critical to the primary task of driving and directly affects its outcome, information delivery can take place in a manner that does not affect the primary task in manner that renders it unsafe. In our study we demonstrate the use of this and other auditory cues along with visual presentations made on wearable device.

3. PARTICIPANTS

30 participants (20 male, 10 female) between the ages of 18 and 34 (Mean: 24, Std. Dev: 4.3) volunteered to take part in the study. Participants reported normal hearing in both ears. No testing was carried out to verify their claims of having normal hearing since there appears to be no known relation between localisation performance and audiogram results unless the hearing loss is profound [25]. All participants were compensated with \$20 shopping vouchers for their efforts.

4. METHOD

4.1. Apparatus

The apparatus used for the study can be divided in to three categories; 'real-world' analogue (for far domain stimuli presentation) and tracking equipment, handheld and worn tracked devices used by the participants and a bone conduction headset (BCH) used to deliver auditory cues.

The real-world analogue used for this experiment was a set of three screens connected to each other at 60°. The screens measured 2400mm x 1830mm (74.65° x 59.53° visual angle) and were mounted 600 mm above the floor. Images were projected on to these screens by three NEC LT265 projectors. The tracking system used for the study comprised of four ARTTRACK2 cameras mounted on top of the screens (see figure 1), paired with the DTrack software [28]. The cameras are capable of tracking objects up to a distance of 4.5m. For detailed specifications of the camera see [29].

Equipment used by participants consisted of a Recon Jet [2], a wearable 'smart glass' and the Steradian S-7X laser tag gun [30] (see figure 2). Both these devices had markers affixed to them to allow their positions to be tracked during the course of the experiment. The laser tag gun was modified such that depressing the trigger on the gun allowed the participant to 'shoot' targets that were displayed on the screens. This was achieved by connecting a pair of leads attached to the trigger inside the gun to the circuit board of a mouse. The Recon Jet has a widescreen 16:9 WQVGA display with images on it set to appear as they would on 30 inch HD display at 7 feet. For more detailed technical specification see [2]. The Recon Jet was connected wirelessly through a router. Tracking data was transferred to the PC using the VRPN software [31] [32].

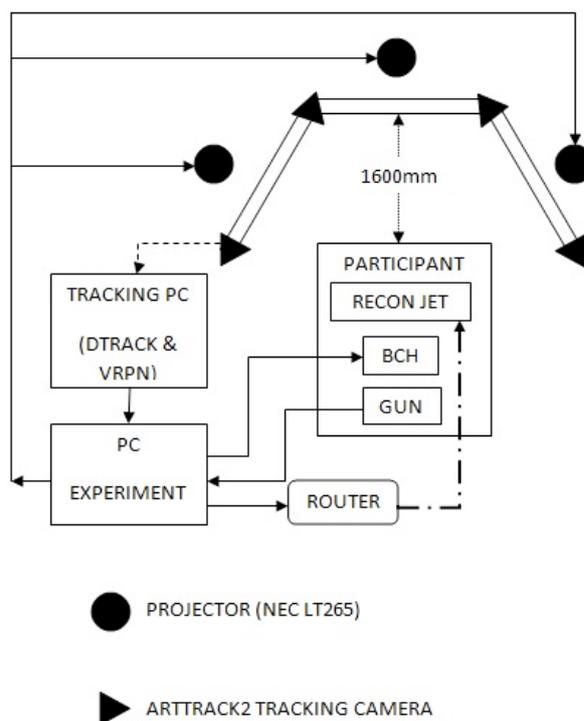


Figure 1: Block diagram of the experimental setup

Participants also wore a pair of bone conduction headsets (Aftershokz Sportz3) [33]. Auditory stimuli for the experiment were reproduced over the BCH. The auditory

stimuli were delivered to the BCH via the PC's on-board sound card.



Figure 2: Participant holding the laser tag gun with markers affixed on top to allow the position of the gun to be tracked. Also seen in the picture are the Recon Jet with markers for tracking, and the BCH.

4.2. Stimuli

Participants were presented auditory and visual stimuli for the experiment. A detailed explanation of the stimuli is given in the following sections.

4.2.1. Visual Stimuli

Visual stimuli were delivered on the screens representing the far field and the Recon Jet display. Stimuli displayed on the screens were targets that appeared at random intervals during the experiment, and a string of numbers at the bottom of the centre screen. Targets consisted of yellow discs of approximately 58mm radius that turned blue when shot and appeared at predefined positions of $\pm 50^\circ$ and $\pm 100^\circ$ (see figure 3). The targets were positioned at the centre of the screens. The numerical string used a black Arial typeface of 65mm size positioned in the horizontal centre and approximately 690mm below the vertical midpoint of the centre screen.

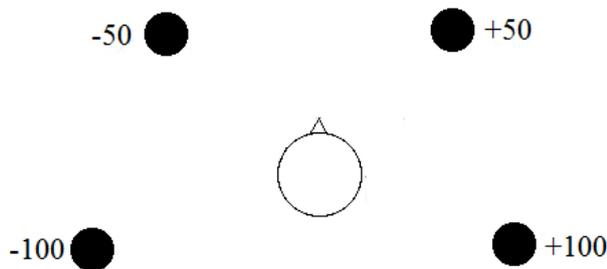


Figure 3: Target Positions

Text displayed on the Recon Jet used a white Arial type face and was positioned in the centre of the screen (see figure 4). Preceding messages were listed above in grey. Visual interrupt signals delivered on the Recon Jet consisted of static cues, pursuit visual cues and a blank screen. The static visual cues consisted of white arrows 1.3° in width and 6.5° in length (see figure 5). The arrows were angled at 40° for targets appearing at $\pm 50^\circ$ and 80° for targets at $\pm 100^\circ$. Pursuit

visual stimuli caused all objects on the screen to move in the direction of the target at 16.2° per second.

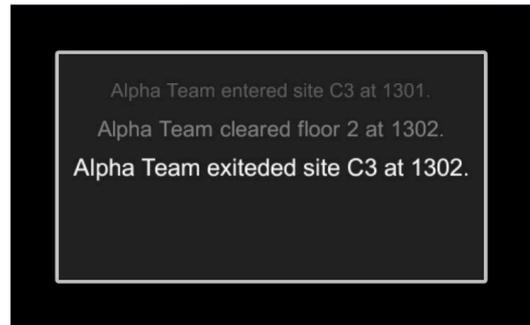


Figure 4: Messages displayed on the Recon Jet screen

In addition to the messages on the Recon Jet and targets projected on the screens, the participants also saw two smaller 'dots' on the screens. These dots represented the position of their head (yellow dot) and the position of the gun (blue dot). Both these moved around on the screen as the participant rotated along the horizontal arc on which the targets appeared.

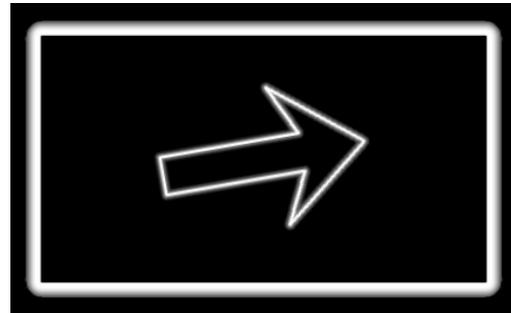


Figure 5: Visual Cue – White Arrow

4.2.2. Auditory Stimuli

Auditory stimuli consisted of a 1 second alarm sound or ping (25ms on set and offset rate). The same sound was used for two of the three types of auditory cues that were delivered. The alarm tone was presented either as a static sound or a binaurally spatialised dynamic audio cue moving in the direction of the target. The binaurally spatialised, dynamic cue simulated the motion of the alarm from the participant's position towards the target on the screen. The cue was designed in accordance with alarm design guidelines prescribed by Walker and Kramer in [34]. The duration and level of the auditory cue were chosen to represent those used by previous researchers [18] [35] [20] [36] [37]. Despite studies demonstrating that wideband noise is easier to localise [18] [20] [36] [38] [39] than most other forms of stimuli, we chose to use a ping for its aesthetic appeal [40]. The third auditory cue consisted only of silence. The static auditory cue was delivered at approximately 70dBA. The dynamic cue on initiation will have had approximately the same loudness level, but fell quickly as the cue moved towards the target. A logarithmic fall off with the addition of the Doppler Effect was modeled to replicate real world auditory percepts.

The visual and auditory cues used here are analogous in that they encompass similar perceptual characteristics, but in different domains (see table 1).

AUDITORY CUES		VISUAL CUES	
S0	No sound (Silence)	V0	Blank Screen
S1	Static alarm	V1	Static (Arrows pointing in the direction of the target)
S2	Binaurally spatialised, dynamic alarm	V2	Pursuit visual cue

Table 1: A list of the auditory and visual cues used in the experiment. A total of 9 cues encompassing a combination of all the cues above were presented to the participants.

The experiment was designed and built within the Unity3D [41] environment. Binaural spatialisation of the auditory cue was achieved using the 3Deception Binaural Engine plug-in for Unity developed by Two Big Ears [42]. We've chosen to adopt the use of a plug-in versus the traditional approach of using individualised HRTFs or HRTF libraries since we believe this lends a greater degree of ecological validity to the study. The plug-in was chosen after an extensive phase of testing and comparisons with existing binaural engines.

5. PROCEDURE

Participants were seated on a rotating chair 1600mm from the central screen (see figure 1). The position was situated approximately on the normal from the central screen such that targets could be presented anywhere on a 200° horizontal arc. Participants were allowed to adjust the height of the chair for comfort.

Participants were then handed the Recon Jet and BCH to put on. If required, they were helped positioning the screen of the Recon Jet so that the text displayed on the screen appeared clear. The BCH was put on such that the drivers of the headset sat in front of the ears on the mandibular condyle. Participants were also handed the laser tag gun. Following this, a short calibration process was run. This was to ensure that participants had a full range of motion that allowed them to reach targets at $\pm 100^\circ$, check if tracking information was being gathered in the right manner and ensure that participants were able to read text appearing on the Recon Jet and the main screen. Following the calibration process, six practice trials were conducted. These trials allowed participants to see the different audio – visual cues that could be presented to them during the course of the experiment.

The main experiment was separated in three blocks. Each block was followed by a five minute break. During each block participants were instructed to read aloud number strings appearing on the central screen and messages appearing on the Recon Jet. Messages displayed on the Recon Jet were chosen at random from one of three structure types; [Alpha] team entered site [C][37] at time [1407], [Alpha] cleared floor [2] at time [1407], or [Alpha] team exited site [C][37] at time [1407] (see figure 5).

During a third of the trials for which messages were displayed on the Recon Jet, a cue would interrupt the participant one second after the message's onset. Simultaneously, a target would appear at $\pm 50^\circ$ or $\pm 100^\circ$. Participants were required to 'shoot' or mark the target using the modified laser tag gun as quickly as possible (see figure 6). Once the target had been shot, participants returned their

gaze to the central screen, and the alternating display of number strings on the central screen and messages on the Recon Jet resumed. Within each block there was one target event trial for each combination of visual and audio cues for each position for a total of 36 target events (9 event types x 4 target locations) and 72 non-event messages. The experiment took on average 65 minutes to complete.

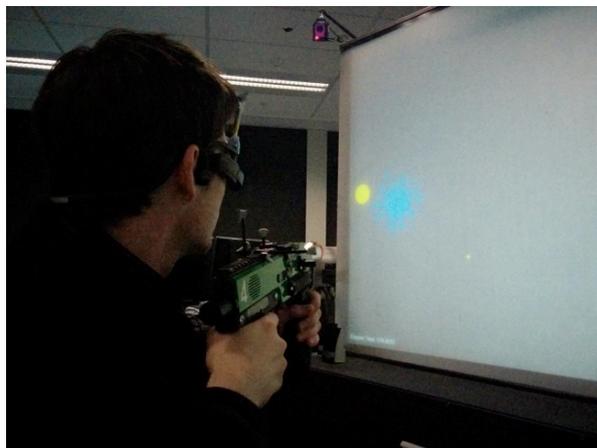


Figure 6: Participant attempting to shoot a target.

6. RESULTS

Two participants were excluded from the analysis – one for not following instructions, while the other had to be excluded due to failure of the on-board sound card to deliver audio signals. Additional technical difficulties with the Recon Jet, primarily associated with power management, meant data from the third block of trials for an additional six participants was recorded incompletely or lost. Data gathered from the third block was excluded from the analysis for all participants to maintain uniformity.

A 3x3x4 (visual cues: 3 auditory cues: 3 target positions: 4) repeated measures analysis of variance (ANOVA) was carried out to test for the main and interaction effects between the factors. Since the data violated Mauchly's test of sphericity, values as determined by the Greenhouse-Geisser correction were used. The results showed significant main effects of all three independent variables; audio cues ($F(1.677, 80.513) = 104.671, p < 0.001$), visual cues ($F(1.767, 84.822) = 60.736, p < 0.001$) and target positions ($F(2.244, 107.732) = 54.592$). The results also showed significant two way interactions between all pairs (visual x audio: $F(3.038, 145.835) = 16.041, p < 0.001$; visual x position: $F(3.432, 164.742) = 13.754, p < 0.001$; audio x position: $F(3.928, 188.535) = 8.248, p < 0.001$). In addition to this, a significant three way interaction between the three independent variables was also observed ($F(5.528, 265.30) = 4.054, p = 0.01$).

To look into the details of the interaction effects, we tested simple two-way interaction effects by fixing the levels of the target position. For the target at -100° , statistically significant interactions were found between the auditory and visual cues ($F(2.942, 158.852) = 11.414, p < 0.001$). Main effects of the visual ($F(1.806, 97.504) = 44.114, p < 0.001$) and auditory cues ($F(1.784, 96.32) = 47.764, p < 0.001$) also showed statistically significant results. Similar results are observed for the target at $+100^\circ$ with interaction effects between the two types of cues being statistically significant ($F(3.258, 169.395) = 5.852, p = 0.001$). The main effects of the cues also show statistically significant results (Audio: F

(1.562, 81.214) = 20.299, $p < 0.001$; Visual: $F(1.927, 100.179) = 30.847, p < 0.001$.

No significant interaction between the audio and visual cues were seen for targets at -50° ($F(2.209, 119.266) = 1.839, p = 0.159$) and $+50^\circ$ ($F(3.1, 161.204) = 0.971, p = 0.410$). For -50° there were significant main effects for both the auditory ($F(1.558, 84.145) = 37.285, p < 0.001$) and visual cues ($F(1.802, 97.331) = 5.982, p = 0.005$), while $+50^\circ$ displayed similar effects only for the auditory cues ($F(1.785, 92.81) = 30.743, p < 0.001$) and not the visual cues ($F(1.717, 89.272) = 1.859, p = 0.167$). The lack of a significant effect for the visual cues suggests that the peripheral vision over rides any of the visual cues when targets appear in these regions. Participants appear to lock onto these targets as result of the auditory cues and peripheral vision, rendering the visual cues ineffective.

The preceding analysis was then followed up with an analysis of variance (ANOVA) for each target position to compare performance between the different cues and their combinations. All positions displayed a significant difference in performance between the different cue types and their combinations ($-100^\circ: F(4.494, 242.66) = 31.008, p < 0.001$; $+100^\circ: F(3.656, 190.112) = 17.182, p < 0.001$; $-50^\circ: F(3.851, 207.944) = 12.063, p < 0.001$; $+50^\circ: F(5.018, 260.93) = 9.672, p < 0.001$). Post hoc tests using the Bonferroni correction of pair wise comparisons between the cueing conditions give a detailed picture of the effectiveness of the cues for each of the four positions. For the auditory cueing conditions (V0S0, V0S1 and V0S2) only, the static (V0S1) and dynamic auditory (V0S2) cues outperform the no cue condition, V0S0, at all target positions ($-100^\circ: p < 0.001$; $+100^\circ$ (V0S2): $p < 0.001$; -50° (V0S1): $p = 0.006$; -50° (V0S1): $p < 0.001$; $+50^\circ$ (V0S1): $p < 0.001$; $+50^\circ$ (V0S2): $p = 0.04$) except $+100^\circ$ V0S1 ($p = 1$). No significant difference was observed between the static (V0S1) and dynamic (V0S2) cueing conditions at -100° ($p = 0.194$), -50° ($p = 1$) and $+50^\circ$ ($p = 1$). A significant difference, though, was observed at $+100^\circ$ ($p = 0.008$). While further investigation is required, the binaurally spatialised auditory cue consistently demonstrates a faster onset of head motion time across all targets (see figure 7).

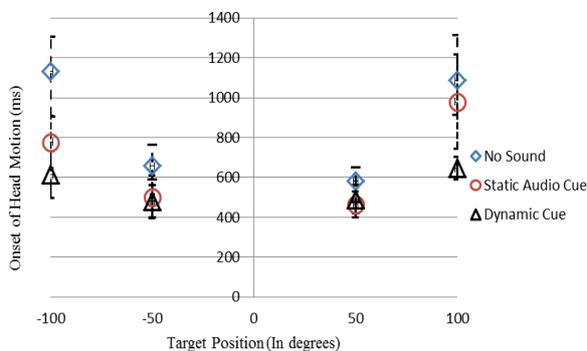


Figure 7: Average time for onset of head motion measured across all auditory cueing conditions.

For the cueing conditions using V1 paired with the auditory cues (V1S0, V1S1 and V1S2), a significant difference is seen between V1S0 and V1S2 at all positions ($-100^\circ: p = 0.02$; $-50^\circ: p < 0.001$; $+50^\circ: p < 0.001$; $+100^\circ: p = 0.001$). Significant differences were also seen between V1S0 and V1S1 at $+100^\circ$ ($p = 0.012$), -50° ($p < 0.001$) and $+50^\circ$ ($p = 0.001$), while -100° showed no significant difference between the cues ($p = 0.06$). This result is similar to the one

obtained with only auditory cues earlier. No significant differences were observed between V1S1 and V1S2 at any of the positions ($p = 1$). Both these conditions show closely matched onset times for head motion, with V1S2 displaying a marginally quicker onset (see figure 8).

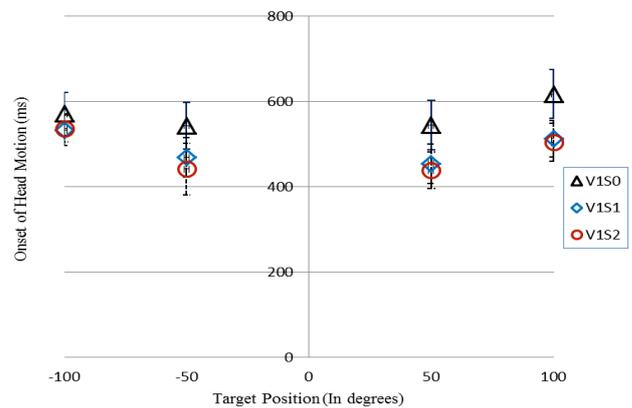


Figure 8: Comparisons between on-set of head motion times for the static visual cue (V1) paired with the auditory cues. A significant difference exists between on-set of head motion times for no auditory cue vs. auditory cueing conditions. No significant difference is seen between the static (S1) and dynamic (S2) auditory cueing conditions when paired with the static visual cue (V1).

For the cueing conditions using V2 paired with the auditory cues (V2S0, V2S1 and V2S2), a significant difference is observed between conditions V2S0 and V2S1 at -100° ($p < 0.001$), -50° ($p = 0.01$) and $+50^\circ$ ($p = 0.003$). No significant difference between the cueing conditions is seen at $+100^\circ$ ($p = 0.261$). Comparisons between V2S0 and V2S2 display significant differences at -100° ($p = 0.001$), -50° ($p = 0.044$), $+100^\circ$ ($p = 0.017$) and $+50^\circ$ ($p < 0.001$). Comparisons between the V2S1 and V2S2 pair does not show any significant difference at -100° , $+100^\circ$, -50° and $+50^\circ$ ($p = 1$). These cueing conditions (V2S1 and V2S2) appear to have similar onset times across all targets (see figure 9).

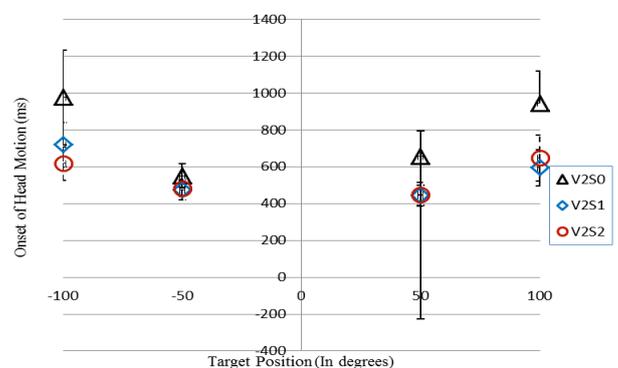


Figure 9: Comparisons between on-set of head motion times for the pursuit visual cue (V2) paired with the auditory cues.

From the analysis that has been carried out, it is clear that the presence of a visual or auditory cue definitely elicits a quicker onset of head motion from the time the target appears at any one of the positions. The lack of significant differences ($-100^\circ: p = 1$, $-50^\circ: p = 0.753$, $+50^\circ: p = 1$ and $+100^\circ: p = 1$) between the spatialised auditory cue (V0S2) and the static visual cue (V1S0), points to the fact that both these cues are

nearly equally good at redirecting attention. A combination of these two cues though, consistently registers the quickest time for the onset of head motion across all positions even though in some cases these differences do not appear statistically significant. This lack of statistical significance appears mainly when this cue (VIS2) is compared with other cues that include either a binaurally spatialised auditory cue (S2) or a static visual cue (V1). While VIS2 appears to be best suited for attention directions tasks, a comparison of the onset of head motion times between $+100^\circ$ and -100° for this cue shows that the cue performs better for the left side i.e. -100° (506.61 ms vs. 529.53 ms) (see figure 10).

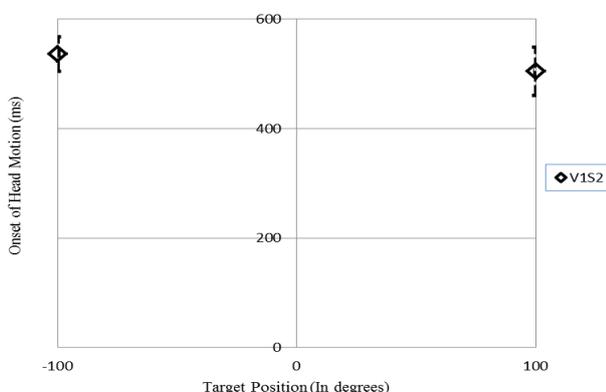


Figure 10: Onset of head motion times for the cueing condition V1S2 (static visual cue and dynamic auditory cue) for $\pm 100^\circ$.

7. DISCUSSION

We've been able to demonstrate the benefits of using auditory cues in an attention redirection task via this study. The results in this case can be categorized in to two distinct types: (1) auditory cues only and (2) audio-visual cues. The first part of the results section falls under the auditory cues category. The results obtained for these cueing conditions tend to suggest that the binaurally spatialised, dynamic auditory cues is effective for redirecting attention to targets that do not occupy the user's field of view i.e. $\pm 100^\circ$. The absence of a significant difference between the static (S1) and dynamic (S2) cueing conditions for targets at $\pm 50^\circ$ is likely due to the fact the targets fall within the user's peripheral vision. The onset of the auditory cues could possibly be the precursor to the participants localizing the target using their peripheral vision. This could be responsible for over-riding both the visual cues, negating their effect. This effect extends across the two auditory cueing conditions S1 and S2 paired with the two visual conditions V1 and V2 for targets at $\pm 50^\circ$. Conversely, when either of the auditory conditions paired with a visual cue was compared to the performance without an auditory cue, a clear difference in the onset of head motion times is observed. This is indicative of the fact that even in the presence of a visual cue delivered during a visually demanding task, an auditory cue is more likely to attract attention and help reorient user attention in the space around him/her. Another observation that points to the effectiveness of the binaurally spatialised dynamic cue is the absence of a considerable difference between targets on the same side i.e. $\pm 100^\circ$ and $\pm 50^\circ$. This result effectively demonstrates that the binaurally spatialised auditory cue is as good at redirecting attention to targets outside the visual field as the visual percept is at acquiring targets at $\pm 50^\circ$ in this experiment.

In the case of combinations of the visual cues, V1 and V2, with auditory cues S1 and S2, the pairing of the static visual cue, V1, with the dynamic auditory cue, S2 appears to provide the best results. As we've demonstrated with the auditory cueing condition only, these results show a superior performance in comparison to other cueing condition pairings when compared with onset of head motion and target acquisition times for targets outside the visual range. This study clearly indicates that the use of auditory cues in conjunction with visual cues to reorient attention is possible. The results from our study compare favourably with those of Perrott et al. [25], Nelson et al. [26] and Rudmann & Strybel [27].

8. CONCLUSION

We've demonstrated the use of a binaurally spatialised, dynamic auditory cue in conjunction with a visual cue to redirect user attention. These reorientation cues appear to be most effective for targets outside the visual field, but have also shown to be of use within the peripheral vision in comparison to having no auditory cue at all. The use of an auditory cue or alarm in a visually demanding task cannot be underestimated. The dynamic auditory cue appears to be able to redirect the user's attention without inducing a frantic search of the visual field, a behavior that was seen with the static auditory cues. Similar to a '3D' auditory cue delivering azimuth, elevation and distance information used by Nelson et al. [26], our dynamic auditory cue exhibits superior performance compared to the static cue. These results also demonstrate that the binaurally spatialised, dynamic auditory cue will be useful in the event that a user does not latch on to a visual cue that may be presented simultaneously. The outcomes from this study also appear to suggest that a 'dual delivery' of cues across two different modalities appears to ensure that the system is somewhat fail safe.

For the purpose of this experiment only four specific targets were used. In the future, it will be worthwhile exploring how both visual and auditory cues will perform in the presence of visual and auditory distractors. It also worth comparing the performance of binaurally spatialised, co-located cues with spatialised dynamic cues moving towards a target. The results from such experiments could provide further vindication for using the BCH as an auditory display incorporated in to a wearable computing interface and also give us an idea as to which of the cues provide better results for binaural spatialisation over a BCH.

9. REFERENCES

- [1] Heinrich, M.J., et al., *Wearable computing device*, U.S.P.a.T. Office, Editor. 2013, Google Inc.: U.S.A. p. 11.
- [2] ReconInstruments. *Recon Jet*. 2015 [cited 2015 01 August 2015]; Available from: <http://www.reconinstruments.com/products/jet/>.
- [3] Walker, A. and S. Brewster, *Spatial Audio In Small Screen Device Displays*. Personal Technologies, 2000. 4(2-3): p. 144-154.
- [4] WHO, *Mobile Phone Use: A growing Problem of Driver Distraction*, in *Decade of Action for Road Safety*. 2011, World Health Organization.

- [5] CDC. *Injury Prevention & Control: Motor Vehicle Safety*. 2015 [cited 2016 3rd Feb]; Available from: <http://www.cdc.gov/motorvehiclesafety/distracted-driving/>.
- [6] DOT. *Distraction: Facts and Statistics*. 2015 [cited 2016 3rd Feb]; Available from: <http://www.distraction.gov/stats-research-laws/facts-and-statistics.html>.
- [7] Wightman, F.L. and D.J. Kistler, *Headphone Simulation Of Free Field Listening. II: Psychophysical Validation*. The Journal of the Acoustical Society of America, 1989. **85**(2): p. 868 - 878.
- [8] Burgess, D.A. *Techniques For Low Cost Spatial Audio*. in *5th Annual ACM Symposium On User Interface Software And Technology*. 1992. Monterey, California, United States: ACM.
- [9] Wenzel, E.M., et al., *Localization using nonindividualized head - related transfer functions*. The Journal of the Acoustical Society of America, 1993. **94**(1): p. 111-123.
- [10] Sawhney, N. and C. Schmandt. *Nomadic Radio: Scaleable and Contextual Notification For Wearable Audio Messaging*. in *SIGCHI conference on Human Factors in Computing Systems*. 1999. ACM.
- [11] Cheng, C.I. and G.H. Wakefield, *Introduction to Head Related Transfer Functions: Representations of HRTFs in Time, Frequency and Space*. Journal of the AES, 2001. **49**(4): p. 231 - 249.
- [12] Wilkins, P.A. and W.I. Acton, *Noise and accidents—a review*. Annals of Occupational Hygiene, 1982. **25**(3): p. 249 - 260.
- [13] Lichenstein, R., et al., *Headphone Use and Pedestrian Injury and Death in the United States: 2004 - 2011*. Injury Prevention, 2012.
- [14] Harma, A., et al., *Augmented Reality Audio For Mobile And Wearable Appliances*. Journal of the Audio Engineering Society, 2004. **52**(6): p. 618-139.
- [15] Tikander, M., M. Karjalainen, and V. Riikonen. *An Augmented Reality Audio Headset*. in *11th Int. Conference on Digital Audio Effects (DAFx-08)*. 2008. Espoo, Finland.
- [16] Labs, D. *Here Active Listening*. 2015 [cited 2016 6 Feb]; Available from: <https://www.hereplus.me/>.
- [17] Oldfield, S.R. and S.P.A. Parker, *Acuity of Sound Localization: A Topography of Auditory Space. II. Pinna Cues Absent*. Perception, 1984. **13**: p. 601 - 617.
- [18] Walker, B.N. and J. Lindsay, *Navigation Performance In A Virtual Environment With Bonephones*, in *International Conference On Auditory Displays*. 2005: Limerick, Ireland.
- [19] MacDonald, J.A., P.P. Henry, and T.R. Letowski, *Spatial Audio Through A Bone Conduction Interface*. International Journal Of Audiology, 2006. **45**(10): p. 595 - 599.
- [20] Walker, B.N. and J. Lindsay, *Navigation Performance With A Virtual Auditory Display: Effects Of Beacon Sound, Capture Radius and Practice*. Human Factors: The Journal of the Human Factors and Ergonomics Society, 2006. **48**(2): p. 265 - 278.
- [21] Lindeman, R.W., H. Noma, and P.G.d. Barros, *Hear-Through and Mic-Through Augmented Reality: Using Bone Conduction To Display Spatialized Audio*. 2007.
- [22] Stanley, R.M., *Measurement And Validation Of Bone Conduction Adjustment Functions In Virtual 3D Audio Displays*, in *School of Psychology*. 2009, Georgia Institute of Technology.
- [23] Valjamae, A., et al., *Binaural Bone Conducted Sound In Virtual Environments: Evaluation Of A Portable, Multimodal Motion Simulator Prototype*. Acoustical Science And Technology, 2008. **29**(2): p. 149 - 155.
- [24] Villegas, J. and M. Cohen, *GABRIEL: Geo-Aware Broadcasting For In-Vehicle Entertainment And Localizability*, in *AES 40th International Conference*. 2010, AES: Tokyo, Japan.
- [25] Perrott, D.R., et al., *Aurally Aided Visual Search in the Central Visual Field: Effects of Visual Load and Visual Enhancement of the Target*. Human Factors, 1991. **33**(4): p. 389 - 400.
- [26] Nelson, W.T., et al., *Effects of Localized Auditory Information on Visual Target Detection Performance Using a Helmet-Mounted Display*. Human Factors, 1998. **40**(3): p. 452 - 460.
- [27] Rudmann, D.S. and T.Z. Strybel, *Auditory Spatial Facilitation of Visual Search Performance: Effects of Cue Precision and Distractor Density*. Human Factors, 1999. **41**(1): p. 146 - 160.
- [28] ART, *DTrack*. 1999, Advanced Realtime Tracking: Munich, Germany.
- [29] ART. *ARTTRACK2*. 1999 [cited 2015 1st Dec]; Available from: <http://www.arttracking.com/products/discontinued/arttrack2/>.
- [30] Technologies, S. *Steradian S7-X Laser Tag Gun*. 2003 [cited 2015 12 Sept]; Available from: <http://www.steradiantech.com/xseries/s7x/>.
- [31] II, R.M.T., *Virtual Reality Peripheral Network (VRPN)*. 1998.
- [32] II, R.M.T., et al. *VRPN: A Device-Independent, Network-Transparent VR Peripheral System*. in

Proceedings of the ACM Symposium on Virtual Reality Software and Technology. 2001. ACM.

- [33] Aftershokz. *Sportz3*. [cited 2015 2nd June]; Available from: <http://aftershokz.com/collections/wired/products/sportz-3>.
- [34] Walker, B.N. and G. Kramer, *Auditory Displays, Alarms and Auditory Interfaces*, in *International Encyclopedia of Ergonomics and Human Factors*, W.K. Informa Healthcare, Editor. 2006, CRC Press. p. 1021 - 1025.
- [35] Walker, B.N. and R.M. Stanley, *Thresholds Of Audibility For Bone Conduction Headsets*, in *International Conference On Auditory Display*. 2005, ICAD: Limerick, Ireland.
- [36] Gardner, W.G., *3D Audio Using Loudspeakers*, in *School of Architecture and Planning*. 1997, Massachusetts Institute of Technology.
- [37] Wersenyi, G., *Localization In A HRTF Based Minimum Audible Angle Listening Test On a 2D Sound Screen For GUIB Applications*, in *115th Convention of the Audio Engineering Society* 2003.
- [38] Stevens, S.S. and E.B. Newman, *The Localization of Actual Sources of Sound*. *The American Journal of Psychology*, 1936. **48**(2): p. 297 - 306.
- [39] Weinrich, S.G., *Horizontal Plane Localization Ability and Response Time as a Function of Signal Bandwidth*, in *98th Convention of the AES*. 1995: Paris, France.
- [40] Walker, B.N., R.M. Stanley, and J. Lindsay. *Task, User Characteristics, and Environment Interact to Affect Mobile Audio Design*. in *PERVASIVE*. 2005.
- [41] Helgason, D., N. Francis, and J. Ante, *Unity3d*. 2005, Unity Technologies.
- [42] Thakur, A. and V. Nair, *3Deception*. 2015, Two Big Ears: Edinburg, Scotland.

RESPONSE TECHNIQUES AND AUDITORY LOCALIZATION ACCURACY

Nandini Iyer, Eric R. Thompson, Brian D. Simpson

Air Force Research Laboratory, 711 Human Performance Wing
2610 Seventh St., B441 Wright-Patterson AFB, OH 45433 USA
nandini.iyer.2@us.af.mil

ABSTRACT

Auditory cues, when coupled with visual objects, have lead to reduced response times in visual search tasks, suggesting that adding auditory information can potentially aid Air Force operators in complex scenarios. These benefits are substantial when the spatial transformations that one has to make are relatively simple i.e., mapping a 3-D auditory space to a 3-D visual scene. The current study focused on listeners' abilities to map sound surrounding a listener to a 2-D visual space, by measuring performance in localization tasks that required the following responses: 1) Head pointing: turn and face a loudspeaker from where a sound emanated, 2) Tablet: point to an icon representing a loudspeaker displayed in an array on a 2-D GUI or, 3) Hybrid: turn and face the loudspeaker from where a sound emanated and then indicate that location on a 2-D GUI. Results indicated that listeners' localization errors were small when the response modality was head-pointing, and localization errors roughly doubled when they were asked to make a complex transformation of auditory-visual space (i.e., while using a hybrid response); surprisingly, the hybrid response technique reduced errors compared to the tablet response conditions. These results have large implications for the design of auditory displays that require listeners to make complex, non-intuitive transformations of auditory-visual space.

1. INTRODUCTION

In the Air Force, operators are routinely required to make complex decisions with an incredible barrage of information, often under severe cognitive load (multitasking, multiple sources to monitor, interoperability issues, etc.). Most information is presented using visual interfaces, and due to increasing complexity of operations, there is a high likelihood that operators might miss critical information if these events occur outside of the locus of visual attention. In these complex operations, the consequences of missing critical information could be greatly reduced by introducing multimodal displays and presenting some of the information through the auditory modality. The auditory modality has the advantage of responding to sounds arriving from anywhere in the environment; thus, while the spatial resolution of the auditory system is coarse relative to visual spatial resolution, its coverage is greater (360 degrees), reducing the possibility that events occurring outside the field of view will go undetected. Auditory cues can also effectively increase awareness of one's surroundings, convey a variety

of complex information without taxing the visual system, increase the sense of presence in immersive environments, cue visual attention, and facilitate cross-modal enhancements.

One of the early obstacles to using auditory cues in operational environments was the feasibility and cost of recreating auditory space over headphones. Using signal-processing techniques, it is now relatively easy to generate stereophonic signals under headphones that recreate the spatial cues available in the real-world; in fact, when these cues are rendered appropriately, it is often difficult to distinguish between sounds presented in the free-field over loudspeakers from those presented virtually over headphones [1]. When coupled with a head-tracker, realistic acoustic environments that respond naturally to dynamic source and head motion can be rendered in 3-dimensional (3-D) space around a person's head. Such displays are not only effective and compelling, but are now sufficiently mature for integration into systems employed in operational environments.

Several studies have now shown that auditory cues can speed responses to targets in a visual search task, i.e., when auditory cues are co-located with visual targets, search times for a visual target are reduced significantly compared to the same search conducted without an auditory cue. Further, some of these studies demonstrated that an auditory cue reduced the deleterious effects on response time that typically occur with increases in visual scene complexity [2, 3]. In all of these experiments, accuracy was fairly high by design and only response time differences were measured and reported. While response time is an important metric in real-world operational environments, research studies are also interested in the question of localization accuracy. In determining localization accuracy, a number of different response techniques have been used; a relatively natural response method requires listeners to turn to the stimulus location and localization responses are obtained by tracking the listener's head in space (head-pointing, finger-pointing or nose-pointing response methods) [4]. Other natural response methods have included using a pistol-like device to "shoot" at the target location [5]. While head-, finger- or nose-pointing responses are most accurate [6], by and large, these other response methods yield comparably similar localization responses; in part, it is due to the fact that these localization responses do not require any mental transformations of the target location and the listeners can use their own anatomical references to localize a sound.

In contrast to more direct or natural localization responses, indirect localization response methods have also been used, such as a verbal reporting technique [7], God's eye localization pointing (GELP) [6], or the large- and small-head response techniques [8]. In verbal response methods, listeners had to be trained to state the perceived azimuth and elevation on sources that were then transcribed by an experimenter. In the GELP method, listeners were



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>



Figure 1: Picture of the facility with the circular speaker array. A listener is shown seated in the center of the array with the tablet and head tracker.

required to make their responses of perceived location on a 22 cm plastic sphere using an electromagnetic sensor. In the large- and small-head response techniques, the sphere in GELP was replaced by an anatomically correct styrofoam head (large or small); the head was placed in front of the subject facing their right side for easy accessibility. All of these procedures were less accurate than direct localization responses, suggesting that some response techniques yield better localization accuracy than others. There are several reasons for decreased accuracy of localization responses with indirect methods; for example, the verbal response technique required correct interpretation and entry by an experimenter which could be a source of error. The remaining three response techniques require listeners to remember target sources relative to their own head and then transform that response onto another representation of possible source location (either a sphere or a styrofoam head). In summary, indirect localization responses are presumably less intuitive than direct pointing techniques.

In future AF operations, we can foresee several areas where operators can utilize natural, intuitive localization judgments to acquire a visual target or avoid a threat; for example, head-mounted displays (HMDs) can be head-tracked to display coupled visual and auditory information allowing operators to access data that are tied to line-of-sight. However, in order for audio-aided cueing to be effective in operational environments, operators might have to make more complex transformations of an auditory cue to an associated visual space; for example, in intelligence, surveillance and reconnaissance (ISR) missions, operators are tasked with trying to find, fix and track a target when presented with a God's eye view of visual targets on the ground. In such scenarios, it is not unreasonable to expect that an audio cue might assist operators to locate a target and track a moving target. However, it is not clear if operators could make the spatial transformations required to best utilize such a cue. That is, can an observer benefit from a 3-D spatial audio cue generated and presented virtually to reduce the acquisition time of a visual object presented on a 2-D display located on a monitor/screen in front of the operator. The current experiment was designed to assess if localization accuracy and response time would vary as a function of the response technique employed in the task. Three response techniques were evaluated: head-pointing, tablet response (with possible target locations displayed using a GUI) and a hybrid method that incorporated head-turning to localize the sound on a tablet. The first technique is very intuitive and

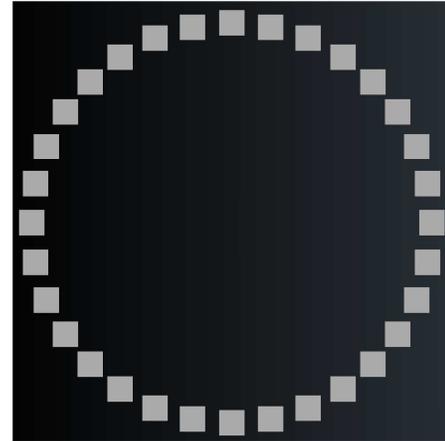


Figure 2: Screenshot of the tablet Graphical User Interface (GUI) used with the tablet and hybrid methods.

natural where a listener has to turn his/her head towards the direction of a sound. The second technique requires listeners to perform a more complex transformation of space; he/she has to associate a sound with an object representing the loudspeakers displayed on a tablet. The third technique was employed to evaluate whether or not turning and acquiring a source might facilitate accuracy with a tablet response.

Another factor that is important while evaluating response techniques is the actual stimulus itself. It is known that high frequency energy (above 8 kHz) is important for accurate auditory localization of non-speech signals [9, 10, 11, 12]. The few studies that have examined speech localization in the horizontal plane using single words presented found no significant differences in localization accuracy with speech and non-speech broadband stimuli; however, they reported an increase frontback confusions with speech stimuli [6]. In the current study, both non-speech broadband stimuli as well as speech were used as stimuli to evaluate whether the two response techniques might have differential effects on the two types of stimuli.

2. METHOD

2.1. Participants

Eleven listeners (six female) participated in the experiment. All had normal audiometric thresholds (<20 dB HL at octave frequencies between 250 and 8000 Hz) and normal or corrected-to-normal vision, and all had prior experience in sound localization studies. They were all paid to participate in the study, and all provided informed consent under a protocol approved by the Air Force Research Lab, 711th HPW Institutional Review Board.

2.2. Stimuli

On each trial, the listeners heard either a burst of noise, or a single word, which were both presented at a sampling rate of 44.1 kHz. The noise bursts had a pink spectrum between 100 Hz and 20 kHz, and a duration of 250 ms. The words were from recordings made in our lab of the PB-50 word lists [13]. The word lists were recorded

with a sampling rate of 44.1 kHz by twelve talkers (six female), none of whom were participants in this experiment. The recordings were post-processed to normalize the duration of each utterance to 500 ms using the Praat software [14].

2.3. Equipment

The experiment was controlled by a PC running the Windows 7 operating system and Matlab (the Mathworks, R2013a). The stimuli were generated in Matlab, and were presented through a RME RayDAT interface, RME M-32DA converter, eight Crown CTs 4200 four-channel amplifiers, and 32 Orb Audio loudspeakers in an evenly spaced circular array, approximately 11.25 degrees apart in azimuth (see Fig. 1). The listener was seated at the center of the array with their ears on the same plane as the loudspeaker array. An Optitrack V120:Trio was used with a custom array of reflecting balls mounted on a headband for tracking head motion. A cluster of four LEDs was mounted in front of and centered on each loudspeaker. The LED array was controlled by PacLED64 boards (Ultimarc). On head-tracked response trials, input was made using a Nintendo Wii remote, with communication to Matlab provided by WiiLab [15]. The tablet interface was a custom app running on a Google Nexus 9 tablet.

2.4. Procedure

Within a block of 32 trials, the stimulus condition (noise or speech) was held constant, and the source location was pseudo-randomly selected so that each listener completed 15 trials per source location per response method. On each trial, the listeners oriented toward a reference location (defined as 0 degree azimuth), heard a stimulus, made a response according to the response input method for that block, and received feedback for the correct location of the source. The details for each response method follow. Each listener completed all of the blocks for one response method before continuing to the next method. The head tracking and tablet response methods were completed first and with six listeners completing the head tracking method first and the remaining completing the tablet response first. Both groups (tablet first or head tracking first) completed the hybrid response as the third condition.

2.4.1. Head pointing method

At the start of each trial, the listener was required to orient towards a reference loudspeaker. The LED cluster closest to their head orientation would illuminate as their head turned. When the LED cluster at the reference location was illuminated, they pressed a button on the Wii remote to start the stimulus presentation. After the end of the stimulus, the head-slaved LED cursor would again illuminate, and the listeners would turn to look in the perceived direction of the sound. When the LED cluster was illuminated on the desired response location, they again pressed a button on the Wii remote to enter their response.

2.4.2. Tablet method

The listeners remained oriented toward the reference loudspeaker for the whole experiment. The tablet showed a circular array of 32 buttons that corresponded to the 32 loudspeaker locations (see Fig. 2). The reference location was at the top of the tablet display. In each trial, they heard the stimulus, and then indicated on the

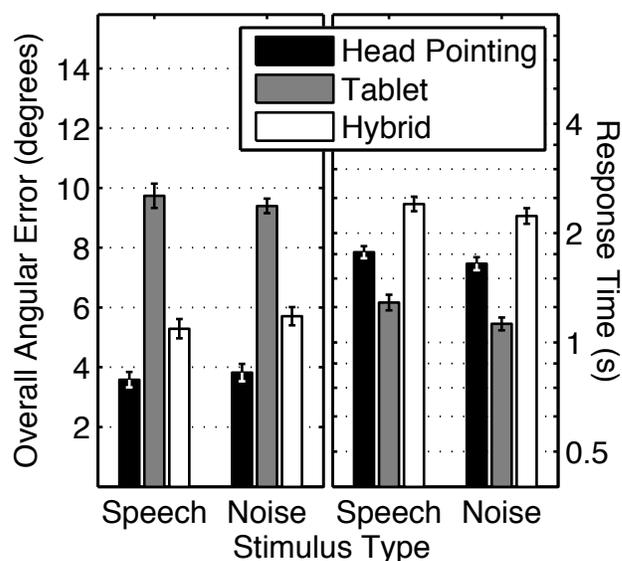


Figure 3: Average unsigned localization error (left panel) and mean response time (right panel: note logarithmic response times) plotted as a function of the two stimuli types, speech and noise. The parameters in the figures represent the three different response techniques used: head pointing (black bars), tablet response (gray bars) and hybrid response (head-pointing then responding using the tablet: white bars). Error bars represent standard error for within-subject measures.

tablet the location from which they perceived the sound to originate. The correct source location was indicated after their response by flashing the background color of the corresponding button location on the GUI.

2.4.3. Hybrid method

The hybrid response method contained elements of the head tracking and tablet methods. The listeners had to orient toward the reference loudspeaker before each trial and press a button on the tablet to begin stimulus presentation. After the stimulus ended, they were instructed to turn their head and look in the perceived direction of the sound source. After identifying the sound source location, they were instructed to return to the reference position and select the response on the tablet (the tablet display did not rotate with the head orientation). As with the tablet method, correct location feedback was provided by flashing the background color of the corresponding button.

3. RESULTS AND DISCUSSION

The data from the experiment are plotted in Figure 3; the left panel depicts average angular error as a function of the stimuli used in the experiment, for the three types of response techniques. As is apparent in the figure, there appears to be no difference in localization accuracy for the two different types of stimuli (Speech and Noise) across response techniques used. The response techniques do, however, influence localization accuracy. Specifically, local-

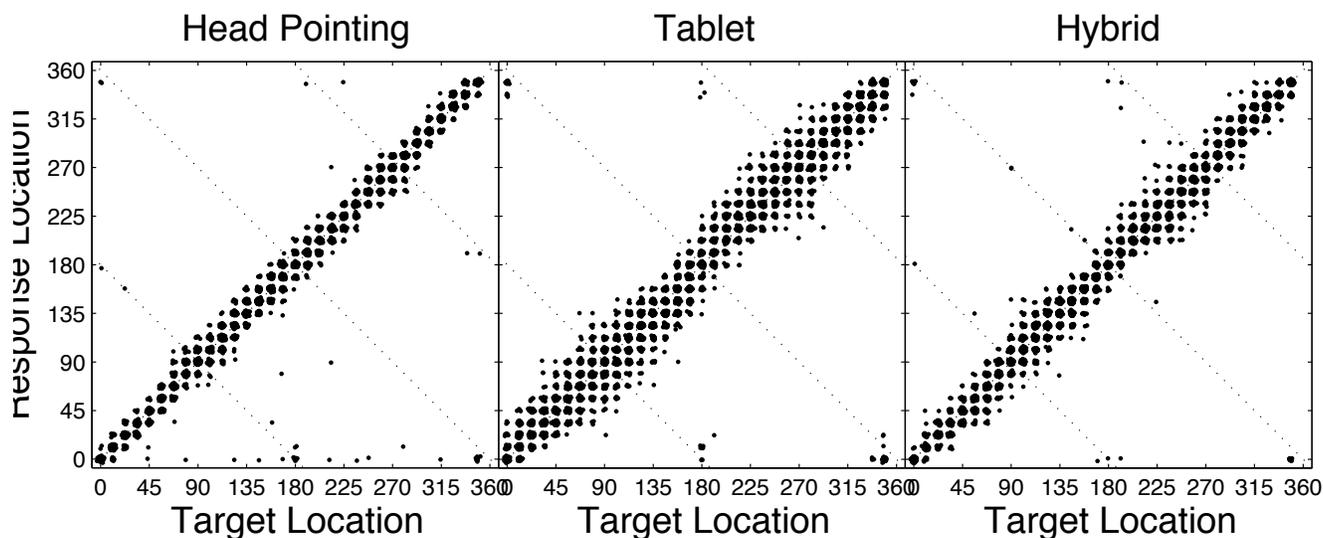


Figure 4: Scatter plot of response location vs. source location for the three response methods. Each data point in the plot is jittered slightly so that more frequently used target/response combinations appear with a larger black patch. The dashed lines with a negative slope in each plot indicate where front-back and left-right reversals would lie.

ization accuracy was best (about 4 deg.) for the head-pointing response technique and increased by at least a factor of two (to about 9.7 deg.) with the tablet response technique. However, when they first turned to look at a loudspeaker before responding on the tablet, errors were smaller when compared to the tablet response technique. The right panel in Figure 3 depicts the geometric mean response times for the three different response techniques in the experiment. As is evident from the figure, listeners were fastest using the tablet response (approx. 1.5 sec on average), slowest using the hybrid technique (approx. 2.3 sec) and took approximately 1.75 sec on average to respond using head-pointing. This was true for both Speech stimuli and Noise stimuli. It is perhaps not surprising that response times were longer using the head-pointing technique, since listeners had to rotate in their chairs and look at a loudspeaker, whereas in the tablet response condition, they responded by clicking on a GUI display of the loudspeaker array. Interestingly, response times were faster in the hybrid condition than the simple sum of response times for the head-pointing and tablet response techniques, especially since they had to reorient towards the reference speaker before they responded; it is possible that the reduction in response time occurred due to reduced transformations listeners made in the hybrid condition compared to the tablet condition. Another possibility is that listeners were generally more efficient at associating a visual target to a source, presumably because the head-pointing response relies on explicit memory of the source's visual location, and the mental transformation onto a tablet response was easier in this condition.

Figure 4 depicts performance for the three response techniques as a scatter plot, where target locations are plotted on the abscissa and response locations are plotted along the ordinate. In these plots, correct responses would fall on the diagonal from lower-left to upper-right. As seen in the figure, head pointing response techniques resulted in a tight distribution with most of the responses falling along the diagonal. In contrast, the response distributions

for the tablet techniques are more scattered along the diagonal. The off-diagonal errors were mostly front-back errors and they were more common for speech, than for the noise stimuli. The increased front-back confusions for speech stimuli have also been reported in other studies, both with real and virtual stimuli. Responses using the hybrid technique were less variable compared to the tablet response technique, but head-pointing localization technique outperformed the two other techniques.

At the outset, we were interested in assessing whether the order in which listeners experienced the head-pointing or tablet response technique influenced localization accuracy in the experimental conditions. The data describing the influence of order are depicted in Figure 5 where average localization errors are depicted as a function of the two types of stimuli used. The left panel depicts data for listeners who responded to Speech or Noise stimuli using the tablet response first followed by the head-pointing, whereas the right panel depicts errors for listeners who responded to the two types of stimuli using head-pointing first followed by tablet responses. A three-way analysis of variance (ANOVA), with one between-subject factor (order of response technique) and 2 within-subject factors (type of stimuli and response technique) showed a significant main effect for response techniques used ($F(2,18)=81.6, p<0.001$). No other main effects or interactions were significant. Thus, it appears that exposing listeners to what should be a more direct response technique first did not afford them any advantage over a group who experienced a more indirect response method first. It is perhaps surprising that the hybrid response technique resulted in lower localization errors compared to the tablet response, because in both cases, the response was the same. Somehow, turning and pointing to a location in space allowed listeners to make the necessary spatial transformation and associate a visual object with a sound with more accuracy. However, it was still not as accurate as the head-pointing technique. Listeners were required to reorient to the boresight speaker be-

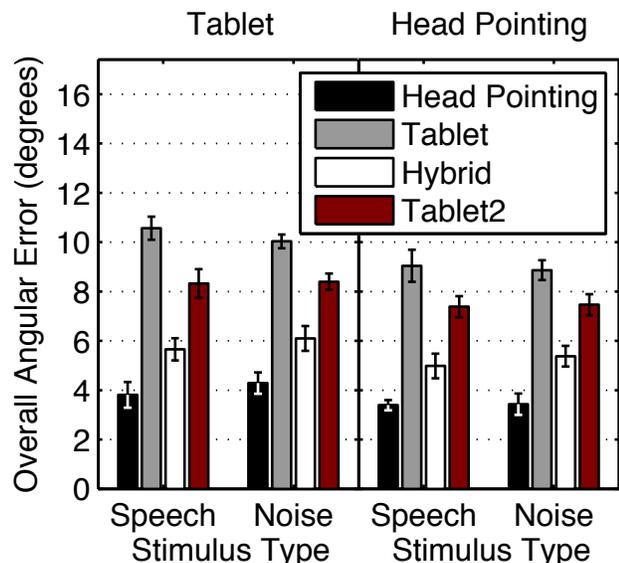


Figure 5: Average localization error plotted as a function of the two stimuli types, speech and noise. The parameters in the figures represent the three different response techniques used, as in Fig. 3. The left panel is for listeners who ran the tablet response condition first, and the right panel is for those who ran the head pointing condition first.

fore responding, which could have resulted in less accurate performance. The hybrid condition was also run as the last condition; it is possible that listeners had gained some familiarity with the tablet response technique so that better performance in the hybrid condition was merely a reflection of habituation with a response technique. In order to assess whether listeners improved in the hybrid condition due to repeated exposure/practice with the tablet response, ten of the eleven listeners from the study were re-run in the tablet response condition for a second time (the eleventh listener was unavailable). The mean accuracy and response times in this condition is plotted in Figure 6, along with data from the first tablet response condition and the hybrid condition. As shown in the figure, accuracy in the second tablet response condition reduced by approximately 2 degrees, and was significantly better than the first tablet response condition ($F(2,18)=81.6, p<0.001$), but still differed from the hybrid condition. However, listeners did not differ in response times between the two tablet response conditions. Thus, repeated exposure to tablet response technique facilitated learning, but the improvement did not explain all of the performance advantage observed with the hybrid response technique.

The data from the current study suggest performance degrades in listening conditions where operators are required to make a mental transformation from a sound emanating in space to an associated aerial view of the objects creating that sound (tablet response technique). The poor performance is reflected in decreased localization accuracy for the tablet response condition. When the auditory source and its corresponding visual object are co-located, listeners can easily identify the source by turning their head towards the auditory source and the visual object simultaneously

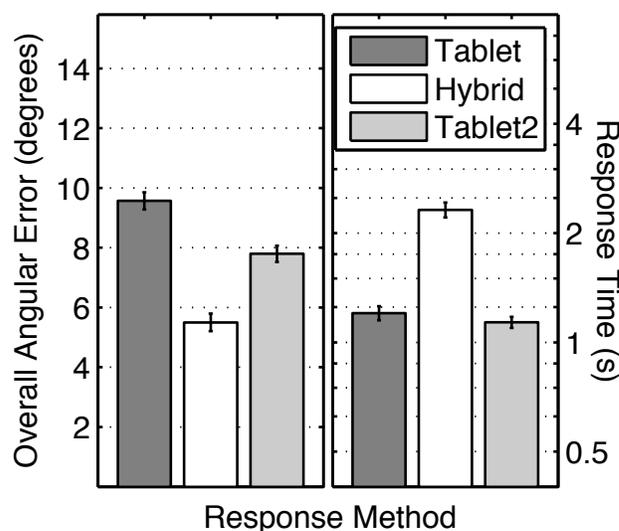


Figure 6: Average localization error (left panel) and mean response times (right panel) plotted for three response techniques, collapsed across the two stimuli type (Speech and Noise). The tablet and hybrid conditions in the figure are replotted from Fig. reffig:angular (dark grey and white bars respectively); the light grey bar depicts average error and mean response times for ten listeners who were rerun in the tablet response condition (denoted as Tablet2 in the figure). Error bars represent standard error for within-subject measures.

(head-pointing technique). Allowing listeners to use a hybrid response technique, during which they can first turn and look at the loudspeaker and then select the appropriate one on the tablet seems to improve their localization accuracy, albeit never to the same accuracy as the head-pointing response technique condition, and at the cost of increased response times. These results suggest that associating an auditory stimuli with a visual object that do not co-exist may not afford any benefits in visual search tasks. It is possible that listeners can learn to associate sound to visual objects if these sounds were perceptually different (i.e., each sound paired with a specific visual object), or if they are trained in the task.

In this task, we measured localization accuracy using three response techniques. We postulate that three possible sources can limit the performance in the task: target locations (all possible locations from which the a target sound can arise), human sensory limitation (listeners ability to localize sounds in space), response technique. From our data, it is clear that head pointing technique resulted in the best accuracy. We argue that for the target locations tested (32 loudspeakers distributed in 360 degrees azimuth), the data obtained with the head pointing technique reflects the lower bound of performance limitations; i.e., the best performance in the task given the perceptual limitations. It is possible that there might be some response transformations that listeners are required to do even in this task; nevertheless, such a transformation is well-practiced, and agrees with accuracy estimates obtained in our lab using a larger density of loudspeaker locations ([1]). Using a tablet response limited performance mainly due to response technique limitations, suggesting that GUIs used to elicit

localization responses should be used cautiously. From an auditory displays perspective, if an auditory stimulus is used to cue a location in space that is aurally displayed, our data suggests that 32 locations might be too dense of a visual field to cue. It is possible that two of the sources of performance limitations can trade-off, so that limitations due to response techniques can be offset by reducing the target location density. Therefore, GUI responses might be perfectly adequate for a more limited distribution of source locations. Further research is needed to validate our claim. Our experiment also suggested that localization accuracy was better when tablet responses were reintroduced for a second time, suggesting that exposure to a response technique could improve performance. Additional research is needed to assess whether systematic training can improve response accuracy in localization tasks using more indirect response methods.

4. CONCLUSIONS

An accumulating body of evidence has shown that auditory cues play a crucial role in everyday life. Spatial auditory information can provide invaluable information to an operator, particularly when the visual channel is saturated [16]. Response times to visual targets associated with localized auditory cues have been shown to decrease [2] relative to those without auditory cues; however, in all these studies, the responses were intuitive and natural i.e., turning to look in the direction from which the sound emanated. In some Air Force operations, the transformation of auditory space to a visual object might not be straightforward, and it might require some mental transformations. The current study attempted to assess the variability associated with response techniques in a simple localization task. The findings suggest that localization errors almost double when listeners have to indicate the location from where a sound emanated using a tablet response technique, compared to a more natural head-pointing localization response. Some of the deleterious effect of making a transformation could be eliminated by allowing listeners to orient their head towards the speaker and then responding using a tablet (the hybrid response technique). Exposure to more natural response techniques did not allow listeners to perform better in conditions requiring some mental transformations. The results of the study suggest that designing auditory displays for military operations where there is not a simple 1:1 matching of the spatial locations of visual and auditory stimuli might not be particularly useful and might require additional training.

5. REFERENCES

- [1] G. D. Romigh, D. S. Brungart, and B. D. Simpson, "Free-field localization performance with a head-tracked virtual auditory display," *IEEE J. Selected Topics in Sign. Proc.*, vol. 9, no. 5, pp. 943–954, 2015.
- [2] D. R. Perrott, T. Sadralodabai, K. Saberi, and T. Z. Strybel, "Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target," *Human Factors*, vol. 33, no. 4, pp. 389–400, 1991.
- [3] R. S. Bolia, W. R. D'Angelo, and R. L. McKinley, "Aurally aided visual search in three-dimensional space," *Human factors*, vol. 41, no. 4, pp. 664–669, 1999.
- [4] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2188–2200, 1990.
- [5] S. R. Oldfield and S. P. A. Parker, "Acuity of sound localisation: A topography of auditory space. I. Normal hearing conditions," *Perception*, vol. 13, no. 5, pp. 581–600, 1984.
- [6] R. H. Gilkey, M. D. Good, M. A. Ericson, J. Brinkman, and J. M. Stewart, "A pointing technique for rapidly collecting localization responses in auditory research," *Behav. Res. Meth. Instr. Comp.*, vol. 27, no. 1, pp. 1–11, 1995.
- [7] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661, 1992.
- [8] D. S. Brungart, W. M. Rabinowitz, and N. I. Durlach, "Evaluation of response methods for the localization of nearby objects," *Percept. & Psychophys.*, vol. 62, no. 1, pp. 48–65, 2000.
- [9] A. W. Bronkhorst, "Localization of real and virtual sound sources," *J. Acoust. Soc. Am.*, vol. 98, no. 5, p. 2542, 1995.
- [10] R. B. King and S. R. Oldfield, "The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays," *Human Factors*, vol. 39, no. 2, pp. 287–295, 1997.
- [11] S. Carlile, S. Delaney, and A. Corderoy, "The localisation of spectrally restricted sounds by human listeners," *Hearing Res.*, vol. 128, no. 1-2, pp. 175–189, 1999.
- [12] A. van Schaik, C. Jin, and S. Carlile, "Human localisation of band-pass filtered noise," *Int. J. Neural Syst.*, vol. 9, no. 5, pp. 441–446, 1999.
- [13] J. P. Egan, "Articulation testing methods," *Laryngoscope*, vol. 58, no. 9, pp. 955–991, 1948.
- [14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," <http://www.praat.org>, 2005.
- [15] J. Brindza, J. Szweda, and A. Striegel, "Wiilab," <http://netscale.cse.nd.edu/twiki/bin/view/Edu/WiiMote>, 2013.
- [16] D. R. Begault and E. M. Wenzel, "Headphone localization of speech," *Human Factors*, vol. 35, no. 2, pp. 361–376, 1993.

IN EAR TO OUT THERE: A MAGNITUDE BASED PARAMETERIZATION SCHEME FOR SOUND SOURCE EXTERNALIZATION

Griffin D. Romigh, Brian D. Simpson, Nandini Iyer

711th Human Performance Wing
Air Force Research Laboratory
2610 7th street, WPAFB, OH 45344, USA
griffin.romigh@us.af.mil

ABSTRACT

While several potential auditory cues responsible for sound source externalization have been identified, less work has gone into providing a simple and robust way of manipulating perceived externalization. The current work describes a simple approach for parametrically modifying individualized head-related transfer function spectra that results in a systematic change in the perceived externalization of a sound source. Methods and results from a subjective evaluation validating the technique are presented, and further discussion relates the current method to previously identified cues for auditory distance perception.

1. INTRODUCTION

Most spatial auditory displays are built around the understanding that, by replicating the natural cues listeners use to determine a sound source's location, any single-channel sound source can be imbued with spatial attributes. These cues are the complex function of space, frequency, and individual listener known as a Head-Related Transfer Function (HRTF). HRTFs capture the acoustic transformation a sound undergoes as it travels from a specific location in space, interacts with a listener's head, shoulders, and outer ears, and arrives separately at the two eardrums [1]. A single-channel sound filtered with the right-ear and left-ear HRTF corresponding to a specific location can then be presented over headphones with directional accuracy and fidelity comparable to a free-field source provided the HRTF measurements were individualized to the listener [2, 3, 4].

Unfortunately, due to logistical issues associated with acquiring individualized HRTF measurements, most spatial displays utilize non-individualized HRTFs or subsets of the cues contained therein (e.g., only gross binaural cues), resulting in less perceptually accurate spatial representations. A frequent bane for headphone-based displays is the problem of poor sound source externalization. Sometimes referred to as "inside-the-head-locatedness" [5], poor externalization (or internalization) is the perceptual phenomena where sound sources are perceived to originate from a location within a listener's head rather than from out in space where the display designer had intended. Blauert [5] was one of the first to thoroughly review the literature associated with

externalization and suggested that the lack of externalization was merely an endpoint on the continuum of perceived auditory distance (i.e. a sound source appears so close to you that it is perceived inside your head).

The theory that perceived externalization is part of auditory distance perception was backed up by the experimental evidence of Hartmann and Wittenberg [6]. They showed that perceived externalization could be manipulated systematically for a harmonic tone complex by zeroing out inter-aural phase differences (IPD) for tonal components above a given frequency; the lower the critical frequency the less externalized the sound source was judged to be up to a cutoff frequency near 1kHz. Hartmann and Wittenberg [6] also showed that externalization was not affected by forcing a single frequency independent interaural timing difference (ITD) cue, and that both monaural magnitude spectra need to be preserved across the entire frequency range to ensure good externalization not just the gross interaural level difference (ILD). Proper externalization (and/or distance perception) has also been linked to a number of other factors including the use of dynamic head-motion cues [7], ratio of direct to reverberant energy [8], and the high frequency roll off [9].

Despite the previous efforts, it is not clear what the role of spectral features are in perceived externalization. At the extremes, there is a very clear relationship between the presence of monaural spectral features and externalization, such that an absence of spectral features when implementing only a frequency independent ILD causes poor externalization, while the full representation of the spectral features when implementing an individualized HRTF produces good externalization. It is less clear however what is perceived with compressed spectra, where the narrowband spectral cues are present yet potentially diminished in magnitude. Based on that question and the desire for a simple parameterization with which externalization could be effectively modulated, the current investigation aimed at determining whether externalization could be reliably controlled through simple modifications to the monaural magnitude spectrum contained in an HRTF.

2. PARAMETERIZATION

From previous literature it is clear that the two extremes of externalization can be attained using methods that differ only in spectral representation of their spatial filters. On one hand, if spatial information is imparted on a sound source using only binaural cues (ILD and ITD), the sound source will appear as though it originates from somewhere along the interaural axis inside the listener's head [10]. On the other hand, well localized *and* externalized sound



This work is licensed under Creative Commons Attribution Non-Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

sources can be created virtually if the source is rendered with an individualized HRTF that preserves the ITD and both left-ear and right-ear monaural spectral cues [11]. These two methods of spatial presentation differ only in the way the spectrum of the sound source at each ear is modified. The parameterization detailed in this section provides a straightforward method to linearly interpolate between the spectra that are known to result in good externalization and spectra which are known to result in strong internalization.

Starting from an individualized HRTF measurement, the following method systematically varies the prominence of spectral features of the left-ear log-power spectrum; an identical procedure is used to modify the right-ear spectrum. If we define $\mathcal{H}_{\phi,\theta}^L[k]$ to be the individualized left-ear log-power spectrum (i.e. decibel scale) corresponding to a location with azimuth $-180^\circ \leq \phi < 180^\circ$ and elevation $-90^\circ \leq \theta < 90^\circ$, the location-specific, frequency-independent average monaural level $\mathcal{A}_{\phi,\theta}^L$ is defined as in Eq. 1.

$$\mathcal{A}_{\phi,\theta}^L = \frac{1}{K} \sum_k \mathcal{H}_{\phi,\theta}^L[k] \quad (1)$$

Here, k is used to represent one of K discrete frequency values in the valid positive frequency range for the HRTF. For the present work $2 \leq k \leq 174$, making $K = 173$, which represents the positive frequencies from approximately 200 Hz to 15 kHz for a 512 length DFT at a sampling rate of 44.1 kHz. This average level is subtracted from the measured HRTF to give the left spectrum $S_{\phi,\theta}^L[k]$ as in Eq. 2, which contains all of the spectral features.

$$S_{\phi,\theta}^L[k] = \mathcal{H}_{\phi,\theta}^L[k] - \mathcal{A}_{\phi,\theta}^L \quad (2)$$

These two components are then weighted and recombined to give the transformed spectrum $\tilde{\mathcal{H}}_{\phi,\theta}^L[k]$ as in Eq. 3

$$\tilde{\mathcal{H}}_{\phi,\theta}^L[k] = \frac{\alpha}{100} S_{\phi,\theta}^L[k] + \mathcal{A}_{\phi,\theta}^L \quad (3)$$

The parameter α in Eq. 3 varies the magnitude of the spectral features contained in the final transformed spectrum $\tilde{\mathcal{H}}_{\phi,\theta}^L[k]$ from full scale for $\alpha = 100$ to nil for $\alpha = 0$. The transformed spectra corresponding to several levels of the α parameter are shown in Fig. 1 for the left and right ears of a representative subject at two locations on the horizontal plane.

3. METHODS

3.1. Subjects

Eight paid listeners (5 males, 3 females) with normal audiometric thresholds participated in the subjective evaluation experiment. Listeners participated in 60 trials broken into self-paced 30 minute blocks over the course of two weeks. All listeners had previous experience with virtual spatial audio in the context of objective localization experiments conducted within the laboratory, however, all listeners were believed to be naive to both the subjective evaluation method presented below and to the formalized concept of externalization at the onset of the experiment. As such, the subjects were presented with the following verbal description of externalization at the onset of every experimental trial to familiarize them with the question at hand.

Externalization: *To what extent does the virtual source sound outside your head?*

In this type of trial you will be asked to judge how each virtual source is positioned relative to

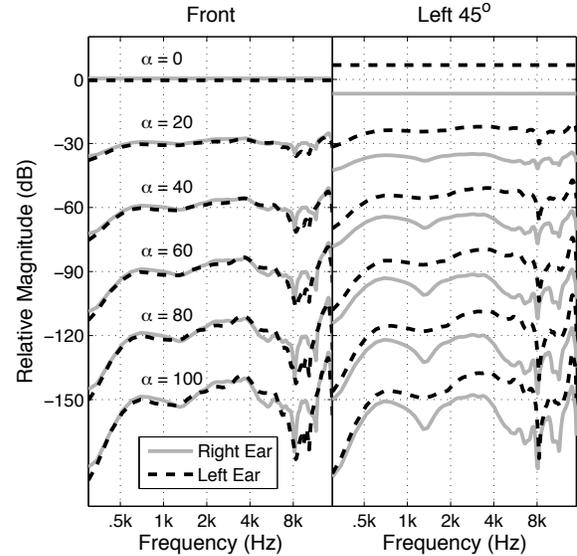


Figure 1: Transformed HRTF spectra for a single subject at to locations on the horizontal plane (panels) as a function of the α parameter. Spectra were given a $\frac{-30\alpha}{20}$ dB gain for plotting purposes.

yourself. When listening over headphones, some sounds may appear as though they originate from inside your head (completely internalized) while others may sound as though they clearly come from a physical location out in space (completely externalized); variations between these two extremes are also possible where a sound might appear to come from on your face, head, or neck or just outside your body.

3.2. Task Description

A single experimental trial consisted of an implementation of the Multi-Stimulus Test with Hidden Reference and Anchor (MUSHRA) [12]. In this task, listeners were presented with the GUI depicted in Fig. 2. By pressing the selection buttons on the GUI (labeled "REF", "A", "B", ..., "F"), listeners were able to selectively listen to each stimulus one at a time as many times as they desired and in any order throughout a trial. Listeners were asked to compare the various stimuli both to each other and to a reference stimulus and provide an externalization rating for each stimulus according to the scale in Table 1 which was always visible to the subjects at the left of the GUI. They were also provided written instructions on the use of the GUI and informed that the reference stimulus should correspond to a rating of 100 on the provided scale. At any time during a trial they could reexamine the verbal description of externalization, adjust the overall stimulus level, and leave comments utilizing the GUI.

3.3. Stimuli

On a given trial seven different stimuli were employed, the reference stimulus and six test stimuli. The reference stimulus always consisted of a virtual sound source rendered with a full HRTF,

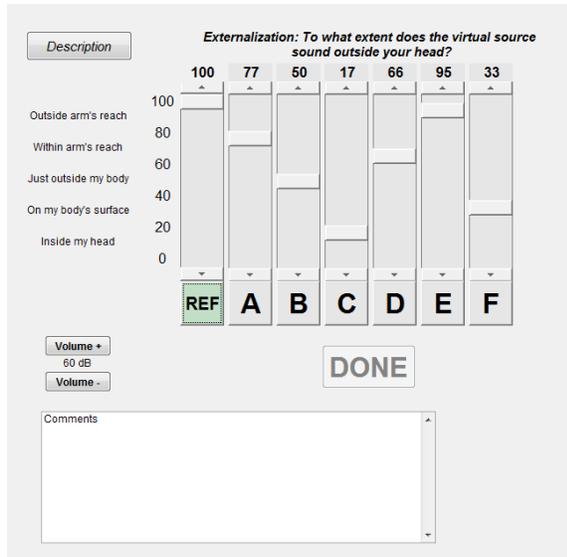


Figure 2: Graphical user interface for the MUSHRA task used during the subjective evaluations.

Rating	Description
80 - 100	Outside arms reach
60 - 80	Within arms reach
40 - 60	Just outside my body
20 - 40	On the surface of my face, head, or neck
0 - 20	Inside my head

Table 1: Verbal descriptions of different levels of externalization and the corresponding rating values used for the subjective evaluation.

while the test stimuli consisted of virtual sound sources rendered with HRTFs that had been transformed as described in Sec. 2 for α values of 0, 20, 40, 60, 80, and 100. The $\alpha = 100$ stimulus was identical to the reference stimulus, therefore acting as the hidden reference and the $\alpha = 0$ stimulus acted as the hidden anchor.

All individualized HRTFs had been previously measured on each listener with the methods described in [3] and consisted of 256 minimum-phase DFT coefficients (sampled at 44.1 kHz) for each ear and a corresponding ITD value at 2 spatial locations, in front of and 45° to the left of the subject). HRTFs were converted to the log-power (decibel) domain, transformed, and ultimately converted back to 256-tap minimum-phase filters. Test stimuli, each 5 s in duration, were generated by convolving one of three single-channel base signals (broadband noise, music, spoken English) with the resulting right and left filters and delaying the contralateral ear by the ITD value. The music and speech samples were taken from the Sound Quality Assessment Material recordings for subjective tests (Track 70 and Track 50, respectively) [13]. All stimuli were normalized post-filtering to have the same average initial level of 60 dB SPL. Each subject participated in ten trials for each location and stimulus type for a total of 60 trials.

4. RESULTS

Across all three base stimuli and both azimuths, two subjects showed a negative correlation with the average trend ($r = -0.78$, $r = -0.81$) computed with their data removed. While it is conceivable that their reference HRTF produced poorly externalized stimuli through some type of measurement artifact (thus resulting in a low externalization rating for $\alpha = 100$), the near perfect reverse ratings exhibited, including rating the anchor, which contained only gross ITD and ILD cues, with a high externalization rating, leads the authors to believe that the subjects misunderstood the task or rating scale. Because of this these two outliers were removed from remaining data and analysis.

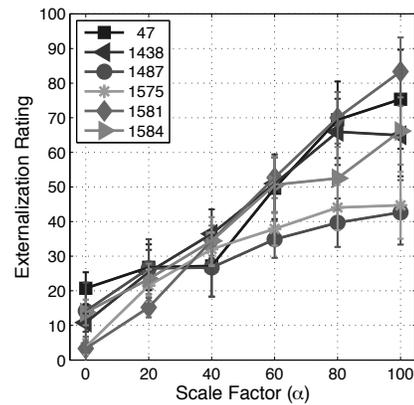


Figure 3: Average externalization ratings for the filtered noise base stimuli at the left 45° position by subject. Error bars represent 95% confidence intervals for the pooled data.

Figure 3 shows the average ratings for the remaining subjects at the left 45° azimuth location averaged across the three base stimuli as a function of the α level. It is clear from Fig. 3 that subjects showed a systematic linear relationship between the α parameter and their externalization ratings. Repeated measures ANOVA results indicate a significant main effect for the α parameter on the externalization rating ($F(5,5) = 9.073$, $p < 0.001$). In general subjects indicated the $\alpha = 0$ condition (containing only ITD and ILD cues) was “Inside (the) head” or “On the surface”. Less consensus is seen in the slope of the functions however, and likewise the rating given to the full HRTF condition ($\alpha = 100$), where ratings varied from “Outside arms reach” to “Just outside (the) body”.

In contrast, ANOVA results did not indicate a statistically significant main effect for base stimulus type ($F(2,2) = 2.305$, $p = 0.150$). Results comparing the ratings for the three base stimuli types averaged across subjects and location are shown in Fig. 4. While not statistically significant, the average data does show a slight trend for externalization ratings to be highest in speech condition compared to the music for bandpass filtered noise.

Figure 5 shows the ratings for the two locations as a function of the α parameter averaged across subjects and stimulus type. Clearly evident in the figure is an interaction with the stimulus location and the α level; the ratings start lower but increase more rapidly for the left 45° location compared to the front. This observation is backed up by ANOVA results which show a significant interaction for α and location ($F(5,5) = 4.891$, $p = 0.003$), but

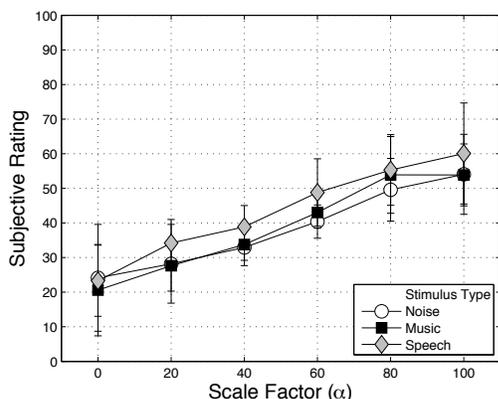


Figure 4: Average externalization ratings pooled over subject and location. Marker color represents base stimulus type. Error bars represent 95% confidence intervals for the pooled data.

no significant main effect for location itself ($F(1,1) = 3.943$, $p = 0.121$).

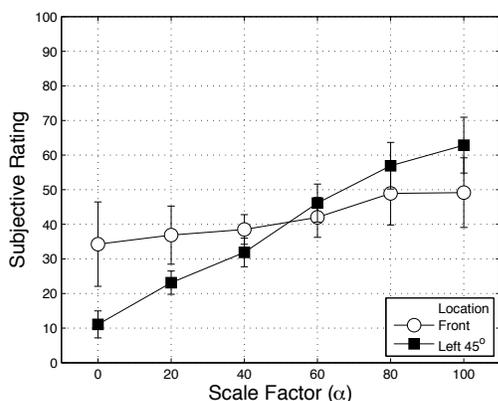


Figure 5: Average externalization ratings pooled over subject and base stimulus type. Marker color represents location. Error bars represent 95% confidence intervals for the pooled data.

5. DISCUSSION

The current results agree with the existing literature for the two extreme α conditions. The internalized ratings given to $\alpha = 0$ condition agree with previous studies utilizing only gross binaural cues, and the full HRTF condition where $\alpha = 100$ showed the expected externalized ratings. More interestingly, the intermediate results suggest that manipulating the strength of narrowband spectral features clearly affects the perceived externalization in a systematic fashion. This implies that the parameterization technique introduced here might be adequate for an externalization or distance based display technology.

Somewhat surprisingly, no effect was found for the different types of base stimuli, despite the fact that they differed significantly in terms of their long-term average spectral profile. Based

on those differences the results of Little *et al.* [9] would suggest that the high-frequency roll off seen for the speech and to a lesser extent music would result in higher externalization ratings compared to the flat spectrum noise. This discrepancy could be explained by the blocked nature the experiment since different base stimuli were never compared directly. It is also likely that the high frequency roll off cue is only used when the different stimuli are assumed to be from the same original sound source, a situation clearly not applicable across base stimuli.

The interaction between the α parameter and the stimulus location may be additional evidence to suggest a relationship between externalization and distance perception. By examining the spectral profiles illustrated in Fig. 1, we can clearly see known low-frequency ILD distance cues present for the lateral location that are not available for the front location. The left 45° profiles clearly show an increase in low frequency ILD cues as α is decreased similar to the near-field HRTF cues observed by Brungart *et al.* [14], and an increase in spectral roll off as α is increased similar to the propagation-related distance cue described by Little [9]. The front location only contains the roll off cue due to its lack of ILD.

In addition to the current positive results, to be valuable as a display technology the parameterization should preserve both the perceived sound quality and localization accuracy. Further research will have to be completed in order to investigate these factors.

6. SUMMARY

The current work describes a simple parameterized method for controlling the perceived externalization of a sound source based on flattening of the monaural HRTF spectra. Examinations show that this method can be related to previously observed cues used for auditory distance perception, and a subjective evaluation demonstrates the technique is capable of producing the desired perceptual results.

7. REFERENCES

- [1] S. Mehrgardt and V. Mellert, "Transformation of the external human ear," *J. Acoust. Soc. Am.*, vol. 61, pp. 1567–1576, 1977.
- [2] A. W. Bronkhorst, "Localization of real and virtual sound sources," *J. Acoust. Soc. Am.*, vol. 98, pp. 2542–2553, 1995.
- [3] D. S. Brungart, G. D. Romigh, and B. D. Simpson, "Rapid collection of HRTFs and comparison to free-field listening," in *International Workshop on the Principles and Applications of Spatial Hearing*, 2009.
- [4] R. Martin, K. McAnally, and M. Senova, "Free-field equivalent localization of virtual audio," *J. Acoust. Soc. Am.*, vol. 49, pp. 14–22, 2001.
- [5] J. Blauert, *Spatial Hearing*. The MIT Press, 1997.
- [6] W. M. Hartman and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, pp. 3678–3688, 1996.
- [7] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, "The contribution of head movement to the externalization and internalization of sounds," *PLoS*, vol. 8, p. 1, 2013.

- [8] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica*, vol. 91, pp. 409–420, 2005.
- [9] A. D. Little, D. H. Mershon, and P. H. Cox, "Spectral content as a cue to perceived auditory distance," *Perception*, vol. 21, pp. 405–416, 1992.
- [10] A. W. Mills, "Lateralization of high frequency tones," *J. Acoust. Soc. Am.*, vol. 32, pp. 132–134, 1960.
- [11] D. J. Kistler and F. L. . Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, pp. 1637–1647, 1992.
- [12] G. A. Soulodre and M. C. Lavoie, "Subjective evaluation of large and small impairments in audio codecs," in *AES International Conference, Florence*, 1999.
- [13] *EBU TECH 3253: Sound Quality Assessment Material recordings for subjective tests*.
- [14] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources; head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 103, pp. 1465–1479, 1999.

QUANTITATIVELY VALIDATING SUBJECTIVELY SELECTED HRTFS FOR ELEVATION AND FRONT-BACK DISTINCTION

Ziqi Fan, Yunhao Wan, Kyla McMullen

University of Florida - SoundPad Lab
313A CSE Building
Gainesville, FL 32611

[fanzq1991|yunhao|drkyla]@ufl.edu

ABSTRACT

As 3D audio becomes more commonplace to enhance auditory environments, designers are faced with the challenge of choosing HRTFs for listeners that provide proper audio cues. Subjective selection is a low-cost alternative to expensive HRTF measurement, however little is known concerning whether the preferred HRTFs are similar or if users exhibit random behavior in this task. In addition, PCA (principal component analysis) can be used to decompose HRTFs in representative features, however little is known concerning whether the features have a relevant perceptual basis. 12 listeners completed a subjective selection experiment in which they judged the perceptual quality of 14 HRTFs in terms of elevation, and front-back distinction. PCA was used to decompose the HRTFs and create an HRTF similarity metric. The preferred HRTFs were significantly more similar to each other, the preferred and non-preferred HRTFs were significantly less similar to each other, and in the case of front-back distinction the non-preferred HRTFs were significantly more similar to each other.

1. INTRODUCTION

3D audio is used in many settings to augment a wide variety of tasks including improving immersion in virtual reality, enhancing speech intelligibility, improving video games, providing spatial cues for assistive technology for persons with visual impairments, enriching positional systems for air traffic controllers, and sonifying multidimensional data. In each of these scenarios, 3D audio minimizes cognitive load, provides spatial cues for degraded visual environments, and provides information redundancy, which significantly improves task performance [1, 2, 3].

3D audio cues are most effectively realized through the use of head-related transfer functions (HRTFs). Proper perceptual fidelity in virtual auditory environments requires the use of individualized or customized HRTFs. The use of a generic or non-individualized HRTF leads to poor elevation perception, decreased externalization, and increased front/back reversal errors [4, 5, 6, 7, 8].

The most accurate HRTFs are obtained by direct acoustic measurement, however the measurement process is very expensive in terms of time and resources [4, 9, 10, 11, 5]. By placing a lis-

tener in an anechoic chamber and positioning a loudspeaker at a known location, it is possible to measure the entire acoustic transformation of a sound by the listener's body. The frequency response is recorded at each ear with microphones placed in the ear canal. When used to synthesize virtual auditory sources, HRTFs are typically realized as the cascade of a minimum-phase FIR filter and an all-pass filter that accounts for the lag in the wavefront arrival time between the two ears [8, 12]. Directly measured HRTFs provide an individualistic 3D sound experience for each person, according to their specific anthropometric features. Although this direct measurement may produce the most accurate measurement, perhaps such costly and resource intensive measurements are not completely necessary to convey 3D sound to a listener.

As a solution, many researchers have proposed less costly methods to alleviate the need for individualization that have been met with varying levels of success. These methods include: HRTF approximation using theoretical computation [13, 14, 15, 16, 17], active sensory tuning [18, 19], machine learning [20, 21, 22, 23, 24], genetic algorithms [25, 26, 27], clustering [28, 29, 30, 31], generic models [4, 5, 32, 7], physical feature measurement [33, 34, 35, 36, 37, 38, 39], pre-measured [40, 41, 10], and subjective selection from pre-measured HRTF databases [42, 43, 44, 45, 46].

Many quantitative and qualitative metrics have been proposed to analyze HRTF similarity [47, 48, 49, 28, 23]. Principal Component Analysis or PCA arises as an objective method to use the common features of an HRTF to decompose it into features that can be varied. PCA is used to describe a data set by using only a few orthogonal components and corresponding weights. For example, Martens was one of the first researchers to show that variations in an HRTF's spectral energy distribution with changing azimuth could be adequately captured by four principal components, quantified in terms of spectral band weighting functions. This finding simplified HRTF analysis by providing a simple (4D) measure of global spectral variation which until then was otherwise difficult to quantify [50]. Following this work, PCA has been used by many researchers to decompose the HRTF.

Though many researchers have used PCA and other decomposition tools to define features and (sometimes) measure HRTF similarity, many if not all of these approaches neglect to address the perceptual validity of the features used to represent HRTFs. Even if the exact mathematical relationships had been discovered, is no perceptual basis to prove that the extracted features affect 3D audio perception. The most straightforward method to determine a listener's preference for HRTFs and how it impacts their 3D audio perception is through a subjective selection procedure in which a listener can choose useful HRTFs from a database of pre-measured



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

filters. This approach is an effective yet inexpensive method to obtain a listener’s HRTF preference, however, little is known about the similarities of the subjectively selected HRTFs. For example, Schonstein and Katz [51] found large variances in judgments for listeners perceptually evaluating HRTFs. The next logical step in this line of research is to analyze the subjectively selected HRTFs chosen by a listener to discover if similarities exist that can be quantified.

The present work uses PCA metrics to decompose a set of pre-measured HRTFs and uses clustering to group them. A given set of listeners perform a subjective selection task in which they indicate their HRTF preferences. Following this, a difference analysis is used to measure HRTF similarity of preferred HRTFs as compared to not-preferred HRTFs. If similarity is observed, this will achieve our goals of (1) demonstrating that the mathematical decomposition of HRTFs using PCA actually has perceptual validity and (2) validating subjective selection as an appropriate method for naive listeners to customize their listening experience. If this work proves to be successful, instead of using the complete HRTF, 3D audio designers could build simplified HRTF models based on the relevant features extracted.

2. BACKGROUND

When clustering a large multidimensional dataset, among the first factors to decide upon is an appropriate metric to use when representing HRTFs. Bondu et. al [28] analyzed several different criteria for clustering HRTFs and discovered that the Avendano criterion performed best both in terms of localization performance and clustering. This would be ideal for our usage, however, Avendano criterion is restricted to the frequency domain. This means that by using this metric, any time domain information would be omitted from the analysis. The present work employs a metric that combines both time and frequency domain information for clustering. The method used in the present work is described in the following sub-sections.

2.1. HRTF Decomposition & Clustering

2.1.1. Principal Component Analysis

PCA is a data reduction method that is used to describe the relevant features of the HRTF. In his seminal study, Martens [50] proposed a method of using spectral band energy to analyze the statistical features of HRTFs. In his study, the spectral area of an HRTF was decomposed into 24 sub-bands centered with 24 frequencies from low to high spectral area. Each HRTF was transformed into a 24 element vector. Then, each vector corresponding to each HRTF was grouped together into a matrix, whose rows corresponded to observations and columns corresponded to frequency elements. Lastly, PCA was performed on the formed matrix to decompose the observations into the combinations of different bases. Based on the study of the shape of the bases, Martens found that the energy bands of frequency should be regrouped into four sets, which are grouped in the following table.

2.1.2. HRTF Clustering

After decomposing an HRTF into relevant features, a *k*-Means clustering algorithm can be used to partition data into *k* mutually exclusive clusters based on their distance to the centroid of a

Group1	166,282,410,543,681,825,980,1158,1368,1616
Group2	1909,2255,2664,3146,3716,4390,5185
Group3	6125,7235,8545,10094,11923
Group4	14083,16634

Table 1: Frequency band grouping, from Martens [50]

cluster. The algorithm forms groupings or clusters in such a way that data within a cluster have a higher measure of similarity than data in any other cluster. The measure of similarity on which the clusters are modeled is defined by Squared Euclidean metric. The *k*-Means algorithm treats each observation in the dataset as an object in a specific location in space. The partition found in the algorithm ensures that objects in the cluster are as close to each other as possible and as far from other objects in other clusters as possible. The *k* centers of the clusters are initialized through Arthur & Vassilvskii’s algorithm [52] and then an iterative algorithm is used to minimize the sum of the distances from each object to its cluster centroid, for all clusters. The algorithm moves objects between clusters until the sum cannot be further minimized. The algorithm runs as follows:

Given cluster number *k* and a set of *n* data points χ .

1. Randomly choose one center c_1 from χ
2. Take a new center c_i , choosing $x \in \chi$ with probability $\frac{D(x)^2}{\sum_{x \in \chi} D(x)^2}$
3. Repeat step 2 until *k* centers $C = \{c_1, \dots, c_k\}$ have been taken together
4. For each $i, j \in \{1, \dots, k\}$, cluster $C_i \in \chi, x_n \in C_i$, if $\|x_n - c_i\| \leq \|x_n - c_j\|$, for all $j \neq i$.
5. For each $i \in \{1, \dots, k\}$, set c_i to be the center of mass of all points in C_i : $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$
6. Iterate steps 4 and 5 until *C* no longer changes

2.2. HRTF Similarity

Once the HRTFs have been decomposed into their relevant features, an HRTF similarity metric can be employed. The similarity for different sets of HRTFs, i.e., preferred, non-preferred, is described by the distance between the average energy for all relevant HRTF azimuths and elevations, as represented in Equation 1,

$$distance_j = \frac{1}{N} \sum_{i=1}^N (E_{j1i} - E_{j2i})^T (E_{j1i} - E_{j2i}) \quad (1)$$

where the subscripts 1 and 2 denote 2 different sets of HRTFs to be compared, N denotes the total number of pairs of HRTFs from the total HRTF database, and i represents all of the HRTF directions for two different HRTF sets. Letter j can be defined to be either left or right ears, E represents the 4 element energy vector, derived from the PCA described in Section 2.1.1. Thus, Equation 1 provides a metric to quantify distance between different sets of HRTFs that can be generalized for any analysis.

It should be noted that since HRTFs consist of both left and right ear filters, we must first redefine the energy vector as the ratio between left and right ears:

$$E = \log \frac{E_l}{E_r} = \log E_l - \log E_r \quad (2)$$

After this step, Equation 1 was used to calculate HRTF similarity.

2.3. Spectral Notch Analysis

The locations of spectral notches within HRTFs convey 3D audio cues that allow listeners to distinguish sounds rendered in the front from sounds rendered in the back and provides and provides elevation cues that allow a listener to determine the height of a spatialized sound source [53, 54]. Notches are characterized by the minimal value points in HRTF magnitude response. To determine the locations of spectral notches within an HRTF, a fast Fourier transform is performed on the head-related impulse response (HRIR). Then, every point in the magnitude response is compared with its adjacent points.

3. METHOD

3.1. Subjective Selection Experiment

The subjective selection experiment methodology used in the present study is thoroughly outlined in in Wan et al. (2015) [55] and summarized in this section.

3.1.1. Subjects

12 subjects (5 male and 7 female) with normal hearing were recruited to participate in the study. Participants were either full-time students or undergraduate summer research students at Clemson University. It should be noted that 3 subjects' performance classified them as outliers. This was because they either (1) preferred all of the HRTFs in the dataset all of the time or (2) Never picked an HRTF consistently. Accordingly, their data was removed from the analysis.

3.1.2. Stimuli

The stimulus used in the study was a 500-ms infrapitch noise, which was constructed by creating pink noise with 200-ms sampling that was repeated 2.5 times.

14 HRTF datasets were used. 13 were from the CIPIC database[9], and one KEMAR dataset. For each trial, the program created impulse responses for the left and right ears by cascading the minimum-phase head-related impulse responses, which were drawn from a given set of HRTFs, with the all-pass impulse responses for the KEMAR ITD.

3.1.3. Procedure

All participants were randomly divided into two groups, A and B, which performed the experiment in different order. In each session, subjects in *Group A* went through a three-stage listening procedure in which they evaluated the perceptual criteria in the following order: externalization, elevation discrimination and front/back discrimination. Listeners in *Group B* evaluated the perceptual criteria in the following order: externalization, front/back discrimination, elevation discrimination. In each stage, the participants judged each HRTF's ability to render the given perceptual cue. Each three stage procedure constituted an experimental session. Each subject completed 3 experimental sessions, occurring between 1 and 2 days apart. To eliminate any potential stimuli judgment bias, subjects were unknowingly deceived by being told

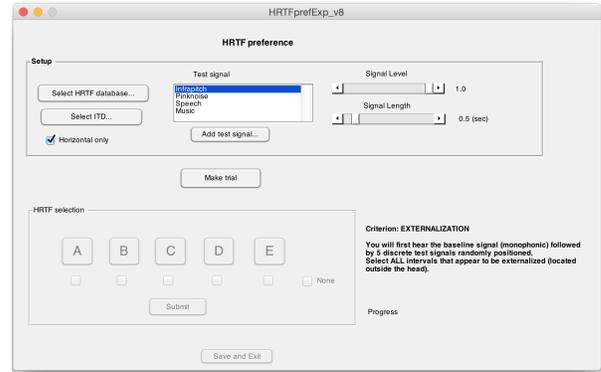


Figure 1: User interface for the spectral coloration selection tasks.

that they would be judging 3 different sets of sounds on each day, when in fact the same HRTFs were used for each session.

The experiment's user interface is shown in **Figure 1**. In each stage, the HRTFs were randomly ordered and presented 4 times per HRTF. In each interval (as shown in Figure 1), 5 HRTFs (marked as letters A, B, C, D, and E) rendered 3D sound. The HRTF corresponding to each letter was highlighted as it was used. The participant could replay any option by clicking its letter. There was a check-box below each HRTF letter that allowed the listener to select whether the HRTF that was used provided adequate cues for the given stage.

In the case of *front/back discrimination*, each trial started with an unspatialized (monaural) reference signal, which was generated by processing the test signal with the HRTF at 0° azimuth, 0° elevation and cross-summing the left and right channels. The purpose for this step was to avoid spectral coloration variability between the raw and the processed test signal. This signal was used as an in-the-head reference, to which the HRTF rendered sounds could be compared. Following the reference signal, the listener heard five consecutive sounds, generated from randomly selected HRTFs, at randomly selected azimuths ($\pm 150^\circ$, $\pm 120^\circ$, $\pm 60^\circ$, $\pm 30^\circ$) on the horizontal (ear level) plane. All of the intervals in each trial used the same sequence of azimuths. If none of the intervals gave the perception of externalization, the listener selected the "None" check-box. After submitting the selections, the results were saved, and the listener proceeded to next trial. The *elevation discrimination* phase proceeded almost identically to the *front/back discrimination* phase. The only difference was that each interval was rendered using a selected HRTF at a random azimuth from: $\pm 150^\circ$, $\pm 120^\circ$, $\pm 90^\circ$, $\pm 60^\circ$, $\pm 30^\circ$; at elevations of $\pm 36^\circ$. The *externalization* phase required the listener to judge the externalization of each interval.

4. RESULTS

4.1. Evaluation Metrics

As described in Section 2.1.1, PCA was performed on each of the CIPIC HRTFs. After band pass filtering the HRTF with center frequencies given in the 1, the energy in the four groups were calculated thus forming a four elements feature vector for each HRTF. PCA is then repeated on the newly derived energy vectors to analyze the energy of the HRTFs.

As described in section 2.1.2, HRTF clustering was also performed to observe any HRTF groupings. Each HRTF’s left and right ear energy vectors for all directions (Table 1) were used for clustering. HRTF similarity was quantified using the metrics described in 2.2.

The metric described in Section 2.3 was used to determine the locations of the spectral notches to evaluate its position and magnitude in each band group. The manner in which to locate the spectral notches was slightly revised due to the fact that CIPIC’s own post-processing of the raw data uses a Hanning window to remove any room reflections [9]. This processing is problematic because the high-frequency components of the HRTF are filtered away. It is for this reason that the notch in the last energy band group was purposefully omitted.

In order to analyze the time dependent HRTF features, the ITD information was extracted and plotted with the notch position to further assess similarity. In these analyses, the notch positions and ITDs are scaled such that their values are between -1 and 1 inclusive. All of the aforementioned metrics in this section were used to assess whether the preferred HRTFs for elevation and front/back distinction are related to their energy features, spectral notch location, or time-dependent features.

4.2. Subjective Selection Preferences

Figure 2 summarizes the results of the experiment described in Wan et al. (2015) [55]. Figures 3 through 7 show the results of the present study. It should be noted that 3 of the original twelve subjects were considered outliers because they did not demonstrate selective behavior. These subjects either did not pick any HRTFs consistently, or they picked almost all of the HRTFs all of the time. Thus, their data was not chosen to be included in the analysis. In addition, the present analysis focuses on the *front/back* and *elevation* distinction stages of the experiment since these stages rely heavily on the listener’s ability to discriminate HRTF spectral features.

4.3. Spectral Similarity of Chosen HRTFs

4.3.1. Elevation Distinction

Figure 3 shows the clustering of the HRTFs used during the elevation distinction stage. The vectors for clustering are the spectral band energy based on the band group in table 1. Each axis represents a frequency band. In the analysis, band groups 2, 3 and 4 are used for clustering. The 4 asterisks in the plot are the centroid of four clusters. The various colors indicate the cluster in which a particular HRTF belongs. The text in the plot represents each of the HRTF datasets used in the analysis. The symbol ‘k’ represents the KEMAR dataset and the numbers ‘1’ through ‘13’ represent the CIPIC HRTF datasets that were used in the experiment, in numeric order. In this figure, the HRTFs preferred in all 3 sessions by Subject 2 are displayed more prominently than the HRTFs that were not preferred.

Table 2 displays the similarity between the preferred and non-preferred HRTFs chosen in the elevation distinction stage (as compared to the average HRTF similarity) for each subject as calculated according to the metric described in Section 2.2. A small value indicates a higher degree of similarity and a larger value indicates dissimilarity. A dash (‘-’) symbol in the table indicates that a listener did not have a specific preference. An ‘0’ in the

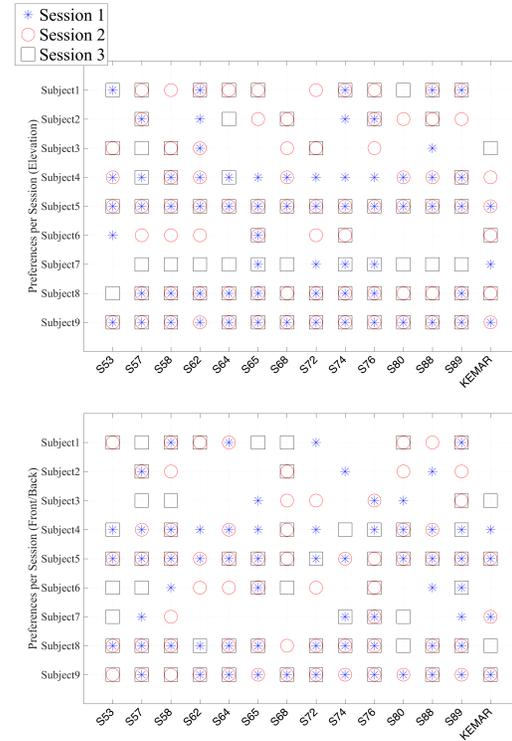


Figure 2: HRTFs chosen by each subject at the end of each session. Along the abscissa is the HRTF identifier and along the ordinate is the subject identifier

table occurs if the listener only consistently preferred one HRTF, and therefore the difference between the HRTFs is 0.

The aggregated results are summarized in Figure 4.

Figure 4 compares the mean distances between preferred HRTFs, non-preferred HRTFs, and the distances between those chosen and not chosen, as compared to the average distance between HRTFs, displayed by the horizontal line in the plot. The results presented were assessed with an ANOVA. Error bars in all figures indicate 95 % confidence intervals. Overall, the preferred HRTFs were significantly more similar than the non-preferred HRTFs. Furthermore, the preferred and non-preferred HRTFs were significantly less similar than the average HRTF similarity [$F_{3,113} = 3.36, p < 0.05$]. The similarity score for the average HRTF group is significantly higher than that for the PvsNP group, suggesting that, for the elevation dimension, PvsNP were more similar than the average HRTFs.

4.3.2. Front/Back Distinction

In a similar fashion as Figure 3, Figure 5 shows the HRTF clustering for the HRTFs used during the front/back distinction stage. In this figure, the HRTFs preferred in all 3 sessions by Subject 1 are displayed more prominently than the HRTFs that were not preferred.

In a similar fashion as Table 2, Table 3 displays the similarity between the preferred and non-preferred HRTFs chosen in the front/back distinction stage (as compared to the average HRTF

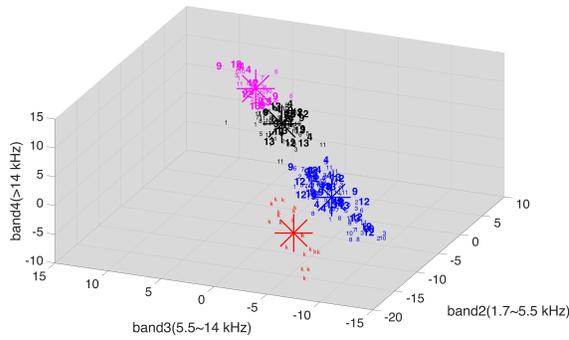


Figure 3: Clustering of the HRTFs used in the elevation distinction stage. Along each axis is the frequency band group. The preferred HRTFs chosen by Subject 2 are highlighted for comparison. The four different colors represents 4 HRTF clusters.

	<i>Preferred</i>	<i>Non-preferred</i>	<i>Average</i>	<i>PvsNP</i>
Subject1	2.01	2.48	4.91	6.03
Subject2	0	4.01	4.91	4.66
Subject3	-	4.70	4.91	4.93
Subject4	1.48	-	4.91	-
Subject5	4.35	-	4.91	-
Subject6	0	3.88	4.91	6.71
Subject7	0	4.03	4.91	7.20
Subject8	5.0	-	4.91	-
Subject9	4.41	-	4.91	-

Table 2: Similarity of preferred and non-preferred HRTFs by subject in elevation distinction stage.

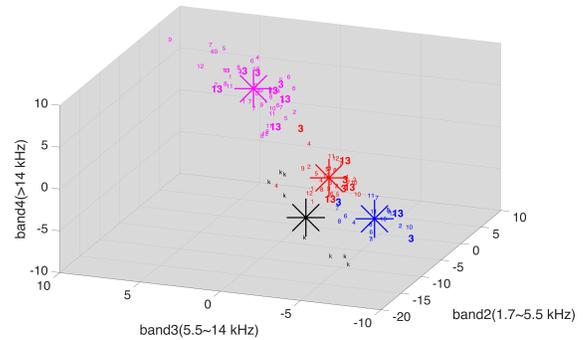


Figure 5: Clustering of the HRTFs used in the front/back distinction stage. Along each axis is the frequency band group. The preferred HRTFs chosen by Subject 1 are highlighted for comparison.

	<i>Preferred</i>	<i>Non-preferred</i>	<i>Average</i>	<i>PvsNP</i>
Subject1	1.75	1.76	4.52	3.73
Subject2	0	3.51	4.53	7.46
Subject3	-	2.52	4.53	3.81
Subject4	2.02	-	4.53	-
Subject5	3.99	-	4.53	-
Subject6	0	1.99	4.53	4.44
Subject7	0	4.28	4.53	4.70
Subject8	4.22	-	4.53	-
Subject9	4.92	-	4.53	-

Table 3: Similarity of preferred and non-preferred HRTFs by subject in front/back distinction stage.

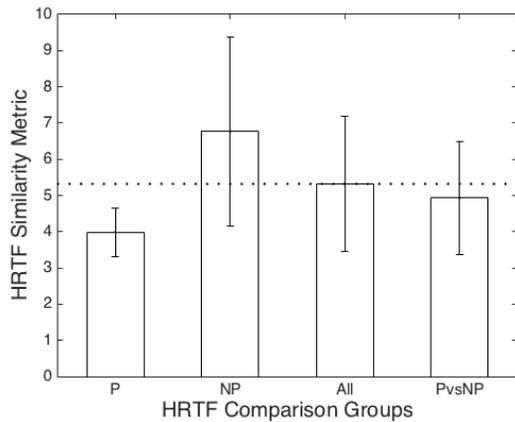


Figure 4: Similarity of preferred (P), non-preferred (NP), All, and preferred to non-preferred (P vs NP) HRTFs used for elevation distinction. Along the abscissa are the comparison groups and along the ordinate is the similarity metric score. Lower values indicate more similarity.

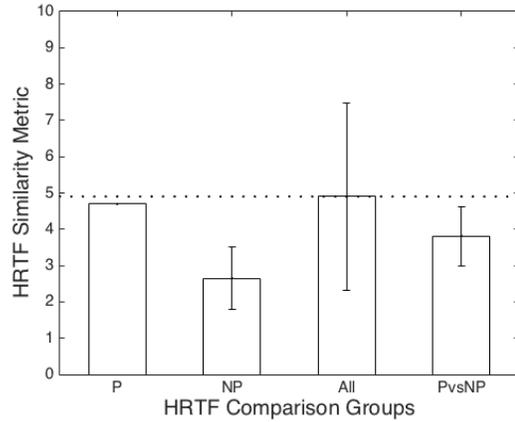


Figure 6: Similarity of preferred (P), non-preferred (NP), All, and preferred to non-preferred (P vs NP) HRTFs for front/back distinction. Along the abscissa are the comparison groups and along the ordinate is the similarity metric score. Lower values indicate more similarity.

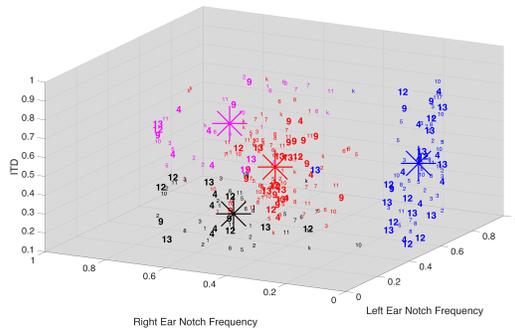


Figure 7: Clustering of HRTF notch locations and ITD for the elevation distinction stage. Along the x-axis is the location of the left ear notch frequency, along the y-axis is the location of the right ear notch frequency, and along the z-axis is ITD. The preferred HRTFs chosen by Subject 2 are highlighted for comparison.

similarity) for each subject as calculated according to the metric described in Section 2.2.

Figure 6 compares the mean distances between preferred HRTFs, non-preferred HRTFs, and the distances between those chosen and not chosen, as compared to the average distance between HRTFs, displayed by the horizontal line in the plot. The results presented were assessed with an ANOVA. Error bars in all figures indicate 95 % confidence intervals. Similar to the results displayed in Figure 4, for each subject the preferred HRTFs were significantly more similar than the average similarity. The non-preferred HRTFs were also found to be significantly more similar than the average similarity [$F_{3,98} = 3.09, p < 0.05$]. What's more, with a comparison between non-preferred and preferred, we find that the non-preferred is more similar than the preferred HRTFs.

4.4. Time Dependent Similarity of Chosen HRTFs

4.4.1. Elevation Distinction

Figure 7 shows the results of the HRTF notches and ITD clustering for the HRTFs used during the elevation distinction stage. The axes on horizontal(xy) plane are locations of HRTF notches for left(x-axis) and right(y-axis) ears. The frequency locations of all notch points have been normalized based on a maximum frequency of 22,050 Hz. On the vertical axis(z axis), the ITD is normalized to scale [0,1]. In this figure, Subject 2's preferred HRTFs are highlighted. HRTF similarity was assessed using the ITD and notch features and no significant differences were found. [$F_{3,113} = 0.18, p = 0.91$]

4.4.2. Front/Back Distinction

Similarly, Figure 8 shows the results of the HRTF notches and ITD clustering for the HRTFs used during the elevation distinction stage. In this figure, Subject 2's preferred HRTFs are highlighted. HRTF similarity was assessed using the ITD and notch features and no significant differences were found. F statistics and p value are given as [$F_{3,98} = 1.72, p = 0.17$]

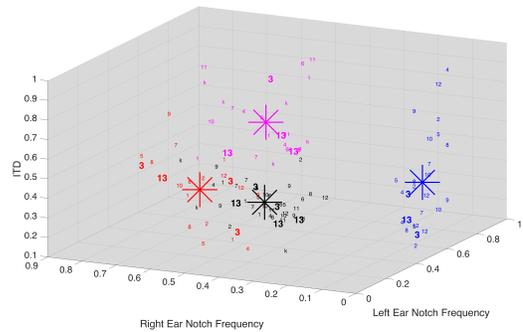


Figure 8: Clustering of HRTF notch locations and ITD for the front/back distinction stage. Along the x-axis is the location of the left ear notch frequency, along the y-axis is the location of the right ear notch frequency, and along the z-axis is ITD. The preferred HRTFs chosen by Subject 2 are highlighted for comparison.

5. DISCUSSION & FUTURE WORK

The goal of the present work was to determine if the PCA features on which HRTFs are clustered have a perceptual linkage and to quantitatively validate the HRTFs chosen in a subjective selection procedure.

In the work we found that HRTFs that were picked were similar in terms of distance from each other. This suggests that PCA decomposition is indeed a valid tool that has a perceptual significance when analyzing HRTFs. This work further validates the subjective selection methodology by showing that there is similarity between chosen HRTFs and novice listeners are capable of using spectral cues to discriminate between features.

Furthermore, the present work shows that HRTFs that were not selected also share similar qualities as they are typically grouped near each other.

The present work highlights the challenge of accurately decomposing an HRTF such that it can be represented as a set of points in space. In creating the current work, many solutions were tried, however it was found that the most informative HRTF representation centered around band energy. This finding suggests that the clusters can be better separated by only considering one direction.

In addition, preferred HRTFs were not significantly more similar than non-preferred HRTFs or as compared to average HRTF similarity. This suggests that the time-dependent features of the HRTF are not as critical in elevation and front/back distinction as the spectral features.

Future work will further delve into the findings and perform a subjective selection experiment in which the user hears a broadband sound coming from a known location, and HRTF features such as notch locations, spectral band energy, and ITDs are manipulated in real-time to interactively tune the HRTF. This experiment will allow us to narrow down on the exact cues that are relevant for 3D audio perception.

Currently, statistical methods of HRTF analysis are frequently limited by the capacity of data set. Up to now, the largest existing publicly available database is CIPIC, which consists of 45 HRTF data sets. Although this dataset is sufficient for many applications, the number of measured HRTFs are not sufficient enough

to perform a thorough analysis of HRTF statistical patterns. Thus, future work will involve a larger HRTF database that is formed by exploring the acoustical basis of HRTFs (as spherical waves) and using acoustic models to create more HRTFs and unify all existing public domain HRTF databases.

6. REFERENCES

- [1] R. Bastide, D. Navarre, P. Palanque, A. Schyn, and P. Dragicevic, "A model-based approach for real-time embedded multimodal systems in military aircrafts," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 243–250.
- [2] J.-O. Nilsson, C. Schuldt, and P. Handel, "Voice radio communication, pedestrian localization, and the tactical use of 3d audio," in *Indoor Positioning and Indoor Navigation (IPIN), 2013 International Conference on*. IEEE, 2013, pp. 1–6.
- [3] V. Y. Nguyen, "Audible assistance on escaping to an emergency exit: A comparison of a manipulated 2d and a 3d audio model," 2004.
- [4] A. W. Bronkhorst, "Localization of real and virtual sound sources," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2542–2553, 1995.
- [5] H. Moller, M. F. Sorensen, C. B. Jensen, and D. Hammershoi, "Binaural technique: do we need individual recordings," *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–469, 1996.
- [6] S. Weinrich, "The problem of front-back localization in binaural hearing," *Scandinavian Audiology. Supplementum*, vol. 15, pp. 135–145, 1981.
- [7] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman, "Localization using non-individualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.
- [8] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," *Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1050–1063, 1997. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/9035397>
- [9] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipc hrtf database," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001, pp. 99–102.
- [10] W. G. Gardner and K. D. Martin, "Hrtf measurements of a kemar," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.
- [11] D. Hammershoi, H. Moller, M. F. Sorensen, and K. Larsen, "Head-related transfer functions: Measurements on 24 subjects," in *92nd Audio Engineering Society Convention*, 1992.
- [12] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. part 1: Stimulus synthesis," *Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867, February 1989.
- [13] V. Algazi, R. Duda, and D. Thompson, "Use of head and torso methods for improved spatial sound synthesis," *Proceeding of AES 113th Convention*, 2002.
- [14] J. Huopaniemi, N. Zacharov, and M. Karjalainen, "Objective and subjective evaluation of head-related transfer function filter design," *Journal of the Audio Engineering Society*, vol. 47, no. 4, pp. 218–239, 1999.
- [15] A. Meshram, R. Mehra, and D. Manocha, "Efficient hrtf computation using adaptive rectangular decomposition," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Aug 2014. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=17366>
- [16] A. Meshram, R. Mehra, H. Yang, E. Dunn, J.-M. Frahm, and D. Manocha, "P-hrtf: Efficient personalized hrtf computation for high-fidelity spatial sound," *IEEE International Symposium on Mixed and Augmented Reality*, 2014.
- [17] K. Terai and I. Kakuhari, "Hrtf calculation with less influence from 3-d modeling error: Making a physical human head model from geometric 3-d data," *Acoustical Science and Technology*, vol. 24, no. 5, pp. 333–334, 2003.
- [18] P. Runkle, A. Yendiki, and G. H. Wakefield, "Active sensory tuning for immersive spatialized audio," *Proceeding of International Conference on Auditory Display*, 2000.
- [19] A. Silzle, "Selection and tuning of hrtfs," in *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.
- [20] H. Hu, L. Zhou, H. Ma, and Z. Wu, "Hrtf personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, vol. 69, no. 2, pp. 163–172, 2008.
- [21] Y. Luo, "Fast numerical and machine learning algorithms for spatial audio reproduction," 2014.
- [22] S. Morioka, I. Nambu, S. Yano, H. Hokari, and Y. Wada, "Adaptive modeling of hrtfs based on reinforcement learning," in *Neural Information Processing*. Springer, 2012, pp. 423–430.
- [23] E. S. Schwenker and G. D. Romigh, "An evolutionary algorithm approach to customization of non-individualized head related transfer functions," in *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.
- [24] M. Washizu, S. Morioka, I. Nambu, S. Yano, H. Hokari, and Y. Wada, "Improving the localization accuracy of virtual sound source through reinforcement learning," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 4378–4383.
- [25] N.-M. Cheung and S. Trautman, "Genetic algorithm approach to head-related transfer functions modeling in 3-d sound system," in *Multimedia Signal Processing, 1997., IEEE First Workshop on*. IEEE, 1997, pp. 83–88.
- [26] E. Durant and G. H. Wakefield, "Efficient model fitting using a genetic algorithm: pole-zero approximations of hrtfs," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 1, pp. 18–27, 2002.
- [27] X. Hu, Y. Tang, and J. Cai, "Optimal approximation of head-related transfer function based on adaptive genetic algorithm," in *Neural Networks and Signal Processing, 2008 International Conference on*. IEEE, 2008, pp. 129–132.
- [28] A. Bondu, S. Busson, V. Lemaire, and R. Nicol, "Looking for a relevant similarity criterion for hrtf clustering: a comparative study," in *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.

- [29] C.-S. Fahn and Y.-C. Lo, "On the clustering of head-related transfer functions used for 3-d sound localization," *J. Inf. Sci. Eng.*, vol. 19, no. 1, pp. 141–157, 2003.
- [30] M. Neal and M. C. Vigeant, "Use of k-means clustering analysis to select representative head related transfer functions for use in subjective studies," *The Journal of the Acoustical Society of America*, vol. 135, no. 4, pp. 2366–2366, 2014.
- [31] S. Shimada, N. Hayashi, and S. Hayashi, "A clustering method for sound localization transfer functions," *Journal of the Audio Engineering Society*, vol. 42, no. 7/8, pp. 577–584, 1994.
- [32] D. Schonstein and B. F. Katz, "Hrtf selection for binaural synthesis from a database using morphological parameters," in *International Congress on Acoustics (ICA)*, 2010.
- [33] N. A. Gumerov, R. Duraiswami, and Z. Tang, "Numerical study of the influence of the torso on the hrtf," *Acoustics Speech and Signal Processing 2002 Proceedings ICASSP 02 IEEE International Conference on*, vol. 2, pp. 1965–1968, 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=1006155
- [34] N. Inoue, T. Kimura, T. Nishino, K. Itou, and K. Takeda, "Evaluation of hrtfs estimated using physical features," *Acoustical Science And Technology*, vol. 26, no. 5, pp. 453–455, 2005. [Online]. Available: <http://joi.jlc.jst.go.jp/JST.JSTAGE/ast/26.453?from=CrossRef>
- [35] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile, *Enabling individualized virtual auditory space using morphological measurements*. IEEE PacificRim Conference on Multimedia, 2000, pp. 235–238. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.113.5338&rep=rep1&type=pdf>
- [36] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, "Estimation of hrtfs on the horizontal plane using physical features," *Applied Acoustics*, vol. 68, no. 8, pp. 897–908, 2007.
- [37] M. Rothbucher, T. Habigt, J. Habigt, T. Riedmaier, and K. Diepold, "Measuring anthropometric data for hrtf personalization," in *Signal-Image Technology and Internet-Based Systems (SITIS), 2010 Sixth International Conference on*. IEEE, 2010, pp. 102–106.
- [38] D. Zotkin, R. Duraiswami, L. Davis, A. Mohan, and V. Raykar, "Virtual audio system customization using visual matching of ear parameters," *Pattern Recognition 2002 Proceedings 16th International Conference on*, vol. 3, pp. 1003 – 1006 vol.3, 2002. [Online]. Available: <http://dx.doi.org/10.1109/ICPR.2002.1048207>
- [39] D. Y. N. Zotkin, J. Hwang, R. Duraiswami, and L. S. Davis, *HRTF personalization using anthropometric measurements*. Ieee, 2003, pp. 157–160. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1285855>
- [40] V. Algazi and R. Duda, "Estimation of a spherical-head model from anthropometry," *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 472–478, 2001.
- [41] E. Blanco-Martín, S. Merino Saez-Miera, J. J. Gomez-Alfageme, and L. I. Ortiz-Berenguer, "Repeatability of localization cues in hrtf data bases," in *Audio Engineering Society Convention 130*, May 2011. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=15888>
- [42] K. McMullen, A. Roginska, and G. H. Wakefield, "Subjective selection of head-related transfer functions (hrtf) based on spectral coloration and interaural time differences (itd) cues," in *Audio Engineering Society Convention 133*. Audio Engineering Society, 2012.
- [43] *User Selected HRTFs: Reduced Complexity and Improved Perception*. Undersea Human Systems Integration, 2010.
- [44] B. Seeber and H. Fastl, "Subjective selection of non-individual head-related transfer functions," *Proceeding of ICAD 2003*, pp. 259–262, 2003.
- [45] Y. Wan, A. Zare, and K. McMullen, "Evaluating the consistency of subjectively selected head-related transfer functions (hrtfs) over time," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Aug 2014. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=17349>
- [46] B. Xie, C. Zhang, and X. Zhong, "A cluster and subjective selection-based hrtf customization scheme for improving binaural reproduction of 5.1 channel surround sound," in *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.
- [47] A. Andreopoulou, A. Rogińska, and H. Mohanraj, "Analysis of the spectral variations in repeated head-related transfer function measurements," 2013.
- [48] A. Andreopoulou and A. Rogińska, *Evaluating HRTF Similarity through Subjective Assessments: Factors that can Affect Judgment*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2014.
- [49] A. Andreopoulou and B. F. Katz, "On the use of subjective hrtf evaluations for creating global perceptual similarity metrics of assessors and assesseees," in *Proceedings of the 21st International Conference on Auditory Display (ICAD 2015)*, 2015.
- [50] W. L. Martens, "Principal components analysis and resynthesis of spectral cues to perceived direction," 1987.
- [51] D. Schönstein and B. F. Katz, "Variability in perceptual evaluation of hrtfs," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 783–793, 2012.
- [52] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.
- [53] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, "Structural composition and decomposition of hrtfs," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001, pp. 103–106.
- [54] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 364–374, 2005.
- [55] Y. Wan, Z. Fan, and K. A. McMullen, "Temporal reliability of subjectively selected head-related transfer functions (hrtfs) in a non-eliminating discrimination task," in *Audio Engineering Society 138th Convention*, 2015.

POSTERS

EXTENDED ABSTRACT:**MUSIC OF MIGRATION AND PHENOLOGY: LISTENING TO COUNTERPOINTS OF MUSK OX AND CARIBOU MIGRATIONS, AND CYCLES OF PLANT GROWTH***Mark Ballora*

School of Music/School of Theatre
 The Pennsylvania State University
 Music Building I, University Park, PA 16802-1901 USA
ballora@psu.edu

1. INTRODUCTION

This extended abstract describes a sonification that was commissioned by a biologist/animal ecologist. The sonification was created with the software synthesis program SuperCollider [1]. The motivation for creating it was to pursue additional levels of engagement and immersion by supplementing the effects of visual plots, as well as to create an informative rendering of a multivariate dataset. The goal is for audiences, in particular students and laypeople, to readily understand (and hopefully find compelling) the phenomena being described. The approach is parameter-based, creating “sonic scatter plots” [2] in the same manner as work described in earlier publications [3], [4].

The work described here is a current experimental project that takes a sonic approach to describing the interactions of plant phenology and animal migrations in Greenland. This area is seen as a predictor of how climate change may affect areas farther south. There is concern about the synchronicity of annual caribou migrations with the appearance of plant food sources, as warmer temperatures may cause plants to bloom earlier and in advance of the caribou arrival at their calving grounds; depleted food availability at calving time can lead to lower populations of caribou.

Parts of this sonification will be applied to a multi-year professional development workshop for middle and high school science teachers. It is hoped that sonifications of plant observations made by teachers and students will enhance student engagement, and possibly lead to greater degrees of understanding of phenology patterns.

2. THE POLAR CENTER AT PENN STATE

The Polar Center at Penn State is an outreach program of the University’s Eberly College of Science. The Center’s mission is to foster understanding, awareness and appreciation of the polar regions through a variety of outreach, education, and research activities. Fine arts as well as the sciences are often employed to communicate the rare beauty as well as the scientific and cultural value of these regions. Its annual Polar Day Symposium, held each spring, features presentations by scientists, writers, and photographers. Since 2014, the event has included sonification presentations of various polar-related phenomena [5], [6].



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

The work described here was created for Polar Day 2016. The sonifications illustrate the interplay of migrations of Greenland caribou, musk ox populations, and the availability of the plant species that are their food sources.

3. DESCRIBING ANIMAL POPULATIONS AND PHENOLOGY**3.1. Background**

Penn State researchers travel to the Russell Glacier, located near Kangerlussuaq, Greenland, each year for approximately two months, from the end of April to the end of June (Julian Days 115-174). They record observations of caribou and musk ox populations, and the dates at which plants appear. Musk ox were placed in the area in the 1960s as a reserve population due to declining populations in other natural habitats. They inhabit the region year-round, and calve in early spring, while the area is still covered in snow. Caribou, which are indigenous to the area, spend the winter in the coastal area of Sisimut, and migrate 250 km inland to the Russell Glacier to calve in early June, when the landscape is green.

The caribou migrations have historically corresponded to the onset of grass and other plant species, which have higher nutrition than winter-growing lichens. Caribou are very conservative in their migration patterns, which date back thousands of years. It is unclear whether they have the ability to change their behaviors and leave their winter grounds earlier to adjust to changes in phenology [7], [8], [9], [10].

3.2. Description of the Dataset

The dataset includes observations taken in 1993, and in the years 2002-2015. Researchers stationed at Kangerlussuaq take daily counts of how many plant species have emerged. There are 9-14 plant species that are food sources for the musk ox and caribou, not all of which emerge each year. At the close of each annual observation period, the timing of the overall available food supply is expressed as a *mean proportion value*, whereby each daily measurement of the number of species that have leafed out is divided by the maximum number of observed species that are observed on Julian Day 174 of that year. This mean proportion normalizes the availability of food supplies to a value between 0 and 1, or from 0% to 100% of the available supplies in a given year. The variability from year to year lies with the first day that is a measurable value is observed, and with the slope of the

progression from 0 to 1, which indicates the rate at which food supplies become available.

Daily counts are also taken of the number of musk ox and caribou observed. For each day, there are three entry types: a value of the total number of each animal observed, a zero if no animal was observed, or a skip in the date sequence if no observations were taken on a given day. To standardize the scaling between food availability and number of animals sighted, the daily counts of musk ox and caribou are also normalized as a mean proportion in the same way that the food supply is expressed as a value between 0 and 1.

3.3. Goals of the Sonification

The goal of sonifying the dataset described above is to investigate whether patterns can be heard that indicate when the plants and animals first appear, as well as changes in their rate of increase.

4. SONIFICATION STRATEGIES

The sound design techniques that are used in this project draw on practices that have been utilized in earlier work, and are based in perceptual principles outlined in sources such as [3], [11], and [12].

4.1. Year and Date

A primary goal of the sonification is to make easily discernible the variations that occur from year to year in the appearance dates and population rates of plants and animals. The dataset consists of 900 total entries, which represent sixty Julian Days, from 115-174, for each year of data. The most basic layer of the sonification is an “auditory calendar,” which marks the dates with a percussive “tap” to represent each day. This type of sound is meant to be unobtrusive, and is discernible at low listening levels. As an undifferentiated stream of taps would be difficult to count and would quickly

become indistinct throbbing (I call this the “woodpecker effect”), regular intervals are demarcated with an accent. This is consistent with the suggestion in [12] that loudness changes may be effectively used as temporal markers. The scientist commissioning the work suggested demarcating each sequence of five days.

This “quintuple meter” is meant to allow listeners to easily hear a difference in arrival times of the plants and animals – at rapid playback rates, it becomes difficult to count the number of days that pass before the phenology activity begins, but it is fairly easy to count the number of accented beats.

Another type of temporal marker annotates the end of each year. A brief pause occurs, followed by a ringing percussion sound that is louder and lasts longer than the tapping sound that represents the days, and indicates the onset of a new year of data.

4.2. GUI

A GUI adds a visual reference (Figure 1). The current year and date are displayed in text boxes at the top of the display, and graphs of the plant, caribou and musk ox data are placed on top of each other. The chart portions illustrate the mean proportion values described earlier. The total numbers of caribou and musk ox observed each year are also printed in their areas of the graph.

A slider moves along the bottom of the display and between each of the graphs to aid the eye in quickly finding the current position in the dataset as it iterates. The GUI also allows adjustment of the time increment between dates and the volume balance between sound streams (described above and in the following section). The current position in the dataset may be adjusted by dragging any of the horizontal sliders.

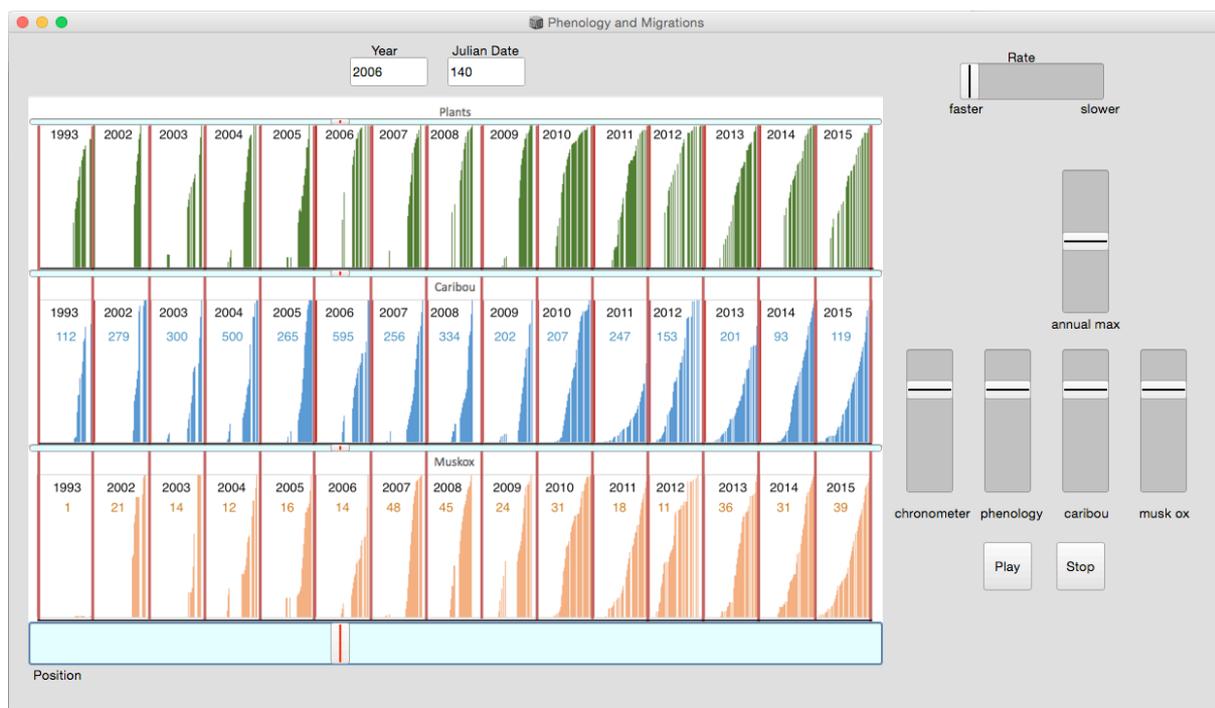


Figure 1: GUI allows control of playback start, pause, stop, volume balances, position within the dataset, and rate of iteration.

4.3. Phenology

The primary information being rendered is the phenology, i.e., the cyclic and seasonal activity of plant growth and animal migrations. Four types of phenology data are sonified: mean proportion values of plants, musk ox, and caribou, and the annual population totals for the caribou. As noted in [12], mapping numeric magnitude to pitch is one of the more intuitive and easily discernible mappings available to sonification designers. (In [3], it is suggested that pitch be considered a *primary auditory cue*, in that small pitch changes are apparent even to untrained listeners, which makes this a strong choice for representing data values.) With pitch as the means of expressing the data values, the values pertaining to plant growth and populations of caribou and muskox are assigned to different “instruments.”

Bearing in mind the caution raised in [12] that it is generally difficult to attend to three or more continually changing streams of auditory information, it seemed appropriate for this sonification to be flexible in how the streams are combined. The GUI, therefore, is meant to serve as a mixing panel, whereby users may adjust the relative balance of the instruments and the position within the dataset at will (as described in the previous section). This is meant to allow users to repeat and refocus as necessary, as one might do when reading and re-reading a passage or studying an image.

As the dataset is iterated, for each date a “note” is played on each of the four phenology-based instruments, plus there is a tap from the percussive chronometer. A QuickTime movie showing a sample run of the sonification may be downloaded from [13].

4.3.1. Pitch Derivation

The mapping of data values to pitch was done as described in [3], whereby a fundamental, f , is selected. The data values, d , multiplied by a scalar, r , are applied as a power of 2, which is multiplied by f , as shown in (1).

$$\text{pitch} = f \times 2^{(d \times r)} \quad (1)$$

With an r value of one, a data value (d) of 0 results in a pitch at f . A data value of 1 results in a pitch an octave higher than f , and a data value of -1 results in a pitch an octave lower than f . This is a “microtonal approach” to pitch mapping, since many data values are likely to result in pitches that fall “between the cracks” of the equal tempered pitches that are found on a concert piano. Since it is easy for untrained listeners to hear variations in pitch that are smaller than the equal tempered half step, this conversion approach has the potential of yielding more nuance in many cases than a coarser approach, such as assigning pitches to MIDI note numbers. The value of r acts as a scalar for the pitch range. The range can be reduced or expanded by changing the value of r to a value above or below 1.0.

For this project, the fundamental, f , was chosen to be 175 Hz. This was a subjective choice. It was a frequency that sounded neither too low nor too shrill to our ears. As outlined above, the observed animals and plants are represented by four “instruments”: plant mean proportion values, musk ox mean proportion values, caribou mean proportion values, and maximum number of observed caribou for each year. (The

mean proportion value was described in section 3.2.) The scalar value (r) is set to one. The result is that as the mean proportion values rise from 0 to 1 over the course of each year’s measurements, each “mean proportion instrument” rises an octave in pitch.

4.3.2. Timbral Characteristics

As described in [3], timbre is useful as a *secondary* (or *supporting*) *auditory cue*, meaning that it is generally not effective for delineating small changes in data values, but can be quite effective for differentiating different streams of information. As timbre is a multi-faceted property, it is recommended in [12] that envelope shape be considered along with harmonic content in creating timbral effects. In this project, the qualities (timbral and otherwise) that are meant to differentiate the instruments include overtone content, attack time, tremolo rate, and pan position. All of these were arrived at by ear, through trial and error, in an attempt to create sound types that were compatible yet also mutually exclusive.

4.3.3. Plant Mean Proportion Values Instrument

The “plant instrument” is a shimmering, percussive sound that is panned center. As the mean proportion values increase from 0 to 1, its changes are mapped to the following parameter ranges:

- The pitch goes from 175 Hz to 350 Hz, an octave higher;
- The relative volume goes from a level of 0.015 to 0.1;
- The tremolo rate goes from 4 Hz to 10 Hz;
- The attack time quickens, going from 0.1 to 0.01 seconds.

4.3.4. Musk Ox Mean Proportion Values Instrument

The “musk ox instrument” is a flute-like sound that is panned to the right. Similar to the plant instrument, as the mean proportion values increase from 0 to 1, its values are mapped to the following parameter ranges:

- The pitch goes from 175 Hz to 350 Hz, an octave higher;
- The relative volume goes from a level of 0.05 to 0.4;
- The tremolo rate goes from 2 Hz to 18 Hz;
- The attack time quickens, going from 0.5 to 0.3 seconds;
- The cutoff frequency of a lowpass filter in the instrument goes from 100 to 1500 Hz.

4.3.5. Caribou Mean Proportion Values Instrument

The “caribou instrument” is a brass-like sound that is panned to the left. As with the other two instruments, as the mean proportion values increase from 0 to 1, its values are mapped to the following parameter ranges:

- The pitch goes from 175 Hz to 350 Hz, an octave higher;
- The relative volume goes from a level of 0.03 to 0.2;
- The tremolo rate goes from 1 Hz to 15 Hz;
- The attack time quickens, going from 0.05 to 0.02 seconds.

4.3.6. Caribou Annual Maximum Instrument

The mean proportion values rendered by the three instruments described above give indicators of the arrival date and rate of growth, but do not contain any information about population fluctuations from year to year. That is, they are based on values that describe percentage of the maximum value. But when they are considered on their own, they can be misleading as there is nothing indicating what a given year's maximum value is. The population totals for caribou and musk ox are printed on the graph displayed in the GUI, as described in Section 4.2. Since the primary focus of this work is to track fluctuations in caribou populations from year to year, an additional auditory stream indicates the maximum numbers of animals observed each year.

A "caribou annual maximum" instrument is created in the form of a complex of five detuned sine waves, with a slight tremolo applied. These sound continuously throughout the playback time of each year, as a background drone. As the population values vary from their minimum of 93 (the value for the year 2013) and their maximum of 595 (the value for the year 2006), the pitch of the sine complex ranges from 175 to 350 Hz, and the tremolo rate is the year's population maximum value multiplied by 0.05, which produces tremolo rates in the range from 4.65 to 29.75 Hz.

Thus, the onset of each year is indicated by two changes: the sound of the bell-like percussive sound and a new annual maximum pitch, which gives an immediate indication of whether populations were greater or lesser than those of the previous year.

5. FURTHER WORK

The intention of this extended abstract is not to present this project as a particularly novel form of sonification, but rather to highlight its context. My suggestion is that this is another small step forward in an ongoing healthy evolution, wherein researchers in a variety of scientific fields are looking to sonification as a means of exploring and presenting data. It has been encouraging to see Penn State researchers become interested in exploring sound as a means of presenting their data, as well as the interest in museum exhibitors in presenting sonified renderings of natural science material as a way of introducing attendees to the dynamics of information being presented [6].

Taking the long view, we are particularly interested in introducing young audiences to science sonifications. By presenting science to a generation of students as something that is understood through listening as well as seeing, we feel that we could add an important dimension to science pedagogy, creating a more holistic and engaging experience than is possible with visual materials alone.

Plans are in place for elements of this sonification to be incorporated into a three-year summer training workshop for middle and high school science teachers [14] that is meant to promote study of phenology and engaged science. By having sonifications of plant growth data as a product of their work, our hope is that the students will have higher degrees of engagement and personal investment in the material. As a federally-funded teacher training program, every facet of it will be subject to assessments, which should give us concrete evidence of the efficacy of sonification in educational contexts.

Enhancement of museum exhibits and incorporation of sonifications into educational materials are the subjects of

other pending grant applications, which will hopefully be the subjects of future publications at ICAD and in other related journals.

6. ACKNOWLEDGMENTS

The data sonifications were commissioned by Eric Post, Professor of Biology at Penn State's Eberly College of Science, and Pernille Sporon Bøving, the Penn State Polar Center's Director for Programming and Engagement, for presentation at Penn State Polar Day 2016. This effort was partially supported by NSF grant PLR 1525636.

7. REFERENCES

- [1] <http://supercollider.sourceforge.net>
- [2] T. Hermann, A. Hunt, J. G. Neuhoff. (Eds.). *The Sonification Handbook*. Berlin: Logos Publishing House, 2011.
- [3] M. Ballora, "Sonification, Science and Popular Music: In search of the 'wow,'" *Organised Sound*, vol. 19, pp. 30-40, April 2014. doi:10.1017/S1355771813000381
- [4] M. Ballora, "Sonification Strategies for the film Rhythms of the Universe." Presented at the 20th International Conference of Auditory Display (ICAD 2014), June 22-25, New York, NY, USA.
- [5] M. Ballora, M. Kenney, "Changes in Antarctic Ice - 400,000 years to present." Presented at PSU Polar Day 2014. <http://www.polar.psu.edu/changes-in-antarctic-ice-400000-years-to-present/>
- [6] M. Ballora, "Two Examples of Sonification for Viewer Engagement: Hurricanes and Squirrel Hibernation Cycles." Presented at the 21st International Conference of Auditory Display (ICAD 2015), July 8-10, 2015, Graz, Austria.
- [7] E. Post, *Ecology of Climate Change: The Importance of Biotic Interactions*. Princeton, NJ: Princeton University Press Monographs in Population Biology, 2013.
- [8] J. T. Kerby, E. Post, "Advancing plant phenology and reduced herbivore production in a terrestrial system associated with sea ice decline," *Nature Communications*, vol. 4, article 3514, October 2013. doi:10.1038/ncomms3514.
- [9] E. Post, U. Bhatt, C. Bitz, J. Brodie, T. Fulton, M. Hebblewhite, J. T. Kerby, S. Kutz, I. Stirling, D. A. Walker, "Ecological consequences of sea ice decline," *Science*, vol. 341, issue 6145, pp. 519-524, August 2013.
- [10] E. Post, P. S. Bøving, C. Pedersen, M. A. MacArthur, "Synchrony between caribou calving and plant phenology in depredated and non-depredated populations," *Canadian Journal of Zoology*, vol. 81, no. 10, pp. 1709-1714, October 2003.
- [11] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.
- [12] J. H. Flowers, "Thirteen Years of Reflection on Auditory Graphing: Promises, Pitfalls, and Potential New Directions." Presented at the 11th International Conference of Auditory Display (ICAD 2005). July 6-9, 2005, Limerick, Ireland.
- [13] <http://www.polar.psu.edu/phenology-caribou-muskox-sonification/>
- [14] Arctic Plant Phenology Learning through Engaged Science (APPLES). bit.ly/arcticapples

BINAURAL SPATIALISATION OVER A BONE CONDUCTION HEADSET: ELEVATION PERCEPTION

Amit Barde, Gun Lee

HIT Lab NZ,
University of Canterbury,
Private Bag 4800,
Christchurch 8140, New Zealand.
amit.barde@pg.canterbury.ac.nz,
gun.lee@canterbury.ac.nz

William S. Helton

Department of Psychology,
University of Canterbury,
Private Bag 4800,
Christchurch 8140, New Zealand.
deak.helton@canterbury.ac.nz

Mark Billinghamurst

School of ITMS,
University of South Australia,
Mawson Lakes, SA 5095, Australia.
mark.billinghurst@unisa.edu.au

ABSTRACT

Preliminary results from an on-going experiment exploring the localisation accuracy of a binaurally processed source displayed via a bone conduction headset are described. These results appear to point to decreased localisation accuracy in the horizontal plane when the vertical component is introduced. There also appears to be a significant compression in the area directly in front of the observer $\pm 15^\circ$ in elevation from 0° . This suggests that participants tended to localise stimuli presented at elevations greater than and less than $\pm 30^\circ$ within a 30° 'window' extending 15° vertically either above or below the horizontal plane defined by the 0° azimuth. The results gathered until now suggest that binaural spatialisation over a bone conduction headset can also reproduce the perception of an elevated source to an acceptable degree of accuracy.

1. INTRODUCTION

Current forms of mobile spatial auditory displays almost all rely on delivery of sound through headphones or earphones. These mediums of sound delivery isolate the ears from the ambient acoustic environment [1]. Besides this, most spatial auditory displays are only capable of providing very basic information such as the arrival of a message, email etc. This is achieved by the use of basic tones or sounds. Inevitably, the user of such a display is forced to engage in a visual interaction with the mobile device in order to retrieve information he/she has been conveyed has arrived [2].

With the widespread availability of information 'on-the-go', the need for spatial auditory displays has become greater. It is now necessary for mobile auditory displays to provide more than just alerts for incoming messages and emails. The mobile auditory display of the future must be seamlessly integrated into a wearable computing system capable of delivering useful and actionable information. For example, a wearable auditory display must not only inform a person of incoming message, but also be able to provide navigation information via a binaurally spatialised auditory beacon. Such

functionality incorporated into an auditory display will be able to reduce the cognitive load that current visually demanding mobile displays exert on their users [2] [3].

In addition to the problem of too much information attempting to be displayed on ever shrinking screens of wearable interfaces, there also exist safety concerns. Mobile devices that constantly attempt to engage the visual faculty may end up being a distraction and divert the user's attention away from the primary task. If that task is an attention critical one such as driving or navigation in hazardous environments, the risk posed to the user is great. The lack of attention to the primary task could prove to be fatal in either of these situations. While spatial auditory displays have been developed in response to the challenges posed by visually demanding displays, they suffer from the issue of sensory deprivation. Most spatial auditory displays involve the use of headphones or earphones to deliver auditory information to the user. The use of these mediums to deliver the sound isolates the user from the ambient acoustic environment by covering the ears or blocking the ear canals [1]. This isolation from the acoustic environment is undesirable, since a lot of our information about the environment outside of our visual field is gathered via the auditory faculty. There is a need to develop auditory displays that allow us to retain our natural acoustic perception of the surroundings while simultaneously being able to provide synthesized auditory cues for information presentation and retrieval. The bone conduction headset (BCH) makes for an ideal candidate for such an auditory display. Its relatively small size and the fact that it does not obstruct the pinnae or the ear canals are design aspects that work in its favour.

We are currently carrying out an experiment as part of larger study to explore the feasibility of the BCH as an auditory display device. In the following sections we will cover the research that has been carried out in to the use of a BCH as auditory display device, the design and execution of our study and preliminary results from the study and what they appear to suggest about auditory perception over a BCH.

2. RELATED RESEARCH

While bone conduction technology has been around for a long time, it is only now gaining some ground, both in the research and commercial fields. There is relatively little known about auditory perception of binaurally spatialised sound sources over a BCH. With a few notable exceptions [1] [4] [5], there are few studies that have closely evaluated binaural spatialisation and localisation performance over a BCH. Our previous study has shown that binaural spatialisation over a BCH induces an acceptable level of externalisation and that localisation performance is within parameters acceptable for a such an auditory display [6].

Up until now research related to the use of the BCH has primarily been restricted to its use as a navigation aid for the visually challenged [1] [7]. Few researchers have explored the use of a BCH as part of an auditory display device for AR or VR environments [8] [9]. A large part of the existing literature also concentrates on the use of individualised HRTFs with BCH based reproduction [4] [10]. McDonald et al's results [4] tend to suggest that the use of individualised HRTFs for BCH based reproduction is able to reproduce spatial resolution that is comparable to or better than that achieved over headphones. Studies with non-individualised HRTFs also have been shown to achieve good results [1] [5]. All this research suggests that the BCH has a great potential for being used as a spatial auditory display device as part of a wearable interface incorporating auditory and visual cues.

The study in to the localisation performance achievable for binaurally presented sources over a BCH is an attempt to explore the limits of the device. Knowledge of the operational limits of the BCH will help with the design and development of an auditory display device that is capable of providing useful information to its user.

3. METHOD

3.1. Apparatus

An ecologically valid approach has been adopted for this study. We have chosen to use inexpensive 'off the shelf' hardware and software components that are representative of those used at the developers' and consumers' ends. The experiment was developed in and run using the Unity3D engine [11]. Binaural spatialisation was achieved using the 3Deception [12] plugin for Unity. The binaural engine was selected after an exhaustive pilot study comparing three popular binaural engines available in the market at the time the experiment was conceived. For the BCH we have used the Aftershokz Sportz3 [13]. This combination of hardware and software is a step away from the traditional form virtual auditory display studies that incorporate the use of individualised HRTFs [4] [10] or non-individualised HRTFs (HRTF databases) [1] [5] [14].

The experiment was developed in the Unity3D environment. Sources were created using the binaural engine and the 'global listener' of the engine was slaved to the main camera in Unity to render a first person perspective to the auditory events in the scene (see figure 1). Sound was delivered to the BCH via a Zoom UAC-2 audio interface. A Dell Inspiron Laptop (Windows 8, 2.2 GHz Intel Core i7) was

used to run the study. The experiment was carried out in a sound proof booth that conformed to the HTML 2045, ISO 8253 and ISO BS EN6189 standards.

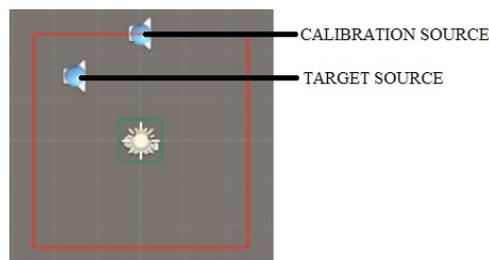


Figure 1: Experimental setup in Unity3D

3.2. Stimuli and Calibration

A single 1 second burst of pink noise (25 ms onset and offset time) was used. Pink noise was chosen since it has been shown that broadband stimulus is easier to localise than one with a restricted spectral range [1] [15] [16] [17]. Calibration of the headset was achieved by asking the participants to adjust the level on the BCH until they felt that it matched the level set on a loudspeaker placed 1m away. During the process participants were asked to look directly at the loudspeaker (PhonicEar AT578-S) and align their heads with it in a manner such that their ears were approximately at the same level as the loudspeaker (see figures 2 and 3). The level of the calibration source played over the loudspeaker was set to approximately 70 dBA measured at 1m from the speaker. The duration and level of the stimulus was chosen to represent that used by previous researchers [1] [15] [18].

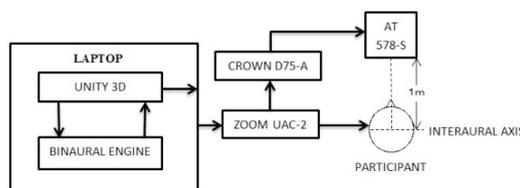


Figure 2: Block diagram of experimental setup

3.3. Participants

We've had 6 participants (1 female, 5 males) between the ages of 19 and 29 (Mean: 23.3, SD: 3.3) take part in the experiment until now. All participants reported normal hearing. No audiometric screening was performed to check for normal hearing function.

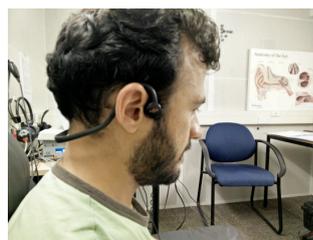


Figure 3: Participant wearing the bone conduction headset

3.4. Procedure

Participants were presented with a single one second burst of pink noise at elevation ranging from -45° to +45° in steps of

15°. This resulted in a total of seven elevation angles. The vertical range within which the stimulus was spatialised was limited since pilot testing showed severe degradation in azimuth perception for vertical angles greater than $\pm 45^\circ$. This is likely due to the fact that azimuth, like longitudes, are compressed as they approach the poles of the imaginary sphere surrounding the participants' heads within which the source is spatialised.

In the horizontal plane (azimuth), locations varied from 0° - 90° i.e. single quadrant size. Step sizes for the azimuth was also 15° resulting in a total of 7 azimuth locations per quadrant. Each elevation was reproduced twice for every azimuth giving us a total of 14 trials per elevation. 98 trials were conducted for every quadrant (14 trails/elevation x 7 azimuth) giving us a total of 392 trials per participants encompassing a complete 360° range in the azimuthal plane.

Trials were divided in to three blocks consisting of a trial block not exceeding 5 minutes and 2 main blocks of trials separated by a 10 minute break. Participants were told to use the two response charts provided to localise the stimulus (see figure 4). There was no compulsion to look or point at the chart to give the response. Positions were to be called out using the signed angles protocol displayed on the response charts. This method of judgement estimation in localisation studies has been validated by previous studies [19] [20] [14] [15]. Participants were asked to face forward during the experiment, and try and keep their head in line with the loudspeaker used for calibration (see figure 3). This wasn't strictly enforced though. No chin brace was employed either to keep the head in a fixed position.

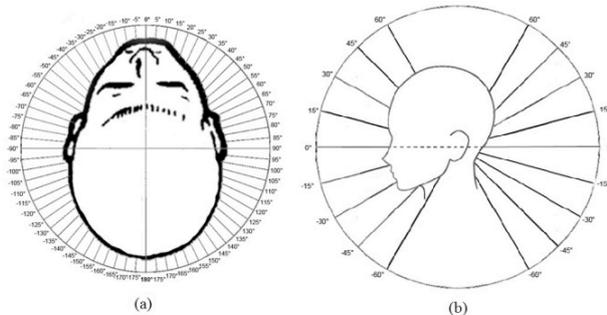


Figure 4: Response charts: azimuth (a) and elevation (b)

Participants were also asked to rate externalization at the end of the experiment. These ratings were based on a method previously employed by Stanley [10] and Gardner [15] in their experiments.

4. RESULTS

Several standard components of localisation such as angular deviation from the source, front-back confusions etc. were measured. The standard front-back and up-down division of the listening space around the user was applied. The division was based of the interaural axis passing through the ear (see figure 5). Participants displayed established phenomena of reversals in the azimuth and elevation (see figure 6). Approximately 82% of the trials in the front resulted in the stimulus being localised to the rear. For trials in which stimulus was presented to rear, only 4% of these were localised to the front. These results are similar to the ones demonstrated in [6], with the exception that the back-front reversals appear to be almost 10% lower in this study.

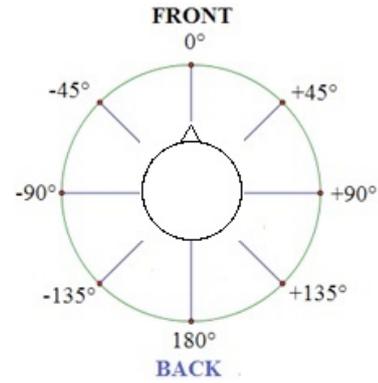


Figure 5: Front-Back division of the auditory space. Line joining -90° and $+90^\circ$ is the interaural axis

Up-down and down-up confusions were also observed during this experiment. Up-down confusions refer to the misrepresentation of a sound source as being below the interaural axis when it is in actuality above it (see figure 6). This is generally caused when an inaccurate representation of the sound source, particularly in the spectral domain, is rendered. This is generally known to occur with non-individualised HRTFs. Down-up confusions can be looked up as phenomena that are exactly opposite of up-down confusions. In this study up-down confusions in the front, -90° to $+90^\circ$, were approximately 24%. At the rear this rate dropped to about 22%. Down-up confusions on the other hand were relatively low. In front they occurred in approximately 5% of the trials for the front and 7% of the trials for the back.

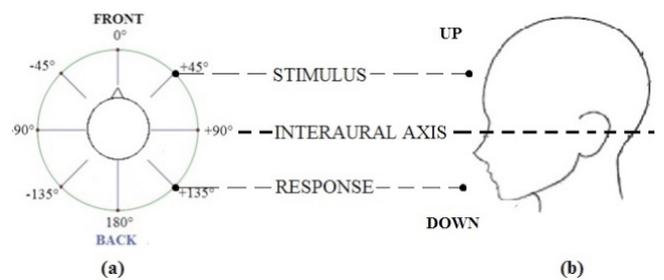


Figure 6: Confusions: Front-Back (a) Up-Down (b)

Angular deviation in the horizontal plane appeared to suffer significantly in comparison with [6]. An average angular deviation from the target of 44.9° was observed for the front. That rose to 51.6° for the rear. These results though appear inflated due one participant who we consider a bad localiser. If this participant's results are excluded, we get an average deviation of 38.9° for the front and 47° for the rear. In the vertical plane, an average error of 21° for front and 20° for the rear was observed. Angular deviation for the horizontal and vertical planes was calculated after resolving confusions. These preliminary results also appear to demonstrate a 'compression' for elevation estimations between -30° and $+30^\circ$. A large number of trials across all elevations appear to consistently be localised within 15° , top and bottom, of the interaural axis. A more in-depth analysis also shows that early angular deviation results appear to be well correlated with those obtained by Wenzel et al. [14] for high and low elevation across the front, side and back for the headphone condition. Four off of the six participants reported externalization. Two of these four participants reported the stimulus to be located at a distance of 1m or more from the surface of the head.

5. DISCUSSION

While the results indicated here are just a preliminary evaluation of an on-going study, it seems that the addition of a vertical component appears to result in poorer localisation when compared to [6]. The ‘compression’ of the localisation within a relatively small area in front appears to be an interesting phenomenon, possibly driven by an evolutionary adaptation. The level of externalisation reported by participants though is encouraging. Based on these preliminary results we could possibly recommend that the element of height not be incorporated in to binaural spatialisation over a BCH. This is because the addition of a vertical component to a task which previously had only requested azimuth ratings appears to result in poorer measured localisation performance when compared to [6]. However, a full analysis of the results and comparisons with existing BCH and headphone studies needs to be carried out to judge the efficacy of the BCH in being able to reproduce a convincing percept of elevation.

6. REFERENCES

- [1] Walker, B.N. and J. Lindsay, *Navigation Performance In A Virtual Environment With Bonephones*, in *International Conference On Auditory Displays*2005: Limerick, Ireland.
- [2] Brewster, S. and A. Walker, *Non-Visual Interfaces For Wearable Computers*. COLLOQUIUM DIGEST-IEE, 1999.
- [3] Walker, A. and S. Brewster, *Spatial Audio In Small Screen Device Displays*. Personal Technologies, 2000. **4**(2-3): p. 144-154.
- [4] MacDonald, J.A., P.P. Henry, and T.R. Letowski, *Spatial Audio Through A Bone Conduction Interface*. International Journal Of Audiology, 2006. **45**(10): p. 595 - 599.
- [5] Lindeman, R.W., H. Noma, and P.G.d. Barros, *Hear-Through and Mic-Through Augmented Reality: Using Bone Conduction To Display Spatialized Audio*. 2007.
- [6] Barde, A., et al., *Binaural Spatialisation over a Bone Conduction Headset: Minimum Discernable Angular Difference*, in *Submitted to the 140th Convention of the AES*2016: Paris, France.
- [7] Walker, B.N. and J. Lindsay, *Navigation Performance With A Virtual Auditory Display: Effects Of Beacon Sound, Capture Radius and Practice*. Human Factors: The Journal of the Human Factors and Ergonomics Society, 2006. **48**(2): p. 265 - 278.
- [8] Valjamae, A., et al., *Binaural Bone Conducted Sound In Virtual Environments: Evaluation Of A Portable, Multimodal Motion Simulator Prototype*. Acoustical Science And Technology, 2008. **29**(2): p. 149 - 155.
- [9] Villegas, J. and M. Cohen, *GABRIEL: Geo-Aware Broadcasting For In-Vehicle Entertainment And Localizability*, in *AES 40th International Conference*2010, AES: Tokyo, Japan.
- [10] Stanley, R.M., *Measurement And Validation Of Bone Conduction Adjustment Functions In Virtual 3D Audio Displays*, in *School of Psychology*2009, Georgia Institute of Technology.
- [11] Helgason, D., N. Francis, and J. Ante, *Unity3D*, 2005, Unity Technologies: Copenhagen, Denmark.
- [12] Thakur, A. and V. Nair, *3Deception*, 2015, Two Big Ears: Edinburg, Scotland.
- [13] Aftershokz. *Sportz3*. [cited 2015 2nd June]; Available from: <http://aftershokz.com/collections/wired/products/sportz-3>.
- [14] Wenzel, E.M., et al., *Localization using nonindividualized head-related transfer functions*. The Journal of the Acoustical Society of America, 1993. **94**(1): p. 111-123.
- [15] Gardner, W.G., *3D Audio Using Loudspeakers*, in *School of Architecture and Planning*1997, Massachusetts Institute of Technology.
- [16] Stevens, S.S. and E.B. Newman, *The Localization of Actual Sources of Sound*. The American Journal of Psychology, 1936. **48**(2): p. 297 - 306.
- [17] Weinrich, S.G., *Horizontal Plane Localization Ability and Response Time as a Function of Signal Bandwidth*, in *98th Convention of the AES*1995: Paris, France.
- [18] Wersenyi, G., *Localization In A HRTF Based Minimum Audible Angle Listening Test On a 2D Sound Screen For GUIB Applications*, in *115th Convention of the Audio Engineering Society* 2003.
- [19] Wightman, F.L. and D.J. Kistler, *Headphone Simulation Of Free Field Listening. II: Psychophysical Validation*. The Journal of the Acoustical Society of America, 1989. **85**(2): p. 868 - 878.
- [20] Wenzel, E.M., F.L. Wightman, and D.J. Kistler, *Localization with non-individualized virtual acoustic display cues*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*1991, ACM: New Orleans, Louisiana, USA. p. 351-359.

SONIFICATION, MUSIFICATION, AND SYNTHESIS OF ABSOLUTE PROGRAM MUSIC

Allan D. Coop

P.O. Box 5160, Braddon, 2612, Australia
allan.coop@gmail.com

ABSTRACT

When understood as a communication system, a musical work can be interpreted as data existing within three domains. In this interpretation an absolute domain is interposed as a communication channel between two programmatic domains that act respectively as source and receiver. As a source, a programmatic domain creates, evolves, organizes, and represents a musical work. When acting as a receiver it re-constitutes acoustic signals into unique auditory experience. The absolute domain transmits physical vibrations ranging from the stochastic structures of noise to the periodic waveforms of organized sound. Analysis of acoustic signals suggest recognition as a musical work requires signal periodicity to exceed some minimum. A methodological framework that satisfies recent definitions of sonification is outlined. This framework is proposed to extend to musification through incorporation of data features that represent more traditional elements of a musical work such as melody, harmony, and rhythm.

1. INTRODUCTION

A conceptual framework is proposed for the organization and description of relationships between a musical work, digital data, and sound. The framework encompasses and organises digital data and acoustic signals that range from noise to musical sound. The goal is to employ this framework to develop a methodology suited to guide the creation and evolution of acoustic signals through sonification to achieve musification.

To this end, acoustic signals, can be better understood by, (i) modelling them as components of a communication system, and (ii) quantitatively locating such signals within the compass of their stochastic and periodic components. It is hypothesised that this is best achieved by employing both basic measures of communication theory such as the efficiency or redundancy of signal components and statistical features of the autocorrelation function of an acoustic signal. In doing so, a methodology is identified whereby any digital content or data stream can be evolved into acoustically organized musical experiences.

Music has a long and complicated history of development in the analog domain. However, its relationship to, development, and analysis within the digital domain is now the focus of intense study. This digital domain is a result of phenomenal advances in computer and communications technology made possible by the continued elaboration of integrated circuit processing technology.



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

It is a technology that has increasingly replicated the functionality of traditional analog devices, including the digital replication of acoustic signal sources such as, for example, those of musical instruments. The impact of communication theory on these developments has been immense [1].

Complex computer-based implementations of digital hardware provide many advantages; programmability, flexibility, additional product functionality such as analog to digital conversion (and vice versa), short design cycles, and good immunity to both noise and manufacturing process tolerances [2].

Such technological change has propelled development and research in sonification [3, 4] and musification [5, 6, 7]. Sonification was originally described as, the use of synthetic non-verbal audio to support information processing activities [8]. Subsequently, sonification has been employed to, (i) transform the inaudible into the audible, (ii) employ audition to gain knowledge, and (iii) develop listening techniques for scientific inquiry [9]. It has more recently been defined as a technique that uses data as input and generates sound signals, with the caveats that, (i) the sound reflects objective properties or relations in input data, (ii) the transformation is systematic, (iii) the sonification is reproducible, and (iv) the system can intentionally be used with different data [10].

Musification has been defined as the musical representation of data [7]. It is designed to go beyond direct sonification and include elements of tonality and the use of modal scales to create musical auralizations. The resulting musical structures take advantage of higher-level musical features such as polyphony and tonal modulation in order to engage the listener [11].

More recently it has been proposed that full realization of the potential of sonification must also include, “the craft and art of music composition” [5]. It is in this sense that here threads of this “second order of intervention” [5] are explored. The aim is to, (1) introduce elements of a compositional framework and methodology for the evolution of sonification into musification as a means of artistic expression and (2) better understand powerful, poorly defined, and still under-examined aspects of organized sound.

2. THEORY

Two propositions inform the framework developed here. (1) Messages within a communication system are transmitted through one or more channels connecting a source to a receiver [1]. (2) A received message exhibits two aspects, one is semantic (having a universal logic, structured, articulable, translatable, and serving the behaviorist conception of action preparation), the other esthetic (untranslatable as there is no language available to translate it to, it refers to the repertoire of knowledge associated with the particular communication system) [12].

The framework is then evolved from the proposal that two axiomatic and mutually exclusive domains together form the organi-

zational foundation of a musical work (c.f. [13]). One domain is internal, cognitively located, subjectively individual, and referred to as programmatic. The other is external, purely physical, collectively objective, quantifiable, and referred to as absolute. This latter domain essentially involves nothing but a section of the theory of the motions of elastic bodies [14].

The internal programmatic domain is physiological and psychological. It comprises the totality of the abstract, conceptual aspects of a musical work, including its intellectual creation and non-acoustic or symbolic representations, whether memorised or physically recorded. Within this domain are found the responsible originating creative processes which are employed to generate, organise, produce, and subsequently interpret and re-fabricate acoustic signals into a musical work. Due to the extreme scale and complexity, thus difficulty, of investigating the processes involved, little is currently understood about the details by which these functions and behaviors are realized (see, for example [15]).

The programmatic domain extends until physical vibration is initiated in the external world. It is this acoustic signal which forms the absolute component of a musical work. It originates as the programmatic component is transferred into the absolute domain. The resultant waveform exists as a temporal sequence of physical vibrations in the world. It ranges from one that: (a) Can be resolved via Fourier's theorem [16] into one or more sine waves each exhibiting a single frequency, to an acoustic signal that is essentially noise as it, (b) exhibits stochastic repetition in the absence of correlation of amplitude or interval of succession [17]. In contrast with the programmatic domain, much detail is known about the physical aspects of the absolute domain (see, for example [18]).

In other words, as a musical sound an acoustic signal is primarily comprised of a structured complex of periodic waveforms that exhibit a non-zero autocorrelation function. It is in this general sense that music may be thought of as organised sound (as proposed by [19]). Alternatively, as noise, an acoustic signal is primarily stochastic and exhibits an autocorrelation function which approaches zero.

Described in this way, it is clear that absolute and programmatic components may be combined to form the complex acoustic signal instantiatable within individual musical works. They originate and exist within two individual and mutually exclusive domains, each required for the complete existence, communication, and experience of a musical work.

Such an interpretation allows a musical work to be characterised as a communication system (figure 1). In this view, a musical work is revealed as a triadic sequence. This sequence is, (A) initiated as a cognitive creation which is then evolved and elaborated as a programmatic component within the programmatic domain. At some moment, this component is, (B) physically instantiated and transmitted acoustically as the absolute component of a musical work within the absolute domain. This transmission may subsequently be received, (C) to cognitively be reconstructed as a programmatic component within the programmatic domain with the potential to animate the experience and behavior of a listener.

Within this communication system, an information source, here the cognitively generated programmatic component of a musical work, is operated on by a transmitter; for example, human cognitive neural activity may be converted into motor activity to generate and transmit an acoustic signal either vocally or with the aid of an instrument. This signal physically propagates as the absolute component of a musical work to any receiver or human listener. Successful reception of the acoustic signal leads to the

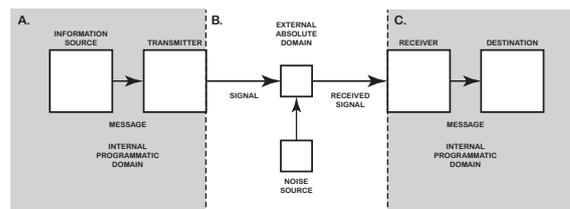


Figure 1: Relationship between the components of a musical work. A: The internally located programmatic domain is the cognitive source of the programmatic component of a musical work. This component is instantiated and released as an acoustic signal into the world by the transmitter. B: The acoustic signal comprises the absolute component of a musical work and propagates externally in the world until it either dissipates or is detected by a receiver. C: At the receiver, the absolute component is captured and reconstructed along with any noise to form a novel programmatic component prior to reaching the cognitively located destination of conscious perception. Note: A and C may be the same individual or one or more nonbiological devices.

inverse operation done by the transmitter being performed on the signal whereby the absolute component is cognitively re-fabricated as a programmatic component to exist and be experienced as a new version of the original pretransmission programmatic component.

It is further proposed that this communication system has increasingly come to be located between two mutually exclusive physical domains. (i) A continuous analog domain embedding the objects of the physical world, including the human brain (or programmatic domain), and more recently, (ii) A discrete mechanical or digital domain which originated with the development of communication technology.

The digital domain relies on the mathematical formalism of communication theory [1]. This theory allows both analog and digital signals to be characterized by measures of entropy and redundancy. Accordingly, a stochastic signal exhibits the highest possible entropy and thus the lowest redundancy. When digitized, such a signal typically exhibits high variability in the record of its sampled waveform. Alternatively, a periodic signal exhibits lower entropy, thus higher redundancy, and the sampled waveform exhibits lower variability.

Autocorrelation is a mathematical tool that can be employed to find and quantify the extent of repeating patterns, particularly the periodicity of a signal. It provides a measure of the similarity between observations or digital samples as a function of the lag between them. A quantitative estimate of signal periodicity (the autocorrelation coefficient) can be obtained from the autocorrelation function (the normalized autocovariance function [20]).

3. SONIFICATION

In the case of sonification, it is possible to base an organisational framework for a musical work on reinterpretation of pre-existing musically related ideas. Thus, for example, when considered as a song or an ode, music has been claimed to consist of three parts—the words, the melody, and the rhythm [21]. However, here the claim is that the only component suited for direct translation by sonification is that of the symbols of a data stream. Although, melodic and rhythmic components might be extracted from a data stream; for the purposes of sonification, it is hypothesised they are

properly considered more abstract second order properties, as they are not events in the same primary way that the numeric values or symbols comprising a data stream are. It is for this reason that melody and rhythm are considered likely more properly left to be treated as components of musification.

3.1. Implementation

The framework and methodology proposed here are claimed to satisfy the four formal requirements for sonification. For the given example these are, (i) the objective data properties of sample redundancy and pitch frequency, (ii) the systematic transformation of data sequences to redundancy and pitch frequency, (iii) the reproducibility introduced by a mathematical model, and (iv) the applicability of the methodology to alternative data sets. The primary motivation for the approach is the apparent lack of general principle suited to guide the sonification process.

While keeping the caveats with regard to periodicity and variability introduced above in mind, importantly, the method now outlined may be generalized to convert and organise any digitized content prior to evolving it to musification. Here, it is developed for application to a symbol stream such as might be found in a file containing digitized text.

The computational implementation proceeds in two steps. (1) A series of partial redundancy values are mapped to a sequence of pitch frequencies. (2) The derived pitch frequencies can then be mapped back to refabricate the original sequence of symbols as a sequence of sounds.

For this analytic approach the absolute component of a given communication system (the data) is assumed to be constrained to the scope of a clearly defined symbol set with ergodic properties.

If this is accepted, the mathematical formalism of communication theory defines the Shannon entropy (h) given in bits as

$$h = - \sum_{i=1}^N p_i \log_2(p_i), \quad (1)$$

where p_i is the probability of the i th value in a data set comprised of N unique values.

As entropy is an extensive quantity, a ‘normalized’ entropy measure is employed to allow more meaningful comparison of different data sets. This measure, the redundancy (r), is calculated as

$$r = 1 - \frac{h}{H}, \quad (2)$$

where H is the maximum or Hartley entropy [22] of the acoustic signal, calculated as

$$H = \log_2(N). \quad (3)$$

3.2. Letters and words

The partial entropy (h_i) of each symbol is obtained from [23]

$$h_i = -p_i \log_2(p_i), \quad (4)$$

while the partial redundancy (r_i) is determined by direct substitution of h_i for h in [2].

Once calculated, the partial redundancy can be employed to order, as required, the unique symbols within a data set.

In the case of data such as text, although not required, one initial simplification is to assume that individual letters and words

are uncorrelated. Here, the letters of a text source can be directly quantified from [1–3].

The probability of each of the letters forming an individual word can be employed to calculate the partial entropy of a given word from

$$h_m = \prod_{j=i}^n p_j \log_2(p_j), \quad (5)$$

where j gives the letter number and n the number of letters in the m th word in a list of unique source words. After substitution, the word probabilities of the originating data set are re-normalized prior to redundancy calculations via equations [1–3].

Alternatively, word probabilities can be obtained directly from a list of source word frequencies obtained from the given data set. For a sufficiently large sample the difference in the partial entropy calculated for words, either from individual letters or for the words them-self, indicates the extent to which letters within a given word are correlated. This provides an estimate of the bias introduced if, as a simplification, it is assumed that the letters forming a word are uncorrelated.

Autocorrelation analysis employed the `xcorr` function in the `signal` package of GNU Octave version 3.8.0-1 [24]. Raw autocorrelation functions were normalized by the correlation at lag zero using the `coeff` flag. The central peak of each scaled autocorrelation function was removed and the amplitude of the largest remaining positive peak was converted to a percentage and reported.

3.2.1. Defining the digital pitch sequence

It is widely accepted that auditory perception extends from 20–20,000 Hz [18], although, for younger people, the auditory range may extend from 16–25,000 Hz [25]. Humans can sense vibrations with a frequency as low as 2 Hz, although a minimum of about 20 Hz is required for perception of ‘tonality’ [26]. Assuming the amplitude is sufficient, lower frequencies are typically felt through their vibratory effect, rather than heard.

The range of perceptually discriminable frequencies is less extensive than the range of detectable frequencies. Within the 16–16,000 Hz range (16.0 kHz–kilohertz) it has been estimated the human auditory system can distinguish about 1,200 [27] to 1,400 [28] distinct pitch levels. Increasing the range to 20 Hz–20 kHz has little perceptual effect as only about 1,500 pitches may be discriminated [29].

For comparison with human audition, one of the largest frequency ranges of any instrument—the piano—has 88 semitones which range from 27.5 Hz–4.185 kHz, while the reproduction of orchestral music with subjectively perfect fidelity requires a frequency range of 40 Hz–15 kHz [30].

For the purposes of developing an absolute pitch sequence for the digital domain, the human sensory capacity can be assumed to extend from silence (by definition 0 Hz) to vibrations with a frequency of 16.384 kHz. For both convenience and simplicity, a lower bound of 1 Hz is chosen (more detail is given below). The upper bound is chosen for several reasons. It is known that human audition deteriorates with age by loss of higher frequencies. The ability to discriminate frequencies is greatly reduced above about 15 kHz, and an upper bound of 16.384 kHz can conveniently be expressed as 2^{14} , where the exponent immediately gives the octave count from 1 Hz.

More technical reasons for the choice of upper bound include, (1) positive powers of 2 are important in computer science—there

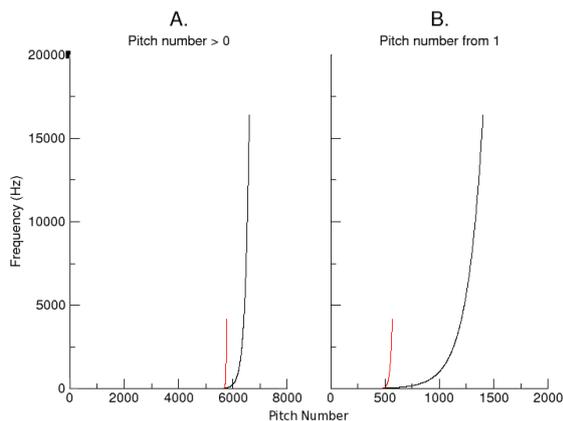


Figure 2: Relationship between absolute and unitary pitch sequences. A: Absolute (black). Frequency profile for 6,601 pitch numbers obtained when the absolute digital pitch is calculated from silence (>0 Hz) to 16.384 kHz. B: Unitary (black). Frequency profile for 1,401 pitch numbers when the unitary digital pitch is calculated from ≥ 1 Hz. A, B: Classical (red). Frequency profile for each of the 88 different pitches represented by the piano keyboard. It has been transposed to match the equivalent frequency range of the associated digital pitch sequence. See text for more details.

are 2^n possible values for a n -bit binary register, (2) more importantly, $\log_b(n)$ bits are required to represent a variable that can take one of n values if n is a power of b (when $b = 2$ the units of Shannon entropy are commonly referred to as bits), (3) the logarithm of a probability distribution is useful as a measure of entropy as it is additive for independent sources, and (4) such an upper bound gives a range of 14 octaves which is similar to the 10 octaves obtained if a frequency range of 20 Hz–20.48 kHz or 16 Hz–16.38 kHz were to be assumed.

The foregoing considerations have led to the selection of frequencies suited for fabrication of what is referred to as a digital pitch sequence. Based on its starting frequency, this sequence may be defined as either absolute (0 Hz–16.384 kHz) or unitary (1 Hz–16.384 kHz).

The relation of these two pitch sequences to the classical pitch sequence, as given by a tuned piano, is illustrated in figure 2. This figure shows the extension of the classical range of pitch sequence to the absolute (A) and unitary (B) pitch sequences.

The absolute pitch sequence illustrated in panel (A) shows that almost 85 % (84.9 %) of frequencies are less than 16 Hz. Alternatively, panel (B) illustrates that, for the unitary sequence, an increase in the start of the sequence from 0 Hz to 1 Hz results in more than 70 % (71.4 %) of the frequencies being greater than 16 Hz.

3.2.2. Mapping redundancy to frequency

In the simplest case, once obtained, the first m partial redundancy values for either words or letters can be sorted into an ordered list and mapped to the equivalent number of values obtained, as appropriate, from either the absolute or unitary frequency ranges or, more conventionally, the Classical pitch sequence of the piano keyboard. Alternatively, partial redundancy values could be mapped

to either the related notes of one or more instruments or to computationally fabricated timbres.

The second and final step in the methodology is to complete the sonification process by mapping the partial redundancy \rightarrow frequency relations back into the sequential order of the original data sequence.

4. MUSIFICATION

The methodological implementation of musification is concerned with both the absolute and programmatic components of a musical work. In the development of absolute program music, the sound content is structured by reference to communication theory. This mathematical formalism is employed as a tool to assist with translation and organisation of the symbols of a digital data set into organized musical sound.

4.1. Melody (and harmony)

A simple melodic component is automatically generated when the symbols of a data sequence (here either letters or words) are mapped to a given pitch sequence. A least two considerations should be taken into account prior to performing this mapping. The first is that frequencies of higher pitch are more difficult to discriminate than frequencies of lower pitch. The second is that the higher the redundancy of a symbol, the less information it communicates. Thus, to transparently increase human interpretability, signals with a high rate or frequency should exhibit less variability (higher redundancy) than signals with a low rate or frequency.

As a consequence, one musification option is to map the most redundant symbols to the highest note pitches or frequencies and vice versa. Although, as for the steps given above, many other mappings might be employed.

An alternative approach is to generate a harmonic component by forming a mapping between pitch and the letters within each word. Here, the sequential partial redundancy values related to individual letters can be mapped to a given pitch sequence. The letters in each word can then provide the synchronous events that comprise individual ‘chords,’ i.e. a chord is given by the letters forming each successive word in a text stream.

An important step in moving from the sonification to the musification of a data stream is the addition of the relevant harmonic series to each note. Further, the acoustic properties of a given note can also be enhanced by the addition of a temporal envelope for the frequency wave forms comprising that note (see for example figure 3).

4.2. Rhythm

In the given example, the most direct source of rhythm is that of the poetic meter of the words in the source text. This can be characterized and applied to organise the rhythmic components of the musification process. Alternatively, where there is no obvious metric source, as might particularly be the case for non-linguistic data, the intervals between repeated symbols provide one immediate option which could be employed to organize temporal patterns. The modulation of these intervals can further be manipulated by the introduction of temporal envelopes. The effect of such a modulation is illustrated in figure 3 for a 1 Hz sine wave. The addition of an exponential temporal envelope reduces the effective duration of the sound from 7 s to about 4.5 s. It is noted that many

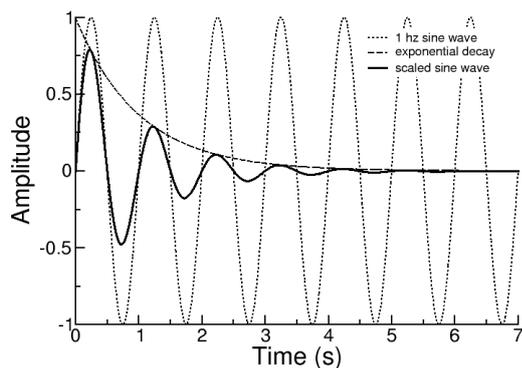


Figure 3: A simple musification technique illustrated by the effect of a temporal envelope on the magnitude and duration of a given waveform. An exponential decay (dashed line) has been applied to a sine wave (dotted line) with a frequency of 1 Hz and duration of 7 s. The resultant waveform is indicated (unbroken line). The audible duration of the modulated sound is reduced from 7 s to about 4.5 s when compared with the duration of the unmodulated sine wave.

other forms of temporal envelope are available for the generation and implementation of rhythmic control beyond the duration of the raw unmodulated sound associated with sonification.

5. RESULTS

To characterize an acoustic signal, the redundancy and the autocorrelation coefficient of each data set, as respective measures of signal variability and signal periodicity, can be calculated and plotted.

Figure 4 illustrates such a plot for a selection of data sets that represent examples of different sounds and musical genres (see symbol \times). As might be expected for a near uniform distribution of white noise (a), the entropy is close to the Hartley entropy and consequently the redundancy is vanishingly small (0.03 %). In the absence of the zero lag peak, the autocorrelation coefficient is also everywhere near zero (largest value 0.011 % at a random lag value of 47.1 s).

Alternatively, a sine wave would be expected to exhibit a lower entropy than white noise due to data clustering at similar values near positive and negative peaks. Thus, a 440 Hz sine wave exhibits a redundancy of 34.7 % while the autocorrelation coefficient is near unity (g).

The remaining analyses of the musical works illustrated in figure 4(b)–(f) form two groups linked by (e) a gregorian chant (*Kyrie Eleison*). One of these groups consists of: (b) a modern electric trio (guitar, bass, and drums), (c) an aria from the *Magic Flute*, and (e) the gregorian chant. These musical works exhibit similar autocorrelation coefficients (respectively 23 %, 23 %, 21 %) but are distinguished by their redundancies (respectively 12 %, 14 %, 20 %). Almost orthogonal to this group is one consisting of (d), (e), and (f). These data respectively comprise one of the first works of musique concrete (*Etude Aux Chemins De Fer*), the gregorian chant, and a minimalist work for piano and violin (*Spiegel Im Spiegel*). These three musical works exhibit

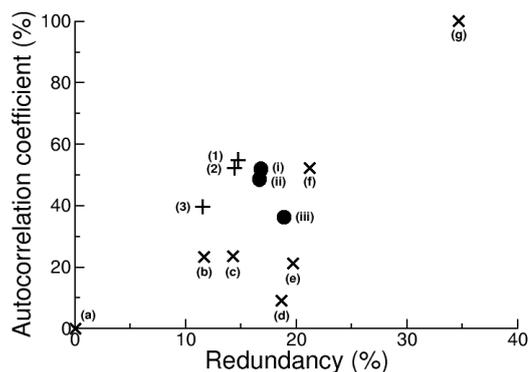


Figure 4: Relationship between the redundancy and autocorrelation coefficient of an acoustic signal. Symbol \times : (a) Stochastic signal: White noise. Musical works, (b) “*****, *****”: *Study For Falling Apart*, (c) Mozart: *The Magic Flute*, Act II, *Der Hölle Rache*, (d) Schaeffer: *Etude Aux Chemins De Fer*, from *Cinq Études De Bruits*, (e) Gregorian Chant: *Kyrie Eleison*, Benedictine Monks of Santo Domingo de Silos, (f) Pärt: *Spiegel Im Spiegel*. (g) Periodic signal: 440 Hz sine wave. Sonified text (+): (1) Classical pitch sequence, (2) Unitary pitch sequence, (3) Mono pitch sequence (1 + 2). Musified text (●—examples of absolute program music), (i) Unitary pitch sequence, (ii) Classical pitch sequence, (iii) Mono pitch sequence (1 + 2). As the reported values for the autocorrelation coefficient and the redundancy lie in the range 0–1, they are given as percentages. See text for further details.

similar redundancies (respectively 19 %, 20 %, 21 %) but are distinguished by their autocorrelation coefficients (respectively 9 %, 20 %, 52 %).

For comparison with the foregoing results, figure 4 also provides results of an analysis of the sonification (+) and musification (●) of a text (the song of Orpheus as reported in [31]). These data (1–3 and i–iii, respectively) exhibit similar periodicity and redundancy to the previously described musical works (a–g). Musification has little effect on the periodicity data when it is sonified (1 c.f. i, 2 c.f. 2, 3 c.f. iii), but increases the redundancy of the signal, particularly when the individual data sets for unitary and classical pitch sequences are combined to form a single acoustic sequence (see figure 4, 3 c.f. iii, and 1 c.f. i and 2 c.f. ii).

6. DISCUSSION

Initially, the idea of a musical work being an instantiation of a communication system is developed. Consequently, a complete musical work can be conceptualized as a triadic sequence comprised of a creatively fabricated programmatic component, a physically propagated absolute component, and a consciously perceived cognitive re-fabrication of the transmitted absolute component that provides a subjectively unique programmatic component for each receiver or listener. These metaphorical descriptors (“programmatic” and “absolute”) are respectively chosen because cognitively a sound may be associated with, for example, any one or more colors, tastes, or emotions, whereas, in the external physical world sound is always an absolute consequence of only one type

of event, atmospheric vibration.

In the absolute domain of the communication channel, it is hypothesised that an auditory signal can be partitioned by two quantitative measures. One is the periodicity and the other the variability of the acoustic waveforms. These may respectively be related to the 'semantic' and the esthetic properties of a given signal, with the type of source and the path taken by an acoustic signal determining the relative contribution made by each of these two measures. Thus, the musicality of an acoustic signal may be determined by the emergence of the organizational periodic or 'semantic' component from the more anarchistic esthetic or stochastic component (see [32]).

It is clear from this view, that the absolute and programmatic components must be appropriately independent and sequentially arranged for a complete musical work to exist. Within the context of the reported framework, the digital record of the acoustic signal of several musical works representing different genres was analyzed by calculation of a redundancy measure of Shannon communication theory and the statistically based autocorrelation coefficient.

Results suggest that, with the exception of an acoustic signal formed by a single sine wave, it is the quantification of autocorrelation or periodicity rather than the redundancy or variability of an acoustic signal that better distinguishes between musical and non-musical sound. It is predicted that an acoustic signal will likely start to lose musical character as the autocorrelation coefficient falls towards about 10% signal correlation or periodicity. This result suggests that in the absolute domain the organizing effect of the periodicity of an acoustic signal may be more important than its variability when a human listener is determining whether to interpret a signal as musical sound rather than noise.

As noted above, the meaning of a received acoustic signal is embedded in its cognitively located semantic and esthetic components. A tick signal indicating radioactive particle decay such as that provided by a Geiger counter is typically stochastic and usually distributed as a Poisson random variable. Within the framework presented here, such a signal would in the normal course of events be predicted to be classified as an esthetic phenomenon as it refers to a repertoire of knowledge associated with the given instrument. However, as the sound source is (cognitively) known to be from a particular instrument, the acoustic signal of the ticks becomes predominantly determined by its semantic associations rather than its esthetic nature. Thus, before reaching its cognitive destination, the stochastic signal is ascribed a meaning that becomes intuitively translatable into action—the greater the frequency of ticks, the more dangerous the location. In this circumstance there is likely little practical reason for cognitively reverting attention to the stochastic esthetic component of the signal. As an audibly sonified stochastic signal, it embodies little musicality, but once the correct association is learned it can be cognitively elaborated into environmentally sourced meaning. This is in concordance with a previous observation that aesthetic investigation of the programmatic domain must, above all, consider the beautiful object, and not the perceiving subject [33].

Furthermore, it is notable that in the normal course of events, a sound file containing a digitized acoustic signal of sufficient quality contains much of the information necessary to reconceptualize the essential elements of the original auditory environment from which the sound was obtained. Much of this environmental information, including aspects of force, rate, and material, etc, is represented in a digital sound recording. The point being that humans

can refabricate significant aspects of the original acoustic environment from high quality digital recordings on the basis of innate cortical scene-analyzing functionality [34]. When represented as a time sequence of amplitudes, such complex three dimensional acoustic data becomes directly amenable to quantification by standard analytic methods and measures such as those proposed here.

Importantly, it is hypothesised that for sonification techniques to successfully create transparent and humanly interpretable acoustic signals that are not mistaken for noise, these signals should exhibit a minimal periodicity. Further, when some such lower bound is exceeded, suitable acoustic elaboration of a sound may provide a digital signal sufficient to support musification. This is not to suggest that the current aim is to provide a method that will prove one sonification is more effective than another.

A principled approach to the development of a musical work has also been presented. It implicitly provides a computationally based metaphor and framework for modelling human creative compositional processes. The approach originates in the recognition of what might be thought of as "objective historical markers," the penumbra of opinion surrounding the ideas of the absolute and the programmatic in music. In particular, here an attempt has been made to reconcile these two approaches to musical theory. On the one hand, there is the nature of music, its place in the cosmos, its physiological and psychological effects, and its proper use and cultural value; whereas on the other hand, there is the systematic description of materials and patterns of musical composition.

As expressed in the historical ideas of the absolute and programmatic interpretations of musical works, a proposal has been made for a sonification methodology that satisfies various suggested requirements. The methodology outlined here appears appropriate for this end. It illuminates and enables the integrated development of what might be referred to as a general sonification methodology. It is an approach that provides a principled precursor for data musification and, putatively, the creation of what might be referred to as absolute program music. Importantly, the obverse is not being suggested, i.e. that musicality is necessary for effective sonification. It is rather that, as mentioned in the Introduction, effective sonification should in the first instance at least satisfy the requirement that it employs data as input and generates sound signals (but see the four associated caveats [10]).

To continue along this line, keeping accepted definitions of sonification in mind—at a technical level converting data to sound to gain knowledge, while at a perceptual level developing listening techniques for scientific enquiry—the approach endorsed here provides at least a principled methodology for further exploration of cognitive function, and elaboration of design principles and purpose, within a formal context or framework.

By initiating such a project at the level of basic nuances of human auditory perception, it is hypothesised that the path to several goals may be clarified. In particular, a more sophisticated understanding of (1) how human perception constrains sonification design, and (2) identification of the consequent guidelines for the creation of effective musification.

Finally, it is noted that it is only since the introduction and widespread availability of sophisticated digital technology and tools that a meaningful approach can be made towards more generalized models of the creative development of musical works through sonification and musification. It is hoped aspects of what is presented here will provide a preliminary contribution towards that goal.

7. REFERENCES

- [1] C. E. Shannon, The mathematical theory of communication, in *The Mathematical Theory Of Communication*, C. E. Shannon and W. Weaver, University of Illinois Press, Urbana & Chicago, IL, U. S. A., 1963. pp. 31–125.
- [2] C. Toumazou, J. B. Hughes, and N. C. Battersby, “Introduction,” in C. Toumazou, J. B. Hughes, & N. C. Battersby (Eds.), *Switched-Currents: An Analogue Technique For Digital Technology*, Peter Peregrinus Ltd., London, U. K., 1993, pp. 1–7.
- [3] T. Hermann, A. Hunt, J. G. Neuhoff (Eds.). “The Sonification Handbook,” Logos Verlag, Berlin, Germany, 2011.
- [4] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, and J. Neuhoff, “Sonification report: Status of the field and research agenda,” Tech. Rep., Int. Community for Auditory Display, <http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1443&context=psychfacpub> (accessed 29 February 2016), 1997.
- [5] S. Gresham-Lancaster, “Relationships of sonification to music and sound art,” *AI & Soc.*, vol. 27, pp. 207–212, 2012.
- [6] V. Straebel, “The sonification metaphor in instrumental music and sonification’s romantic implications.” in *Proc. 16th Int. Conf. on Auditory Display*, Washington, DC, U. S. A., June, 2010.
- [7] J. Edlund, “The Virtues of the Musifier: A Matter of View,” <http://musifier.com/index.php/tech/musification-and-view>, (accessed 29 February 2016), 2004.
- [8] S. Barrass, and G. Kramer, “Using sonification,” *Multimedia Syst.*, vol. 7, pp. 23–31., 1999.
- [9] A. Schoon, “Sonification in music”, in *Proc. 15th Int. Conf. on Auditory Display*, Copenhagen, Denmark, May , 2009.
- [10] T. Hermann, “Taxonomy and definitions for sonification and auditory display,” in *Proc. 14th Int. Conf. on Auditory Display*, Paris, France, June 2008.
- [11] F. Visi, G. Dothel, D. Williams, and E. Miranda, “Unfolding Clusters: A Music and Visual Media Model of ALS Pathophysiology,” in *Proc. SoniHED Conf. Sonif. Health Env. Dat.*, York, U. K., 2014.
- [12] A. Moles, *Information Theory And Esthetic Perception*, University of Illinois Press, Urbana, IL, U. S. A., 1966, p. 129.
- [13] L. Goehr, *The Imaginary Museum of Musical Works: An Essay In The Philosophy Of Music*, Clarendon Press, Oxford, U. K., 2007.
- [14] H. L. F. Helmholtz, *Sensations Of Tone As A Physiological Basis For The Theory Of Music*, A. J. Ellis (trans.), London, UK, 1912. p. 3.
- [15] M. A. Arbib, ed., *Language, Music, And The Brain: A Mysterious Relationship*, The MIT Press, Cambridge, MA, U. S. A., 2013.
- [16] J. Jeans, *Science And Music*, Cambridge University Press: London, UK, p. 83, 1961.
- [17] A. Moles, *Information Theory And Esthetic Perception*, University of Illinois Press, Urbana, IL, U. S. A., 1966, p. 81.
- [18] T. D. Rossing (ed.), *Springer Handbook of Acoustics*, Springer, New York, NY, U. S. A., 2007. pp. 747–748.
- [19] E. Varèse, and W. Chou, “The liberation of sound,” *Perspectives of New Music*, vol. 5, no. 1, pp. 11–19, p. 18, 1966.
- [20] D. L. Hartmann, “ATM 552 Notes: Time Series Analysis – Section 6a,” http://www.atmos.washington.edu/dennis/552_Notes_6a.pdf, (accessed 29 February 2016), p. 128, 2016.
- [21] Plato, *The Republic*, Book III, 398, in B. Jowatt (trans.), *The Dialogues of Plato*, Volume III, 3rd ed., The Clarendon Press, Oxford, U. K., 1892. p. 83.
- [22] R. Hartley, Transmission of information. *The Bell Syst. Tech. J.*, 1928, 7: 535–563.
- [23] T. Xu, Chiu D and I. Gondra, “Constructing target concept in multiple instance learning using maximum partial entropy,” in P. Perner (ed.), *Machine Learning and Data Mining in Pattern Recognition*, Springer, Berlin, Germany, 2012, pp. 169–182.
- [24] J. W. Eaton, D. Bateman, and S. Hauberg, *GNU Octave Version 3.8.0 Manual: A High-Level Interactive Language For Numerical Computations*, CreateSpace Independent Publishing Platform.
- [25] R. Cogan & P. Escot, *Sonic Design: The Nature of Sound and Music*, Prentice Hall, Inc., Englewood Cliffs, NJ, U. S. A., 1976. p. 442.
- [26] S. Gelfand, *Hearing: An Introduction To Psychological And Physiological Acoustics*, Marcel Dekker Inc, New York, NY, U. S. A., 1998. p. 281.
- [27] J. E. Cohen, Information theory and music, *Behav. Sci.*, 1962, vol. 7, no. 2, pp. 137–163.
- [28] [25], p. 444.
- [29] D. Crystal, *Prosodic Systems And Intonation in English*, Cambridge University Press, Cambridge, UK, 1969. p. 110.
- [30] H. F. Olsen, *Music, Physics, and Engineering*, 2nd ed., Dover Publications, Inc, New York, NY, U. S. A., 1967. p. 383.
- [31] Ovid, The story of Orpheus and Eurydicé, in *Ovid’s Metamorphoses*, Book X. J. Dryden (trans.), Proprietors of the English Classics, London, UK, 1826. pp. 246–274. pp. 246–247.
- [32] E. Wind, *Art and Anarchy*, 3rd ed., Gerald Duckworth & Co. Ltd., London, U. K., 1985.
- [33] E. Hanslick, *The Beautiful In Music*, G. Cohen (trans.), Novello and Company, London, U. K., 1891. p. 17.
- [34] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, A Bradford Book, The MIT Press, Cambridge, MA, U. S. A. 1990. p. 455.

INVESTIGATION OF ITD SYMMETRY IN MEASURED HRIRS

Andrea F. Genovese, Jordan Juras, Chris Miller, Agnieszka Roginska

New York University
 Music and Audio Research Lab
 35 West 4th Str. 10012, New York, NY, United States
 genovese@nyu.edu

ABSTRACT

The *Interaural Time Difference* is one of the primary localization cues for 3D sound. However, due to differences in head and ear anthropometry across the population, ITDs related to a sound source at a given location around the head will differ from subject to subject. Furthermore, most individuals do not possess symmetrical traits between the left and right pinnae. This fact may cause an angle-dependent ITD asymmetry between locations mirrored across the left and right hemispheres. This paper describes an exploratory analysis performed on publicly available databases of individually measured HRIRs. The analysis was first performed separately for each dataset in order to explore the impact of different formats and measurement techniques, and then on pooled sets of repositories, in order to obtain statistical information closer to the population values. Asymmetry in ITDs was found to be consistently more prominent in the rear-lateral angles (approximately between 90° and 130° azimuth) across all databases investigated, suggesting the presence of a sensitive region. A significant difference between the peak asymmetry values and the average asymmetry across all angles was found on three out of four examined datasets. These results were further explored by pooling the datasets together, which revealed an asymmetry peak at 110° that also showed significance. Moreover, it was found that within the *region of sensitivity* the difference between specular ITDs exceeds the just noticeable difference values for perceptual discrimination at all frequency bands. These findings validate the statistical presence of ITD asymmetry in public datasets of individual HRIRs and identify a significant, perceptually-relevant, region of increased asymmetry. Details of these results are of interest for HRIR modeling and personalization techniques, which should consider implementing compensation for asymmetric ITDs when aiming for perceptually accurate binaural displays. This work is part of a larger study aimed at binaural-audio personalization and user-characterization through non-invasive techniques.

1. INTRODUCTION

One of the most crucial binaural cues in spatial sound is the *Interaural Time Difference* (ITD). For a particular location in space, an ITD describes the difference of time-of-arrival between the two ears for a sound source's direct path. ITDs are an important factor in how the human brain localizes a sound source and

it is the primary binaural cue for low frequencies [1]. ITDs are contained within measured *Head-Related Impulse Responses* (HRIRs), which can be used to transfer the perceptual cues of a sound source at some point in space about a listener's head to any mono sound file through time-domain convolution [2].

It is generally accepted that a users' experience is significantly improved by the use of individually measured HRIRs rather than the HRIRs recorded on mannequin dummy heads [2]. Different pinnae and head morphologies affect both ITD and spectral cues individually for each user. Only listeners with a morphology close to that of the mannequin will experience a satisfactory binaural reproduction, while most listeners will experience a degraded spatial auditory image.

Recently, many modeling techniques for parameterizing individual HRTFs (the Fourier transformation of HRIRs) from pictures and scans came to the attention of the sound engineering community [3] [4]. One such technique [5], generates a measure of the user's head size through a photographic technique and a subsequent adaptation of a general HRIR subset by the insertion of the user's head-size related ITDs, calculated via a spherical head model (first developed by [6]).

However, these models assume a symmetrical morphology of the listener's head and pinnae, implying symmetrical cues for sound localization between the left and right sides of the horizontal plane. This assumption of symmetry presumes an equal ITD value for a source placed, for example, at an azimuthal location of $\theta = -30^\circ$ and its respective mirrored-equivalent at $\theta = +30^\circ$ in the opposite hemisphere. In literature, there is an indication that this assumption of symmetry does not always hold true [7], motivating the need for more exploratory studies aimed at analyzing and assessing available spatial audio data in order to evaluate the severity of asymmetric ITDs. While Zhong et al. [7] have explored the asymmetry of HRTFs, the scope of this paper will focus on ITDs only.

Evidence of significantly asymmetrical ITDs would motivate further exploration on the correlations between morphological asymmetries and mirrored ITDs. An analysis of this relationship could be of significant value to the future development and improvement of HRIR/HRTFs modeling techniques that could adjust or compensate modeled personalized HRIRs for a user with asymmetric cues. At this stage, no perceptual test which could validate the relevance of physical ITD asymmetries has yet been performed, nonetheless the ITD differences between mirrored locations can be compared to established *Just Noticeable Difference* (JND) values [8] for location discrimination. Specifically, the goal of this paper is to explore whether ITDs in individual-HRIR public



This work is licensed under Creative Commons Attribution-Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

databases are statistically asymmetric and whether there is a pattern in the asymmetry across angles in the horizontal plane. This exploratory analysis is part of a bigger study by New York University's Music and Research Laboratory (NYU-MARL), which is focused on extracting personalized spatial audio cues through non-invasive techniques.

2. DATABASES

For this analysis, publicly available databases of individual HRIRs were collected. The datasets were chosen on the basis of their frequent use within the field and their diversity in terms of sample rate and measurement methodology. Four databases of individual HRIRs measurements were found to be suitable for this analysis. Due to the different characteristics and measurement methods of the databases, it was decided that the analysis of ITD asymmetry would be performed separately and independently for each dataset. The four selected datasets are: LISTEN [9], CIPIC [10], FIU [11] and MARL [12].

2.1. LISTEN database

This database was recorded for the LISTEN project [9] initiated as a collaboration between IRCAM and AKG. This set was recorded on 51 subjects¹ at 44.1 kHz sampling rate. The set consists of 187 locations per subject, measured at different azimuth resolutions for 10 respective elevation angles spaced at 15° from $\phi = -45^\circ$ to $+90^\circ$.

The advantage of this set lies in its measurement technique which made use of a crane structure, to precisely move a loudspeaker along a rig, and of a software-controlled rotating chair with a headrest used to rotate the subject to the desired azimuth degree. Using a single loudspeaker avoided the issue of having to compensate for each loudspeaker position. In comparison with other datasets in this list, the method used for LISTEN limited the possibility of measurement errors caused by human misalignments with the target angle, which would likely affect ITDs. Furthermore, the microphone capsules used for recording allowed for blocked-meatus conditions to prevent resonance of the ear-canal.

The drawback of the LISTEN database is the 44.1 kHz sampling rate which gives a low temporal resolution for estimating the magnitude of ITD asymmetry (distance between samples is $22\mu s$ as opposed to $10\mu s$ in the 96 kHz case).

2.2. CIPIC database

The CIPIC database [10] was compiled and made publicly available by UC Davis. This set consists of 25 azimuth locations recorded between -80° to $+80^\circ$ in azimuth and 50 different elevations from -45° to $+230.625^\circ$ (steps of 5.625°), for a total of 1250 locations recorded per subject (45 subjects). HRTFs were measured using Golay-code signals at 44.1 kHz, where the subject seated in a 1m radius hoop aligned to the subject's interaural axis. Subjects were not constrained but were able to monitor their head position.

One particularly interesting aspect of this measurement set is the inclusion of a detailed collection of anthropometric data of head and pinna morphology parameters which allows for the possibility of a future correlational study between ITD asymmetries

¹“Subject 1034” was later removed from the set due to inconsistencies in measurement data. So, 50 subjects were ultimately considered.

and anthropomorphic asymmetries. Pinna measurements were collected on both sides of the head, providing useful information about morphological asymmetries within subjects' ear characteristics. However, the CIPIC dataset was constructed using a non-optimal sample rate (44.1 kHz) for a high-precision detection of onsets.

2.3. FIU database

The Florida International University DSP Lab measured the individual HRTFs of 15 different subjects² at 12 azimuth locations (30° spacing) and 6 elevations [11]. Recordings were conducted at 96 kHz and were accompanied by anthropometric data measured via a 3D scan of the pinnae. The set includes measurements for 6 different elevations spaced at 18° apart from -36° to 54° . The measurements were conducted using the HeadZap system from AuSIM 3D using Golay-Code. The recording methodology relied on a rotating chair and a laser pointer to align the subject, and is thus not fully reliable as tiny head movements between measurements might give rise to ITD asymmetries. It was found that one of the angle mirrored pairs ($\pm 150^\circ$) had to be dropped due to problems in the set composition where the HRIR data from -150° was identical to that at $+150^\circ$.

2.4. MARL database

The last database to be analyzed was the NYU MARL (Music and Audio Research Lab) dataset collected by Andreopolou et al. [12] in 2013 and formatted to the MARL repository standard described in [13]. Four participating subjects had their personal HRTFs measured ten times each using different alignment methods (rotating stool with laser pointer or magnetic tracker) at a 48 kHz sampling rate. The MARL dataset thus allows for the study of HRTF feature variability across measurements. HRTFs were measured with a resolution of 10° azimuth and 15° elevation angles (going from -30° to $+30^\circ$ of elevation) for a total of 180 filter pairs per set. 40 sets (10 sets per 4 subjects) were collected — reduced to 32 after removal of corrupted sets. HRTFs were recorded with different techniques across repeated measurements using Golay Code, Maximum Length Sequences, and sine sweep signals.

In this paper, each repeated measurement on the same subject was treated as if it were a different independent subject, thus pooling together different measurement techniques within the dataset. The shortcomings of this dataset for our purposes are due to the low-precision of the alignment techniques and non-uniformity of the measurement method. These factors make it hard to discern between the presence of asymmetry in ITD and simple misalignments due to head movements during recording. These key aspects make MARL a very different dataset that has to be interpreted more carefully than the rest. The actual number of subjects is low, so a low variance of asymmetry is expected. However, it is interesting to explore how the non-uniformity of techniques will impact measurement error and therefore the results.

3. ANTHROPOMETRIC ASYMMETRY

Table 1 illustrates a pre-analysis exploration on the symmetry of individuals' anthropometric data included in the CIPIC database for a sample of their measured subjects. The provided data [10] includes a variety of pinnae measurements for both the left and

²Later reduced to 14 due to a formatting problem

Anthropometric Feature	Mean	σ
Cavum Concha Height (cm)	0.1213	0.1077
Cymba Concha Height (cm)	0.0910	0.0642
Cavum Concha Width (cm)	0.1357	0.0883
Fossa Height (cm)	0.1551	0.1421
Pinna Height (cm)	0.3207	0.2737
Pinna Width (cm)	0.1613	0.1251
Intertragal Incisure Width (cm)	0.0715	0.0551
Cavum Concha Depth (cm)	0.1295	0.1152
Pinna Rotation Angle (deg°)	0.0957	0.0415
Pinna Flare Angle (deg°)	0.0813	0.0638

Table 1: Mean and Standard deviation of the left-right difference of pinnae features measurements for 37 subjects in CIPIC

the right hemisphere. The table reports the values of the average feature difference between the left and right ear measure across subjects, and the standard deviation of these differences. The most varied feature across subjects is the “Pinna Height”, with an average difference of 0.3207 cm. The presence of these morphological asymmetries supports the hypothesis that ITDs will also be asymmetrical.

4. ANALYSIS AND RESULTS

All of the available data was analyzed to produce a series of plots intended to illustrate the presence and severity of asymmetry specific to ITDs. Due to the mismatch in angle resolutions, sample rate, and reliability of measurement techniques, the first-stage of the analysis did not pool the data but was rather conducted separately for each database. Each set is therefore analyzed in its own context, allowing us to observe whether different measurement methods yield different results.

In the context of this document, the ITD was calculated as a measure in samples (then translated into seconds) between the two points of maximum time-domain cross-correlation, between the left and right ear HRIR signals. For consistency, the ITD cross-correlation calculation function was applied to each database, including those who already provided ITD data. This was not possible, however, for the FIU database, as the HRIR measurements were only available in a minimum phase format — therefore the provided ITD values were used. To account for clear measurement errors, a pre-analysis inspection of the data led to the discarding of strong outlier measurements that could affect the results.

The ITD symmetry for every subject was calculated as the absolute value of the difference of ITD magnitude with that of their respective mirrored counterpart. A difference of 0 would indicate

complete symmetry between hemispheres, while a higher absolute difference would point to a higher level of asymmetry. The values were averaged across each set of N subjects for every available angle $\theta \in [0; 180]$ within each dataset:

$$\bar{S}(\theta) = \frac{1}{N} \sum_{n \in N} ||ITD_n(\theta)| - |ITD_n(360^\circ - \theta)|| \quad (1)$$

The mean and standard deviation of the absolute asymmetry across subjects was measured for those angles on the horizontal plane that possessed an ITD value at both hemispheres. Thus, the 0° and 180° angles were excluded from the analysis as they lack a mirrored correspondent across hemispheres.

DATASET	N	θ_P	$\bar{S}(\theta_P)$	\bar{S}'
LISTEN (a)	50	105°	$84\mu s$	$51\mu s$
CIPIC (b)	45	115°	$119\mu s$	$49\mu s$
FIU (c)	14	120°	$72\mu s$	$32\mu s$
MARL (d)	31 (4)	150°	$47\mu s$	$35\mu s$
D1 (a+b+d)	126	110°	$80\mu s$	$45\mu s$
D2 (a+b)	95	110°	$94\mu s$	$49\mu s$

Table 2: Peak ITD asymmetry values, peak positions across examined databases and non peak curve averages

DATASET	df	t	p
LISTEN (a)	49	3.91	$1.6e-04$
CIPIC (b)	44	3.90	$3e-03$
FIU (c)	13	3.27	$1.4e-04$
MARL (d)	30	1.51	0.07
D1 (a+b+d)	125	4.46	$8.9e-06$
D2 (a+b)	94	4.70	$4.4e-06$

Table 3: One-tail t-test results. Three out of four datasets (in bold-face) showed significance at 95% confidence level.

The plots in Figure 1 illustrate the average ITD difference across subjects between the left and right hemispheres on the horizontal plane (0° elevation) for each dataset. Error bars show the 95% confidence interval boundaries for each angle. In Table 2, the peaks of the ITD average curves are reported to the nearest microsecond, along with the associated azimuth angles. Ignoring the MARL set, once notices the proximity of the peak azimuths placed around $110^\circ \pm 10^\circ$. In the case of MARL, no clear peaks can be identified and the curve looks somewhat flatter.

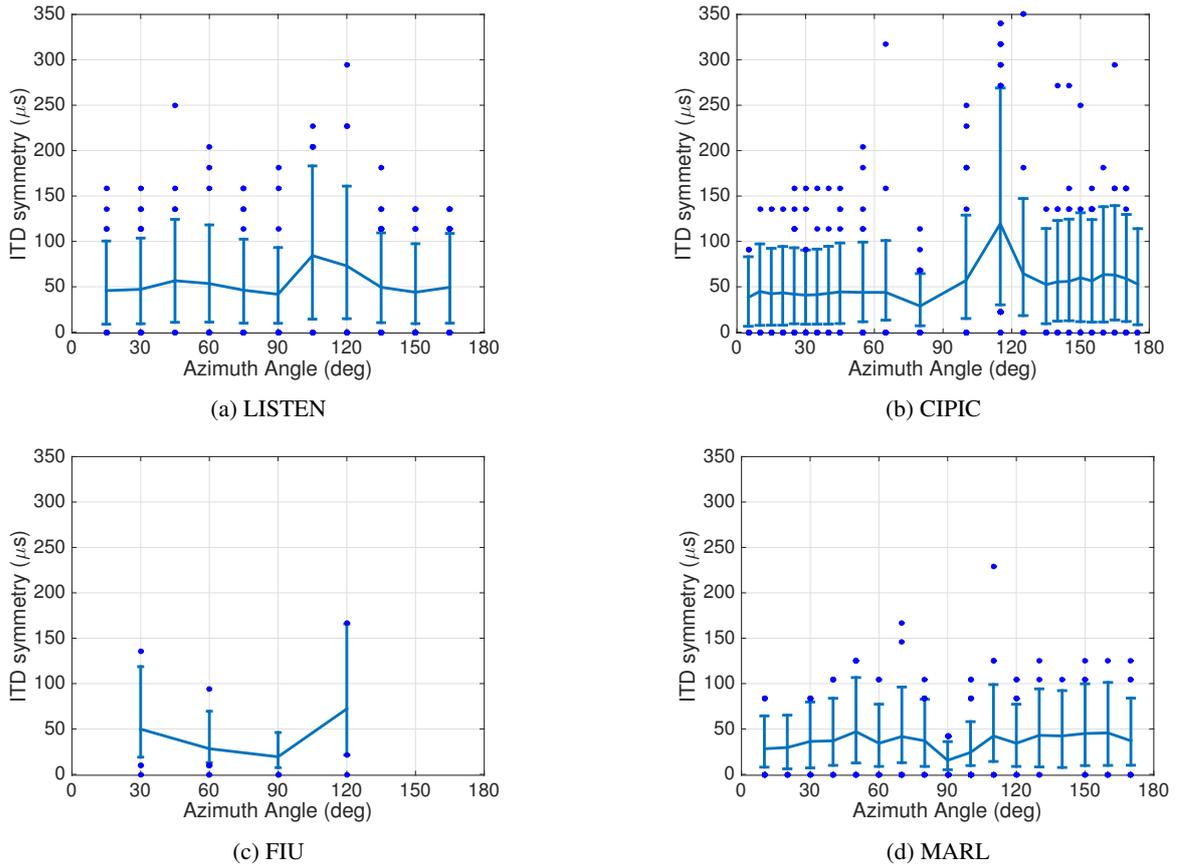


Figure 1: Average ITD symmetry and standard deviation across subjects for the available datasets calculated using (1). The azimuthal resolution depended on the set. All the ITDs were taken at zero elevation.

In Figure 2, we notice a region of higher ITD asymmetry magnitude. For three out of the four examined databases, this region ranges roughly in the lower bound around 90° and 130° in the upper bound. Unfortunately, the precision of this statement is undetermined by the absence of coherent resolution and sample rates between databases, though the general shape of the asymmetry curves across databases is similar (Figure 2).

The presence of this region motivates further exploration into whether HRIRs at certain angles are more sensitive to morphological asymmetries than others. To inspect this hypothesis, one-tailed t-tests between the ITD symmetry values at the peak azimuth θ_P for each subject, and the curve's average asymmetry value excluding the peak \bar{S}' , were conducted for each set (details in Table 2). The results of this test highlight whether the asymmetry at the peak is significantly higher than the asymmetry at the rest of the angles, with significant difference show for three of the sets at 95% confidence ($p < .05$) (Table 3).

From Figure 2 we also note a consistent minimum ITD asymmetry at 90° azimuth indicating a possible interaction with the angle of incidence.

4.1. Data Pooling

After observing the results for each database, the data was pooled in order to extract ITD asymmetries representative of the entire sample population and increase confidence in the results obtained thus far. The calculated ITDs for the databases were linearly interpolated to the whole 360° horizontal plane. The FIU set was excluded due to excessively missing data. Since the MARL dataset consists of repeated measurements for four subjects and its measurement techniques are less reliable from CIPIC and LISTEN, two versions of data-pooling sets were created, one with MARL (D1, $N = 126$) and one without MARL (D2, $N = 95$).³

Results in Figure 3 show that, regardless of the inclusion of the MARL set, the asymmetry peak is centered at 110° . In both cases, the t-tests values shown in Table 2 show a significant difference (95% confidence level) of the peak region's asymmetry against all other values.

³ITD angle-alignment based on Left-Right contours was considered to correct for misalignment, however that would have meant to assume a spherical head model. Preliminary tests showed that the asymmetry magnitudes were not reduced by the alignment, on the contrary they increased.

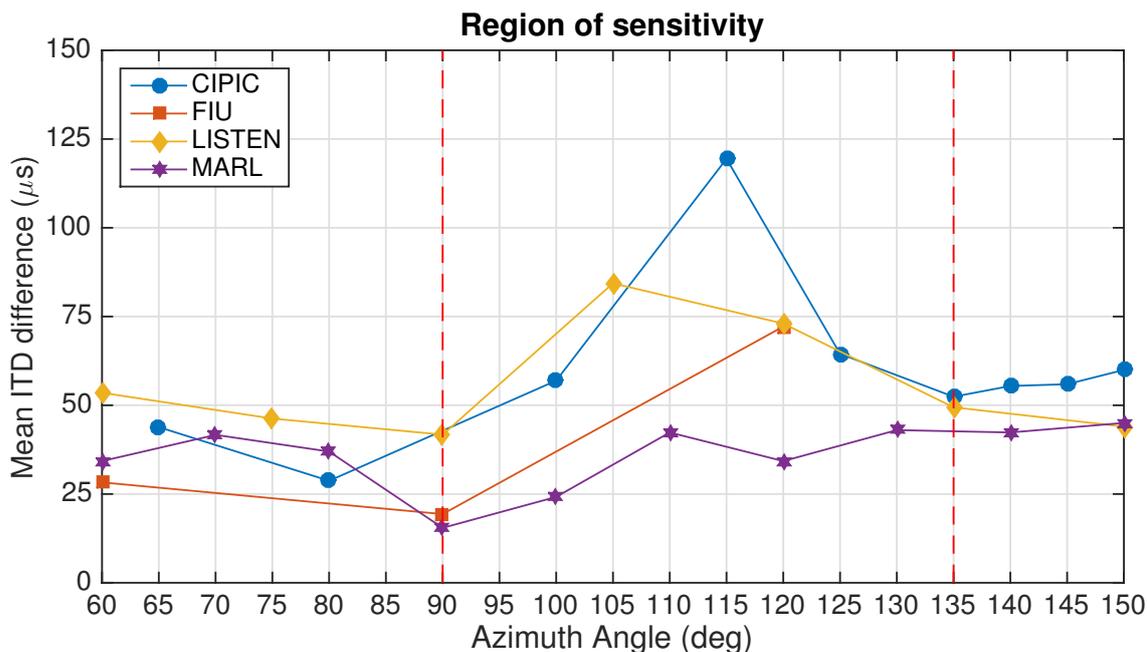


Figure 2: Close-up superposition of all mean ITD asymmetry curves across 60° and 150°. An identified region of sensitivity spans between 90° and 130°

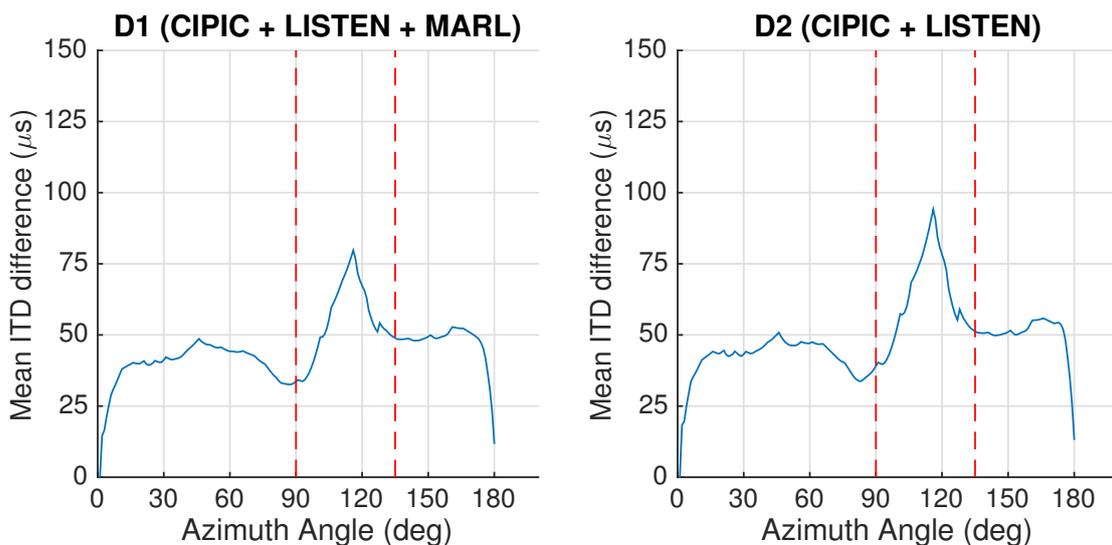


Figure 3: Average asymmetry across the horizontal plane for pooled datasets D1 (CIPIC + LISTEN + MARL) and D2 (CIPIC + LISTEN). The dotted lines represent the sensitivity region

4.2. Summarization Table

Table 4 illustrates an attempt to summarize and characterize the nature of the ITD asymmetry curve of each database and parameterize its distribution across the horizontal plane. The table shows values approximated to the nearest microsecond for the grand mean, standard error, skewness and kurtosis. The standard error for the FIU database is much greater due to the smaller number of subjects. Table 4 provides a quick reference of the severity of the asymmetry in all databases. The skewness value describes the off-

set of the sensitivity region, while the kurtosis indicates its width.

5. DISCUSSION

The previous section identifies a “region of sensitivity” located between 90° and 130° azimuth, whose presence is validated by the fact that all datasets showing a significant difference between the peak asymmetry values and the average asymmetry values find their peaks located within 15°. The MARL dataset is the only

Dataset	N	Mean	Median	SE	Skewness	Kurtosis
LISTEN (a)	50	54 μ s	50 μ s	1.8 μ s	1.43	3.756
CIPIC (b)	45	52 μ s	49 μ s	2.6 μ s	2.524	11.087
FIU (c)	14	42 μ s	39 μ s	6.3 μ s	0.341	N/A
MARL (d)	31	36 μ s	37 μ s	1.5 μ s	-0.946	3.142
D1 (a+b)	126	46 μ s	45 μ s	1 μ s	0.407	4.754
D2 (a+b+d)	95	49 μ s	47 μ s	1.3 μ s	1.032	5.395

Table 4: Summarization of symmetry curves in each database

dataset showing non-significant results, but is less suited to explore population asymmetry parameters due to its focus on a small number of subjects. It is, however, rather suited for exploring asymmetry caused by measurement error. The four MARL subjects display more symmetric features on average, though no morphological data is provided to verify this. The lack of strong asymmetry peaks in MARL may result from the small pool of subjects, and seems to be independent of measurement method.

indicate that measurement errors are not the cause for the strong asymmetries in the other datasets. We therefore hypothesize that the asymmetry observed in the three other sets is independent of the measurement technique.

One reason to keep the datasets separate was to observe whether different measurement techniques have an impact on the magnitude of ITD asymmetry. The pooled data looked for more general statistics regarding average asymmetries and the *sensitivity region* that would approach a representative population curve. As depicted in Figure 3, a significant peak occurs at $\theta = \pm 110^\circ$, which can be regarded as the point of maximum asymmetry in the azimuth plane — observed within the proposed *region of sensitivity*. For D1 (MARL set included), the average asymmetry was about 45 μ s and the peak asymmetry 80 μ s. Excluding MARL (D2) the average asymmetry was 47 μ s and the peak asymmetry 94 μ s. T-tests for both cases showed significant difference between peak and average asymmetry ($p < .05$). The inclusion of the MARL database did not have a significant impact on the results.

Is the magnitude of the ITD asymmetry large enough to cause perceptual asymmetry? Per Klumpp and Eady [8], the ITD's JND is highly frequency-dependent. The study found that for a 1 kHz pure tone the JND was on average about 11 μ s, but for a 90 Hz tone the JND increased to 75 μ s, while for band-limited random noise the average JND drops to 9 μ s. As depicted in Figure 3, the peak values of the average ITD asymmetry range between 80 – 94 μ s, exceeding the JND values for all frequency bands. Therefore, the localization error between two mirrored locations is significantly more noticeable in the region of sensitivity. Figures 1 (a) and (b) both present peaks higher than most JNDs. The asymmetry results of Figure 1 (c), and the t-test results are less reliable due to the low number of subjects (14) and error-prone methodology, but nevertheless the curve is indicative of a similar range of sensitivity between 90° and 130°.

These results are useful for guiding the study of the influence

of head and ear morphologies on sound perception. Since the peak asymmetries were found on the rear-lateral angles, it is possible to hypothesize that asymmetrical pinnae might interfere with the direct path of the sound source to the tympanus. A likely factor is illustrated in Table 1 which depicts the *Pinna Height* as the most asymmetric anthropometric feature in the pool of subjects of CIPIC.

From an engineering perspective, the impact of these findings relates to the trade-off between the complexity and accuracy of a binaural audio reproduction system. For the angles within the sensitivity range, the localization accuracy of a listener is likely to degrade if the system uses standardized non-individual HRIRs or symmetrical-head models for personalization. Depending on the application, localization accuracy might be more or less important for the correct delivery of the intended user experience. Applications aimed for entertainment, such as music or movie reproduction, are unlikely to necessitate precise localization and may prefer a simpler processing environment which assumes a spherical head model. If, instead, localization accuracy is key (i.e. binaural navigation systems or psychoacoustics studies), then ITD asymmetry compensation should be considered.

6. CONCLUSIONS

This study used a number of publicly available personalized HRIR databases to investigate the presence and severity of asymmetric ITD measurement data between the left and right hemisphere on the horizontal plane. A consistent azimuth region of higher sensitivity, approximately between $\pm 90^\circ$ to $\pm 130^\circ$, was identified in all datasets. T-tests confirmed that the asymmetry in the region's peak is significantly different than the average asymmetry at all other angles for three out of four datasets. These results were further explored by merging the datasets in interpolated pooled repositories, which identified a general ITD peak asymmetry at 110°, also significantly different than the average asymmetry.

From a perceptual standpoint, the asymmetry peaks in set D1 and D2, which range from 80 μ s to 94 μ s, are higher than the ITD JND thresholds found in [8]. At least for the examined datasets, the peak asymmetries in the sensitivity region are likely to be severe enough to be noticeable at all frequencies.

These findings validate the statistical presence of perceptually-noticeable asymmetric ITDs in individually measured subjects. The likely causes of ITD asymmetries are the anthropometric

asymmetries of the head and ears which lead to morphological differences across hemispheres. These results are of interest for HRIR personalization and modeling techniques that aim for accurate localization, especially if making use of the same public repositories utilized for this study.

6.1. Further Work

An immediately apparent next step would expand this analysis to more datasets. It would also be interesting to further explore the role of other variables, such as elevation and measurement technique, in ITD asymmetry. However, the lack of a common format in the used databases created a non-ideal situation where low angle-resolutions made it harder to precisely refine the location of the asymmetry peaks, having to resort to interpolation for missing angles. Moreover, low sampling rates are an obstacle to precise ITD values. It is hoped that in the future more public datasets will adhere to a defined standard such as the SOFA conventions [14], developed as HRTF data exchange format and able to make this type of analysis easily implementable.

Further analysis on the morphological characteristics that give rise to the ITD asymmetry could highlight a possible role of the outer ear pinnae in modifying the direct path from the source to the ear canal. In the context of binaural personalization, there is motivation to establish an exact relationship between anthropometry and sound perception. It would be interesting to run a regression test and look for correlations between physical ITD asymmetry and morphological asymmetry for a pool of subjects representative of the general population, perhaps starting with *Pinnae Height* (Table 1). Binaural audio models could then improve the listening experience by implementing an additional layer of signal compensation based on the observed anthropometric parameters. For example, if the asymmetry found in the HRIR measurements were to be correlated to morphological asymmetries, the mismatch between hemispheres could be easily parameterized and accounted for by using a more comprehensive individualization technique (i.e. photographic information), using only the identified relevant features. This type of study is currently a challenge due to the limited availability of morphological data across public datasets.

As a final note, the importance of these results has to be assessed in light of the fact that physical asymmetry may not coincide with perceptual asymmetry of spatial sound. In fact, perceptual compensation via neural internal-delay has been previously hypothesized [15] [16]. Formal listening tests are required in order to confirm the indications given by the JND values and to inspect if there is a direct correlation between physical and perceptual asymmetries.

7. REFERENCES

- [1] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, 1992.
- [2] M. Morimoto and Y. Ando, "On the simulation of sound localization." *Journal of the Acoustical Society of Japan (E)*, vol. 1, no. 3, pp. 167–174, 1980.
- [3] N. Gupta, A. Barreto, and M. Choudhury, "Modeling head-related transfer functions based on pinna anthropometry," in *Proc. of the Second International Latin American and Caribbean Conference for Engineering and Technology (LACCEI)*, 2004.
- [4] M. Dellepiane, N. Pietroni, N. Tsingos, M. Asselot, and R. Scopigno, "Reconstructing head models from photographs for individualized 3d-audio processing," in *Computer Graphics Forum*, vol. 27, no. 7. Wiley Online Library, 2008, pp. 1719–1727.
- [5] J. Juras, C. Miller, and A. Roginska, "Modeling itds based on photographic head information," in *Audio Engineering Society Convention 139*, Oct 2015.
- [6] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 472–479, 2001.
- [7] X.-L. Zhong, F.-c. Zhang, and B.-S. Xie, "On the spatial symmetry of head-related transfer functions," *Applied Acoustics*, vol. 74, no. 6, pp. 856–864, 2013.
- [8] R. Klumpp and H. Eady, "Some measurements of interaural time difference thresholds," *The Journal of the Acoustical Society of America*, vol. 28, no. 5, pp. 859–860, 1956.
- [9] O. Warufsel, "Listen hrtf database, ircam," 2002. [Online]. Available: <http://recherche.ircam.fr/equipements/salles/listen/>
- [10] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipc hrtf database," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001, pp. 99–102.
- [11] N. Gupta, A. Barreto, M. Joshi, and J. C. Agudelo, "Hrtf database at fiu dsp lab," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 169–172.
- [12] A. Andreopoulou, A. Rogińska, and H. Mohanraj, "A database of repeated head-related transfer function measurements," 2013.
- [13] A. Andreopoulou and A. Roginska, "Towards the creation of a standardized hrtf repository," in *Audio Engineering Society Convention 131*, Oct 2011. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=16096>
- [14] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, *et al.*, "Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.
- [15] S. E. Boehnke and D. P. Phillips, "Azimuthal tuning of human perceptual channels for sound location," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1948–1955, 1999.
- [16] P. Joris and T. C. Yin, "A matter of time: internal delays in binaural processing," *Trends in neurosciences*, vol. 30, no. 2, pp. 70–78, 2007.

Implementation and Evaluation of 10.2 channel Microphone for UHDTV Audio

Daeyoung Jang, Jae-hyoun Yoo, Tae Jin Lee

Electronics and Telecommunications Research Institute of Korea(ETRI),
Audio Lab.,
P.O.Box 34219, 218 Gajeong-no, Yuseong-Gu, Daejeon, Korea
{dyjang, jh0079, tjlee}@etri.re.kr

ABSTRACT

As broadcasting environments change rapidly to digital, user requirements for next-generation services that surpass the current HDTV service quality become more demanding. The next-generation of broadcasting services will change from HD to UHD and from 5.1 channel audio to more than 10 audio channels, including a height channel for a high quality realistic broadcasting service. In accordance with the estimated trends of future broadcasting services, we propose a 10.2 channel audio format for a Korean UHDTV broadcasting service. It can create almost similar spatial sound images as 22.2 channel audio with half the number of speakers. In this paper, we propose a 10.2 channel audio acquisition system for the creation of UHDTV content, and measurements and preliminary evaluation are carried out to determine whether the performance is acceptable for broadcasting.

1. INTRODUCTION

Most advanced countries have completely changed their broadcasting services from analog to digital. Other countries are also preparing digital broadcasting services due to its transmission channel efficiency. In particular, ultra-high definition (UHD) video services have been launching for cinema and broadcasting areas since 2014. Immersive audio services are also required to provide spatial synchronization with a widened video display above 100 inches for a UHD video service. Additionally, both horizontal and vertical sound expression are necessary to provide spatial envelopment for a reasonably immersive audio service. Usually, many reports have insisted that an audio format with more than 10 channels is acceptable for immersive sound representation [1–4].

ETRI proposed 10.2 channel audio as an immersive sound format for a UHDTV broadcasting service in Korea in 2011 [5]. This format consists of seven channels of horizontal loud-speakers, three channels of ceiling loud-speakers, and two woofers. The seven-channel horizontal loud-speaker layout is the same as the Dolby/DTS 7.1 channel format for cinema.

This paper describes a 10-channel spherical microphone that can be used to record a 10.2 channel audio signal for broadcasting. First, section 2 illustrates a 10.2 channel format for UHDTV in Korea. Section 3 depicts the design and implementation of a 10-channel microphone system and section 4 describes characteristics of the implemented 10-channel microphone; finally, section 5 makes a conclusion about future work for further plans regarding evaluations and verifications.

2. 10.2 CHANNEL FORMAT FOR UHDTV

2.1. Layout of 10.2 channel audio format

The ideal layout of a 10.2 channel audio format as a beyond 5.1 channel realistic sound representation for UHDTV is depicted in Figure 1. The 10.2 channel audio format contains three channels of horizontal front loudspeakers, two side surround loudspeakers, two back surround loudspeakers, three ceiling loudspeakers, and two Low Frequency Effect (LFE) channels.

The horizontal front and side surround loudspeakers are compatible with a 5.1 channel loudspeakers setup, and the back surround loudspeakers are compatible with Dolby/DTS 7.1 channel loudspeakers for cinema sound. These two channels of back surround loudspeakers are important for reducing front–back confusion, the phenomenon in which listeners confuse whether a sound source is in front or behind them, because the Inter-aural Time Difference (ITD) and Inter-aural Level Difference (ILD) are similar.

An additional two front ceiling channels are located in the upper position of the screen at the adjacent direction of horizontal front left and right loudspeakers. These channels can represent ceiling sounds such as airplanes, lightning, etc. Another ceiling channel is located at a very high position relative to the listeners head or somewhat behind that. To support compatibility with a NHK 22.2 channel sound system and the establishment of a ceiling loudspeaker in a common listening room, it is recommended that the back ceiling loudspeaker is located between the vertical 90° and 135° position from the listener.

The LFE channels are two channels at the adjacent position of front left and right loudspeakers, and provide a flatter front sound image. Usually, these two channels have the same signal to provide powerful effect sounds for cinema, but a 90° phase shift from each other provides the envelopment of low frequency sounds for music reproduction.

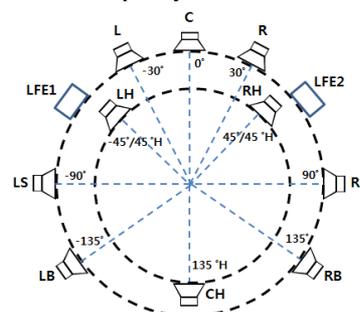


Figure 1: 10.2 channel loudspeaker layout.

2.2. Sound localization performance of 10.2 channel

We compared the 5.1 and 22.2 channel format through objective and subjective evaluation methods to evaluate the sound localization performance of the 10.2 channel audio format [6]. The objective sound localization method used in this paper is the Auditory Process Model (APM) by M. Park [7]. And evaluation test results are followed in subsections.

2.2.1. Objective test

APM is a mathematical model of a human's three-dimensional sound perception process that consists of a peripheral process to simulate neural transduction, a binaural process to simulate ITD, ILD characteristics, and central processes to determine sound localization. In these processes, APM can get similar results to subjective sound localization tests.

The objective horizontal sound localization performance of the right side of a 10.2 channel system was calculated using APM, and those of 5.1 and 22.2 channel systems were also calculated for comparison. The layouts of the loudspeakers of the 5.1, 10.2, and 22.2 channel formats under test were depicted in Figure 2.

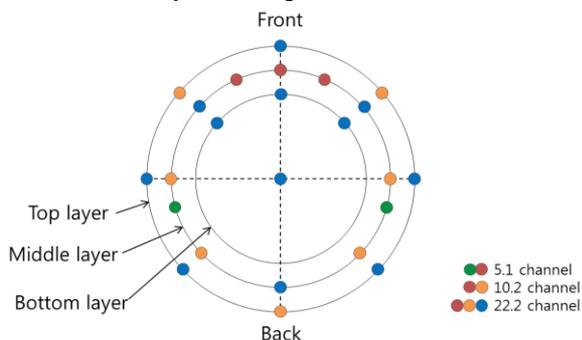


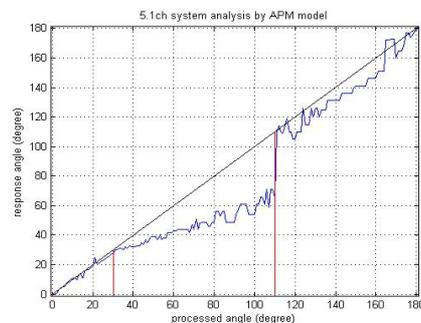
Figure 2: The loudspeaker layouts for tests.

The side-right surround loudspeaker of the 5.1 channel format is located at 110° from the center channel position. The subset of 22.2 channel loudspeakers are also used in the 10.2 channel format, because the 10.2 and 22.2 channel formats are compatible in the loudspeaker position. Sound source are localized by panning with adjacent loudspeaker pairs to obtain the APM parameters for every direction on the right side of the listener. Then, a test signal is reproduced and APM was calculated with stereo signals that were acquired at the sweet spot.

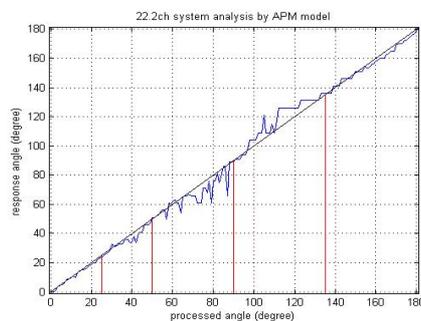
Figure 3 shows the results of APM for the horizontal right side channels of 5.1(a), 22.2(b), and 10.2(c) channels. In the graph of Figure 3, the horizontal axis represents the reproduced sound image and the vertical axis represents the sound image calculated by APM. Therefore, the sound localization performance is better when the curve is more coincident with the diagonal line. In Figure 3(a), the 5.1 channel system has degraded the sound localization performance for 30–110° of the side surround and 110–180° of the back surround.

In Figure 3(b), the 22.2 channel audio system shows very good sound localization performance because it uses enough loudspeakers for the horizontal surround sound image.

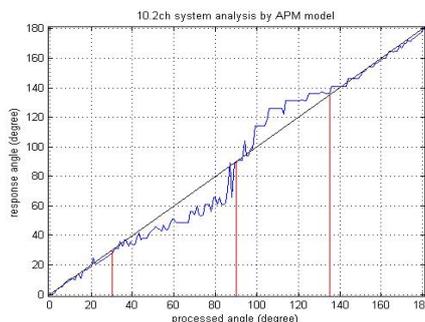
Compare Figure 3(c) to 3(a) and 3(b); the 10.2 channel audio system shows reasonable sound localization performance, meaning that the 10.2 channel audio system has better sound localization than the 5.1 channel audio system. However, the 10.2 channel audio system has a somewhat poorer sound localization performance than the 22.2 channel audio system, but a quite stable sound localization performance was obtained compared to the 5.1 channel audio system.



(a) 5.1 channel



(b) 22.2 channel



(c) 10.2 channel

Figure 3: Compare of sound localization by APM

2.2.2. Subjective test

Subjective testing with a 22.2 channel audio system was also conducted to confirm the sound localization performance of the 10.2 channel audio system. A 22.2 channel audio system was installed in the listening room, and the 10.2 channel audio system was implemented as a subset of the 22.2 channel audio system.

The subjects used for this listening test are nine spatial audio experts and three non-experts. In the test, the subjects heard several localized sound sources rendered by the 10.2 and 22.2 channel audio systems.

Table 1: Direction of stimuli for sound localization test.

	Elevation Angle (°)	Azimuth Angle (°)
A	70	330
B	90	15
C	80	80
D	60	260
E	70	165

Five stimuli of directional sounds for sound localization tests are defined in Table 1. Each sound source was rendered using the Vector Base Amplitude Panning (VBAP) method designed by V. Pulkki [8]. The subjects evaluated the sound localization performance of the 10.2 and 22.2 channel audio systems by listening to the submitted sound sources and pointing in the perceived sound direction with a laser pointer.

Figure 4 shows the subjective test results for the sound localization performance of the 10.2 and 22.2 channel audio systems. The 22.2 channel audio system had a 7.8° average difference for the perceived sound position, and the 10.2 channel audio system had a 10.6° average difference for the sound position. This 2.8° difference is an angle at which humans cannot usually perceive a difference in direction with their auditory system. Therefore, the results show that the sound localization performance of the 10.2 channel audio system is not degraded compared to that of the 22.2 channel audio system.

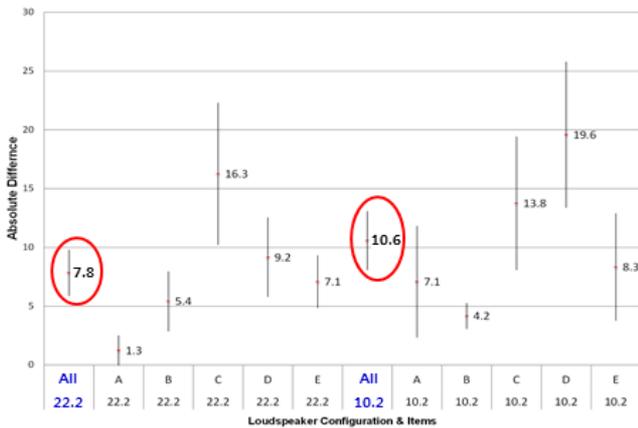
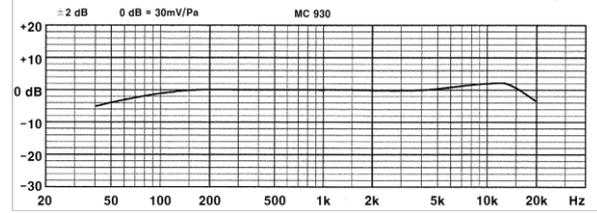


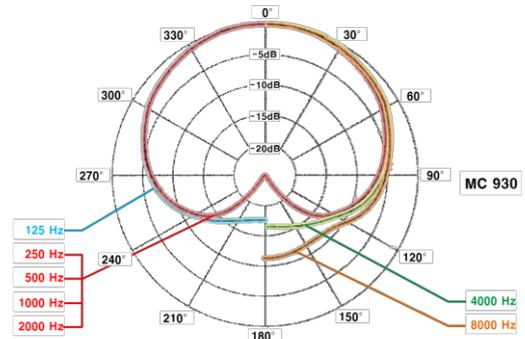
Figure 4: Subjective sound localization test results for the 10.2 and 22.2 channel audio systems.

3. IMPLEMENTATION OF A 10.2 CHANNEL MICROPHONE

The microphone units used in 10.2 channel microphone systems are cardioid directional microphones (Beyerdynamic MC930). The microphone unit has very flat frequency response, as shown in Figure 5(a), and has a stable directivity pattern for a wide frequency range, as shown in Figure 5(b).



(a) Frequency response

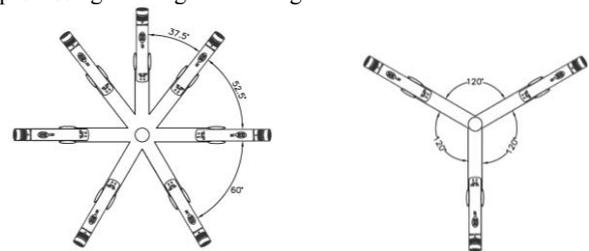


(b) Directivity patterns

Figure 5: Characteristics of the MC-930 (Beyerdynamic).

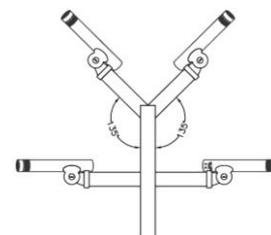
A 10-channel one point spherical microphone set was designed, as shown in Figure 6. The actual shape of the implemented microphone system is shown in Figure 7. The diameter of the microphone system is designed at about 55 cm, providing sufficient inter-channel signal separation while still being easy to carry.

Each microphone holder is a floating type for shock-and-vibration-free recording, with elastic string nets. The front left and right microphones have a wider angle of 37.5° from the center microphone for a lower correlation. In addition, the LFE channel can be generated by additional post-processing filtering and mixing.



(a) Lower channels layout

(b) Height channels layout



(c) Vertical section view

Figure 6: Diagram for the 10-channel microphone array.



Figure 7: Picture of an implemented 10-channel microphone.

4. MEASUREMENTS AND FIELD RECORDING FOR LISTENING

4.1. Measurements of directivity characteristics

The directivity of the implemented microphone system was measured in the listening room by the frontal sound source. Because the 10-channel microphone system has a symmetric shape for microphone directions, only the left side microphones were measured. Figure 8(a) shows the relative frequency responses for the frontal sound stimulus of white noise. The front center channel has the highest gain for almost all of the frequency range, and about 3 dB of the relative higher gain with front left channel.

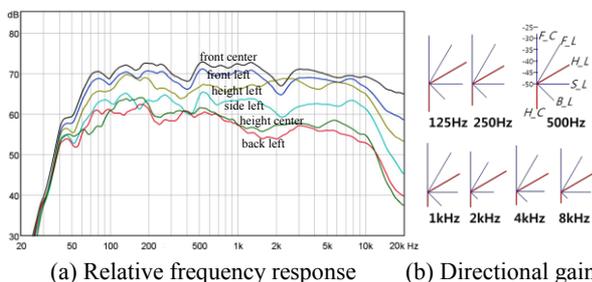


Figure 8: Directivity characteristics of the left half-sphere of the 10-channel microphones for a frontal sound source.

Figure 8(b) shows the directional gains for several typical frequencies. These directional gains represent the directivity characteristics of a 10-channel microphone system. The measurement results show that the implemented microphone system has very stable directivity characteristics throughout the whole frequency range.

4.2. Field recording and listening

Field recording was conducted using a 10-channel microphone system in the Korean traditional music hall (Figure 9(a)) and concert hall (Figure 9(b)). The Korean traditional music hall is an almost rectangular shaped room and has a somewhat reverberant effect. The recorded music pieces were interior music “cheon nyon manse,” Korean string instrument “hae geum solo” and the “song of Korean poetry.”

Several recording systems were used in the concert hall for comparison. Several stereo microphone setups and a 5.1 channel microphone were installed in the room, as shown in Figure 8(b). The recorded music contents included “Hungarian dances No. 1” by Brahms, “voice of spring” by Johann Strauß II, and “violin concerto” by Mendelssohn.



Figure 9: 10-channel field recording.

Recorded music contents were evaluated by informal listening tests performed by recording engineers and architectural acoustics experts (Figure 10). When adjusting the reproduction loudness levels of recorded music content similar to that of the recording place, we could feel similar envelopment and sound images. In addition, the frontal sound images of 10-channel sounds are more stable than those of 5.1 channel sounds.



Figure 10: Listening room for the field recording sounds.

5. CONCLUSION

This paper proposes a 10-channel microphone system for acquiring 10.2-channel high quality spatial sounds for Korean UHDTV broadcasting. A 10-channel microphone system was designed, implemented, and verified using directivity measurements and listening to field-recorded 10-channel audio content. The results of directivity measurements show that characteristics of a 10-channel microphone system are reasonable for spatial sound recording. It is identified through additional informal listening tests that an implemented 10-channel microphone can be used for the live recording of music programs for broadcast.

We have additional future plans to verify the performance of our 10-channel microphone system for use in broadcasting contents production. First, formal listening tests for 10-channel microphone systems will be conducted to confirm the performance of microphone systems for UHDTV broadcasting. Then, additional field recording will be followed for several programs such as sports, street scenes for live news, and wild life documentaries. Furthermore, we have to generate LFE channel signals from 10-channel signals recorded with a low-pass filter, mixing, etc.

6. ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (B0101-16-0295, Development of UHD Realistic Broadcasting, Digital Cinema, and Digital Signage Convergence Service Technology)

7. REFERENCES

- [1] Kimio Hamasaki et al., "Development of a 22.2 Multichannel Sound System," NHK STRL Broadcast Technology, No. 25, Winter 2006.
- [2] Kazuho ONO et al., "Portable spherical microphone for Super Hi-Vision 22.2 multichannel audio," AES 135th Convention, Oct. 2013.
- [3] Report ITU-R BS.2159-4, "Multichannel sound technology in home and broadcasting applications," ITU-R BS.2159-4, May. 2012.
- [4] White paper, "Dolby® ATMOS™ Next-Generation Audio for Cinema," Dolby Laboratories, Inc., 2013
- [5] TTA.KO-07.0098, Audio Signals for UHD Digital TV, 2011.
- [6] Taejin Lee et al., "Multichannel Audio Reproduction Technology based on 10.2 ch for UHDTV," JBE Vol. 17, No. 5, September 2012.
- [7] Munhum Park, Phillip A. Nelson and Kyeongok Kang, "A Model of Sound Localisation Applied to the Evaluation of Systems for Stereophony," ACTA Acoustica, Vol. 94, pp. 825~839, 2008.
- [8] Ville Pulkky, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," J. Audio Eng. Soc., Vol. 45, No. 6, June 1997.

LISTEN TO YOUR DRIVE: SONIFICATION ARCHITECTURE AND STRATEGIES FOR DRIVER STATE AND PERFORMANCE

Steven Landry¹, David Tascarella², Myoungsoon Jeon^{1,2}, & S. Maryam FakhrHosseini¹

Mind Music Machine Lab

¹Department of Cognitive and Learning Sciences

²Department of Computer Science

Michigan Technological University

1400 Townsend Drive, Houghton, MI 49931, USA

{sglandry, datascar, mjeon, sfakhrho}@mtu.edu

ABSTRACT

Driving is mainly a visual task, leaving other sensory channels open for additional information communication. As the level of automation increases in vehicles, monitoring the state and performance of the driver and vehicle shifts from the secondary to primary task. Auditory channels provide the flexibility to display a wide variety of information to the driver without increasing the workload of driving task. It is important to identify types of auditory displays and sonification strategies that provide integral information necessary for the driving task, and not overload the driver with unnecessary or intrusive data. To this end, we have developed an in-vehicle interactive sonification system using the medium-fidelity simulator and neurophysiological devices. The system is intended to integrate driving performance data and driver affective state data in real-time. The present paper introduces the architecture of our in-vehicle interactive sonification system and potential sonification strategies for providing feedback to the driver in an intuitive and non-intrusive manner.

1. INTRODUCTION

Automobiles and in-vehicle safety systems have been improved over the past several decades [1]. Researchers classify driving safety strategies into two categories: Passive and Active/Primary safety applications. Passive strategies, such as airbags, help people stay alive and uninjured in an accident [1], while active strategies, such as collision warnings and electronic stability control, aid in the prevention of car crashes and improve drivers' performance [2-3]. These active strategies reduce drivers' errors and improve their performance especially when they are engaged in secondary or tertiary tasks.

Since the human visual system is mostly busy with the driving task, designing active safety systems based on other sensory channels (e.g., auditory modality) is recommended. Auditory channels provide the flexibility to display a wide variety of information to the driver continuously. According to multiple resource theory [4], each task has a vector that shows the number and qualitative level of the attentional

resources. Assuming that driving includes visual, spatial, and manual resources, the amount of load within each resource depends on the task demands. Therefore, it is important to identify the appropriate auditory displays and sonification strategies that can provide integral information necessary for the driving task, and not overload the driver with unnecessary or intrusive data.

Sonification is a method for presenting data with sound [5]. The idea behind it is that people can perceive changes and draw conclusions easier and faster by listening to acoustic data in real time. A previous study [6] suggested that sonifying aggressive drivers' emotion will help them improve their driving performance in two ways: regulating their affect and providing appropriate feedback. This retrospective method can be used to return angry drivers' mood to a safer neutral mood. In contrast, in the present paper, we attempt to sonify drivers' performance and affective states in a more proactive way. Thus, this real-time sonification can also make a *feedforward* loop in terms of driving behavior and road safety.

2. DATA ROUTING FOR DRIVING PERFORMANCE AND DRIVER STATE

2.1. Data routing software from the driving simulator

As our driving sonification research platform, we use a medium-fidelity driving simulator, NADS (National Advanced Driving Simulator) MiniSim (Figure 1). Since the simulator provides all the driving data afterwards, the first step to sonify driving behavior in real-time was to develop a program that can relay the driving data in real-time. The first stage of our data routing software initializes the majority of the programs and lets the user set up the session. This includes selecting the driving variables (e.g., speed, pedal force, lane deviation, steering angle, etc.) of interest that our program will extract, display, record, and route from the driving simulator. The second stage listens for packets to come through on the created port. Then, based on the choices the user made in the first stage, the program parses the received packet for the variables that were marked to be observed and forwards the results to the third stage. The third stage takes the parsed data and handles the file and network I/O.



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License.

The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>



Figure 1. NADS MiniSim simulator as a research platform

2.2. Data gathering for estimating driver affective states

The ability to dynamically detect drivers' affective states is crucial in predicting drivers' behavior and performance and guiding them to safer driving. So far, our team has developed a real-time affect detection system using facial expression [7], and empirically tested heart rate [8] and oxygen concentration level of prefrontal lobe of the drivers [9] as an index of affective states (e.g., angry). Our team has intensively used eye-tracking devices to detect a driver's distraction [10], and we also expect that the eye-tracker can detect drivers' cognitive tunneling while they are absorbed in a certain affective state. We would like to extend a range of affective states and appropriate sensors. All these techniques have been used separately. To more effectively estimate drivers' affective states, we plan to integrate all these sensing data by developing data fusion algorithms.

3. SONIFICATION MAPPING

This novel system provides researchers with the ability to develop and test auditory displays that reflect situations in the driving simulator in real-time. In the sonification mapping process, we also need to consider other variables, such as human factors and system factors [11] as well as musical parameters (Table 1). For real-time sonification, we use Pure Data (Pd), which is the free real-time graphical dataflow programming environment. Other researchers can also use our Pd patch to test their own driving sonification. The data obtained from the simulator and neurophysiological devices are sent to Pd via Open Sound Control (OSC), which is a commonly used protocol for networking between sound synthesizers and other data.

To make our sonification algorithms, we first define the observation states (i.e., affective states and driving behaviors) and sonification parameters as input measures, and sonification as an output measure. We can define the total observation states function as

$$ObservationStates = AS(FacialExpression, EyeMovementPattern, HeartRate, Respiration, Skinconductance, BrainWaves, Brainactivities) \times DB(LaneDeviation, SteeringWheelAngle, Speed, PedalForce, Collision)$$

Next, we can define the total sonification parameters function as

$$SonificationParameters = MP(Genre, Key, Tempo, Chord) \times HF(Familiarity, Preference, Expectation) \times SF(Type, Duration, Timing, Regularity, Interference)$$

Our goal is to find the best profile of variables for each observation state and then, identify the best profile of sonification parameters to effectively manage driver affective states and driving performance. Therefore, sonification outputs function can be defined as

$$SonificationOutputs = f(ObservationStates \times SonificationParameters)$$

We also plan to expand our system to interact with any number of driving simulators of any level of fidelity and many more physiological sensors and eye trackers. For the applications to real cars, we plan to gather driving data via CAN (controller area network) bus.

Table 1. Mapping variables for observation states and sonification parameters

Observation States		Sonification Parameters
Affective States (AS)	Driving Behaviors (DB)	
- FacialExpression	- LaneDeviation	Musical Parameters
- EyeMovementPattern	- SteeringWheelAngle	- Genre
- HeartRate	- Speed	- Key
- Respiration	- PedalForce	- Tempo
- SkinConductance	- Collision	- Chord
- BrainWaves		Human Factors
- BrainActivities		- Familiarity
		- Preference
		- Expectation
		System Factors
		- Type
		- Duration
		- Timing
		- Regularity
		- Interference

4. SONIFICATION STRATEGIES

Researchers in the 1st in-vehicle auditory interaction workshop at ICAD 2015 [12] have discussed different types of sonification strategies appropriate for in-vehicle situations. Based on the discussion, we classify different strategies into three main areas: continuous soundscapes, discrete auditory displays, and target matching auditory displays.

4.1. Continuous Soundscapes

Continuous soundscapes are background or atmospheric sounds embedded with information describing mostly low urgency or non-time-sensitive information. Soundscapes allow for a large amount of data to be aggregated and summarized via sound parameter mappings. They have been found to be easily distinguished from background noise, while non-intrusive enough to be easily ignored by the listener when other tasks demand attention [13]. Soundscapes are often reported as informative and relaxing, especially when comprised of natural sound samples (i.e., running water, animal calls, wind rustling tree leaves, etc.).

For example, certain animal sounds could represent traffic in adjacent lanes. The sound of a swarm of crickets could appear panned and faded in the direction of the vehicle to make the driver aware of the potential obstacle when changing lanes.

4.2. Discrete Auditory Displays

Discrete auditory displays come in many forms, but can represent a specific action or event. For example, an earcon representing a takeover requests for a semi-autonomous vehicles would fall into the category of discrete auditory displays. A few ambitious vehicle manufacturers have already introduced lane departure warnings and lane change assistance auditory displays. Before these and other type of hazard detection auditory displays become a standard in all new vehicles, further testing is required to design a system that is both optimally informative and non-intrusive.

4.3. Target Matching Auditory Displays

Target matching auditory displays can inform the driver of both the current and optimal state of individual driving tasks. Lane keeping and economical driving are two areas that would benefit from this type of sonification strategy. For instance, the driver's music could be panned to either side to indicate the vehicle's deviation from the center of the lane. The target would be for the driver to keep the music panned to the center indicating that the vehicle is in the proper lateral position (center) of the lane. The music, or any type of audio the driver is listening to, would be panned in the opposite direction of the vehicle's position in the lane. For instance, when the vehicle drifts towards the right side of the lane, the music would be panned to the left, to indicate both the current and target (optimal) lateral lane position. Of course, this can be created based on the principle of gamification.

5. CONCLUSION & FUTURE WORKS

Although our in-vehicle interactive sonification system is in its nascent stage, the potential application for sonification research is immense. Not many driving simulators provide real time data routing for sonification purposes, and even fewer provide that data in an easy to use manner for non-programmers. Our software will be further developed to include an easy to use GUI for the user to select variables of interest, and network or port destinations for the data to be channeled to. The integrative data server will be developed to receive, log, and route other data streams from neurophysiological equipment. Considerable interest is directed toward affective computing in the driving domain given the performance deficits of particularly emotional drivers. Future works include a variety of evaluations of different sonification scenarios. Driving performance as well as user acceptance will be collected, as user opinion of the sonification system will be critical for widespread use.

6. REFERENCES

- [1] J. D. Lee, "Fifty years of driving safety research," *Human Factors*, vol. 50, no. 3, pp. 521-528, 2008.
- [2] A., Jarašūniene, and G., Jakubauskas, "Improvement of road safety using passive and active intelligent vehicle safety systems," *Transport*, vol. 22, no. 4, pp. 284-289, 2007.
- [3] R., Schoeneburg, and T., Breitling, "Enhancement of active and passive safety by future PRE-SAFE systems," In *2005 Conf. on ESV*, Washington DC, pp. 05-0080, 2005.
- [4] C. D., Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, pp. 159-177, 2002.
- [5] T., Kramer, "An introduction to auditory display," *Auditory Display: Sonification, Audification, and Auditory Interfaces*, In G. Kramer (Ed.), Addison-Wesley, 1-77, 1992.
- [6] S. Fakhrosseini, P., Kirby, and M., Jeon, "Regulating drivers' aggressiveness by sonifying emotional data. In *Proc. of the 21st Int. Conf. (ICAD 2015)*. New York, USA, July 2015.
- [7] M. Jeon, and B. N. Walker, "Emotion detection and regulation interface for drivers with traumatic brain injury," In *Proc. of the Int. Conf. (CHI'11), Diversity Workshop*, Vancouver, BC, Canada, May 7-12, 2011.
- [8] S. Jansen, A. Westphal, M. Jeon, and, A. Riener, "Detection of drivers' incidental and integral affect using physiological measures," In *Adjunct Proc. of the 5th Int. Conf. (AutomotiveUI'13)*, pp. 97-98, Eindhoven, The Netherlands, October 27-30, 2013.
- [9] S. M. Fakhrosseini, M. Jeon, and R. Bose, "Estimation of drivers' emotional states based on neuroergonomic equipment: An exploratory study using fNIRS," In *Proc. of the 7th Int. Conf. (AutomotiveUI'15)*, Nottingham, UK, September 1-3, 2015.
- [10] S. M., Fakhrosseini, M. Jeon, P. Lautala, and D. Nelson, "An investigation on driver behaviors and eye-movement patterns at grade crossings," In *Proc. of the Joint Rail Conf. (JRC2015)*, San Jose, CA, March 23-26, 2015.
- [11] M. Jeon, "A systematic approach to using music for mitigating affective effects on driving performance and safety," In *Proc. of the 14th ACM Int. Conf. (UbiComp'12)*, ACM Press, pp. 1127-1132, Pittsburgh, USA, September 5-8, 2012.
- [12] M. Jeon et al., "Proceedings of the in-vehicle auditory interactions workshop," In *Proc. of the 21st International Conference on Auditory Display (ICAD2015)*, Graz, Austria, July 8-10, 2015.
- [13] K. Wolf, G. Gliner, and R. Fiebrink, "A model for data-driven sonification using soundscapes," In *Proc. of the 20th Int. Conf. (IUIIC)*, ACM Press, 2015.

ACCESSIBLE SPECTRUM ANALYSER

Fiore Martin, Oussama Metatla, Nick Bryan-Kinns and Tony Stockman

Centre For Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, E1 4NS
UK
t.stockman@qmul.ac.uk

ABSTRACT

This paper presents the Accessible Spectrum Analyser (ASA) developed as part of the DePic project (Design Patterns for Inclusive collaboration) at Queen Mary University of London. The ASA uses sonification to provide an accessible representation of frequency spectra to visually impaired audio engineers. The software is free and open source and is distributed as a VST plug-in under OSX and Windows. The aim of reporting this work at the ICAD 2016 conference is to solicit feedback about the design of the present tool and its more generalized counterpart, as well as to invite ideas for other possible applications where it is thought that auditory spectral analysis may be useful, for example in situations where line of sight is not always possible.

1. INTRODUCTION

The Design Patterns for Inclusive Collaboration (DePIC) project aimed to develop new ways for people to interact with each other using different senses, so reducing barriers caused by visual and other sensory impairments (depic.eecs.qmul.ac.uk).

The development of the Accessible Spectrum Analyser, ASA, came out of our collaboration with visually impaired (VI) audio engineers and musicians, some of the results of which, including the development and evaluation of an Accessible Peak-level Meter, APM, were reported at the ICAD 2015 conference [1]. The requirement for the ASA came as a result of VI audio engineers explaining the need for an accessible means of examining the power in specific frequency bands of the audio signal being edited. However, this is only one possible application area for an ASA. Visually impaired school students studying GCSE and higher level Physics, electronics and engineering courses have the need to be able to view spectra of various kinds. Clearly spectral analysis, as a technique, is applied in numerous scientific and engineering applications. We are planning to develop a more generalized version of the tool for use as a sensory substitution device in science and engineering education and practice. Furthermore, there may be potential for applications of the tool by sighted users who need to monitor spectra without having line of sight to the visual display.



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

The ASA we have developed is a real-time spectrum analysis plug-in that allows visually impaired users to examine spectrograms using data sonification. The aim of reporting this work at the ICAD conference is to solicit feedback about the design of the present tool and its more generalized counterpart, as well as to invite ideas for other possible applications where it is thought that auditory spectral analysis may be useful.

2. SONIFICATION DESIGN

The sonic representation of spectrograms in this software is primarily designed to enable the user to monitor a specific frequency band to determine when, and to what extent power in that band exceeds a user-specified threshold.

The sonification uses a pitch mapping with an up-up polarity, that is, the higher the pitch the higher the peak level of power within the chosen frequency band. The user specifies the frequency range to be monitored, and sets the threshold level in dB, and, as soon as the energy of any frequency in the selected frequency band goes past the threshold, the ASA sounds a short beep. The beep starts at 440 Hz and it is raised one semitone for each dB of difference between the frequency magnitude and the user specified threshold. In case more than one frequency within the selection is higher than the threshold, the highest one is represented in the sonification.

If the ASA is used to monitor a stereo signal, the two channels are mixed together before being analysed. In the ASA, the sonification is panned from left to right, where the pan position represents the location in the spectrum of the peak frequency with respect to the whole spectrum, ranging from 20 Hz on the very left, to 20050 Hz on the very right. For example, if the selected spectrum peaks at 50 Hz, then the beep is presented towards the left, whereas if the peak is at 20 kHz, the beep will occur towards the right.

It has been helpfully pointed out by a reviewer of this submission that parameter mappings which make use of panning must be used with care. It restricts the end use cases to where a listener is positioned within the sweet spot, or tethered to a DAW via headphones. This will need to be borne in mind if we wish to optimize the mapping for situations where 'line-of-sight' is compromised, because in that case panning may fail to produce a meaningful display for the listener, if they are not facing the stereo speaker array or are too far away for headphones. One solution may be to encode the data redundantly with another parameter. For example the rate

of beeping could increase as frequency peaks approach the upper end of the spectrum.

3. PARAMETERS

The plug-in comes with five tweakable parameters:

1. **Threshold:** sets the threshold in dB for the spectral sonification. If any frequency within the selection is higher than the threshold, then the plug-in will emit a beep;
2. **Selection Start:** sets the starting point, in Hertz, of the selection. Frequencies within the selection will be monitored for peaks;
3. **Selection Size:** sets the size of the selection, from the starting point. For example, if the selection starts at 1000 Hz and the selection size is 500 Hz, then all the frequencies between 1000 and 1500 Hz will be monitored for peaks;
4. **Dry:** controls the level of the input audio, namely the audio content to be analyzed;
5. **Wet:** controls the level of the sonification.

The Accessible Spectrum Analyser provides access to the parameters by exposing them to inspectors - such as the ReaAccess plug-in (used with the Reaper DAW) or the Cakewalk Sonar inspector - in a clear and well formatted way.

4. EVALUATION

Because the development of the ASA came at the very end of the DEPIC project that funded it, a full evaluation of the tool has not so far been possible. However, information about the plug-in has been widely distributed on email lists for visually impaired musicians and audio engineers. The take up of the tool has been fairly limited, but those musicians and audio engineers who have adopted it have been very enthusiastic about its functionality, citing the fact there are currently no similar tools that provide direct access to signal spectra through audio currently. The few criticisms that have been received have reflected the reviewer feedback described above, that the panning functionality is not welcomed by all users.

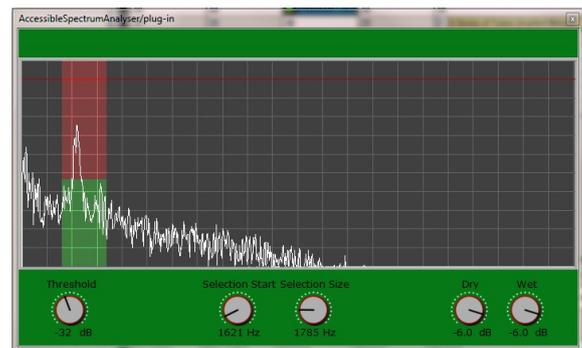
The accessible spectrum analyser was downloaded a total of 109 times from outside the Queen Mary campus in February and March. The bulk of these downloads occurred in March, after the announcement of the plug-ins release on relevant forums and mailing lists at the end of February. This provides unspectacular but solid evidence of the demand for this early-stage version, which has so far only been advertised to one subset of the possible target population.

5. FUTURE WORK

As mentioned above, we hope to develop a more functional and flexible version of the ASA, capable of providing auditory access to other commonly used signal measures, such as auto and cross correlation, coherence, cepstrum etcetera. To this end, we will investigate currently available open source, freely available packages for signal analysis, such as Octave or those available from Physionet, to provide a signal analysis engine on which to base a more general accessible signal analysis front-end.

6. OBTAINING THE SOFTWARE

The binaries for the ASA software, along with other software releases from the DEPIC project, can be downloaded from [2]. This work is licensed under the Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at [3]. The source code of the Accessible Spectrum Analyser is available in the SoundSoftware repository, located at [4]. It is released under the Cokos WDL license, which in short means you can alter it and redistribute it freely, even without providing the source code of your derivative work.



(a) Spectrum Analyser

7. REFERENCES

- [1] Metatla, O. Bryan-Kinns, N. Stockman, T and Martin, F. Sonifications for Digital Audio Workstations: Reflections on a Participatory Design Approach, Proceedings of ICAD 2015.
- [2] <http://depic.eecs.qmul.ac.uk/?q=software>
- [3] <http://creativecommons.org/licenses/by-nc/4.0>
- [4] <https://code.soundsoftware.ac.uk/projects/asa>

MUSIFICATION OF SEISMIC DATA

Ryan McGee

Media Arts and Technology
University of California, Santa Barbara
ryan@mat.ucsb.edu

David Rogers

College of Fine Arts
University of New South Wales
dvr@allshookup.org

ABSTRACT

Seismic events are physical vibrations induced in the earth's crust which follow the general wave equation, making seismic data naturally conducive to audification. Simply increasing the playback rates of seismic recordings and rescaling the amplitude values to match those of digital audio samples (straight audification) can produce eerily realistic door slamming and explosion sounds. While others have produced a plethora of such audifications for international seismic events (i.e. earthquakes), the resulting sounds, while distinct to the trained auditory scientist, often lack enough variety to produce multiple instrumental timbres for the creation of engaging music for the public. This paper discusses approaches of sonification processing towards eventual musification of seismic data, beginning with straight audification and resulting in several musical compositions and new-media installations containing a variety of seismically derived timbres.

1. INTRODUCTION

Supported by the Australia Council for the Arts Music Board, the original goal of our research was to accentuate sonic differences in the audification of individual seismic events around the world whilst searching for musical qualities within the sounds. Work began during a residency at the AlloSphere Research Center at the University of California, Santa Barbara - a unique venue for multi-channel sound spatialization¹. In addition to a highly spatialized, multi-channel work for the AlloSphere, two pieces of stereo, electronica-style music were produced that could be used to engage the public and have since received over 45,000 plays on SoundCloud. These electronica pieces, coupled with a new live-streaming seismic sound engine have been used in long-term public installations (Section 4).

Starting by replicating previous seismic audification techniques[1] [2] [9] using MATLAB, our sound processing grew to include a number of granular and frequency-domain effects to obtain a greater variety of sound timbres produced from a single seismic recording. We then used extremes of granular processing and frequency-domain filtering to accentuate sonic differences between separate seismic events as well as separate recordings of the same event.

¹<http://allosphere.ucsb.edu/>

 This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

2. AUDIFICATION OF SEISMIC DATA

Since seismic recordings are a form of physical data[3] obeying the general wave equation, the process of making them audible is simply a matter of rescaling seismometer data recordings to the range of digital audio samples, [-1.0, 1.0], and playing them at rates fast enough enter our range of hearing (20Hz - 20kHz). The typical range for seismic waves is 0.1-3 Hz, so increasing the playback rate by a factor of 100-1000X is common.

Seismometers record activity 3-dimensionally along vertical, East-West, and North-South axes as shown in Figure 1. Seismic recording stations categorize their recordings by sample rate, gain sensitivity, and orientation. For instance a channel code of **BHZ** would indicate a **b**road band, **h**igh gain, vertically (**z**-axis) orientated recording. Broad band channels are indicative of a 10Hz-80Hz sampling rate (specified in the header of each recording), which are desirable for audification as they are the highest available sample rates for any event. Likewise, high gain channels are desirable to produce more amplitude resolution in the resulting sounds. Differences amongst the audification of separate axes are subtle (Figure 2), so the vertical, Z, channel is typically used by default, but it is possible to synchronize and mix the audifications of all 3 orientations together to produce a slightly fuller sound.

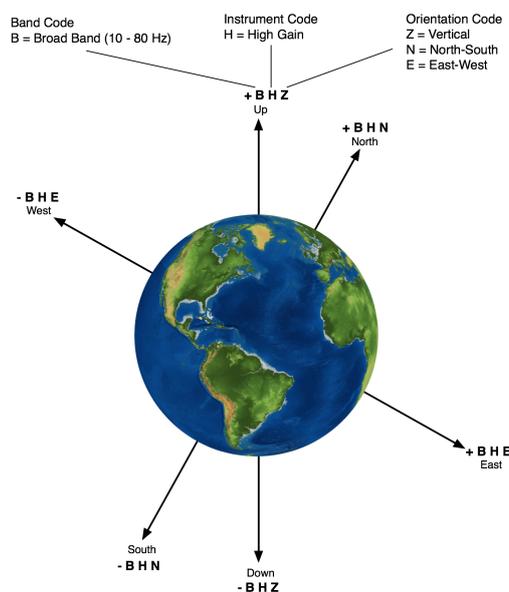


Figure 1: Orientation of Seismic Recording Channels

Our experiments began with data collected from the February 21st, 2011 magnitude 6.1 Christchurch, New Zealand Earthquake. IRIS (Incorporated Research Institutions for Seismology)², provides an online interface for accessing a database of current and past seismic events around the world³. Figure 2 shows the original waveform produced from audification at 276 times the original speed of the Christchurch event followed by examples of the sound processing techniques described in the following sections.

3. SEISMIC SONIFICATION TECHNIQUES

Wanting to explore more sonic variation for each seismic event without straying from the data, we devised several means of audio processing without using parameter mapping[4] so that the seismic data sets would remain as the only sound generators. This is an important distinction within the field of sonification since most techniques involve the mapping of data to parameters of subjectively chosen sound generators. With this work, the original seismic waveforms are accelerated and scaled to generate sound, and variety is achieved by resampling, filtering, granulation, time-stretching, and pitch shifting. Since the definition of audification limits processing to resampling, scaling, and minor filtering[3] we consider any further modifications to the sound as entering the realm of sonification. The following processes were used heavily in the creation of the *Christchurch* and *Haiti* compositions (Section 4).

Several sound examples to accompany the following sections are online at <http://i-e-i.wikispaces.com/Auditory+Display>.

3.1. Synchronous Granulation

Granulation of sound is the process of slicing a sound into several sound “grains” creating segments lasting 1 to 100 milliseconds[5]. If the grains are played back in order then the original sound results. One can repeat each adjacent grain a number of times to result in a new sound 5 times longer than the original. Choosing an arbitrary duration for each repeated grain will result in several discontinuities in the sound. For example, a grain’s amplitude may start at 0.23 and end on -0.72. When repeated, this jump in amplitude would produce an undesirable click in the sound. To solve this a short amplitude window (envelope or ramp) is applied to each grain so each always starts and ends at amplitude 0. Next, the grains are overlapped so there is less audible beating from the windowing. In this work, synchronous granulation refers to methods of granulation that involve several grains of identical duration and windowing. Audified earthquakes can be characterized by an initial high frequency, high amplitude sound that decays over time like hitting a snare drum. Synchronous granulation has the effect of time-stretching these sounds, repeatedly emphasizing each grain, which emphasizes the unique decay of each earthquake. However, a drawback is that the amplitude windowing can produce additional low-frequency beating artifacts in the sound.

3.2. Asynchronous Granulation Based on Zero-Crossings

Asynchronous granulation implies that each sound grain will have different characteristics. In our case, the duration of each grain varies over time based on an algorithm that chooses the start

and end points for each grain based on the location of zero crossings within the sound, which are points where the wave’s amplitude is equal to 0. Zero crossings may occur often, sometimes even less than 1ms apart, so a minimum duration for each grain is also specified. The convenience of using zero crossings is that windowing is not needed since the grains will already start and end on 0. Another quality of zero crossings is that they usually indicate the beginning of an impulse or large transient within the sound. When asynchronous grains are repeated multiple times, the transient, impulsive portions of each earthquake are emphasized, creating stuttering rhythms unique for each event.

3.3. Time-Stretching, Pitch Shifting, and Filtering via Phase Vocoding

The phase vocoder[6] is a complex process used for spectral analysis and resynthesis, allowing for frequency-domain filtering. Its process breaks a sound into multiple segments of equal duration and uses a Fast Fourier Transform to analyze the frequency spectrum of each segment. One may interpolate multiple spectra between two segments to extend the duration of a sound while maintaining its frequency content. If one time-stretches a sound in such a fashion and then alters the playback rate, the result becomes a change in pitch without a change in duration (unlike audification). The spectra of each segment can also be manipulated to apply filtering effects. Removing all frequencies below a certain amplitude threshold has the effect of de-noising a sound, leaving only the most prominent frequencies. This de-noising can be taken to extremes to leave only a few partials in each sound, ultimately producing unique tones and chords for seismic events.

4. SEISMIC COMPOSITIONS AND INSTALLATIONS

4.1. *Christchurch* (2012)

<https://soundcloud.com/seismicsounds/christchurch-earthquake>

Christchurch uses a single seismic recording from the nearest station to the February 21st, 2011 Christchurch earthquake. The piece begins with a build-up of several reversed audifications of the event, time-stretched at different speeds. Then, the strong impact from the raw audification is heard, followed immediately by a chaotic granulated version emphasizing the loudest points in the impact. A tone fades in that is an extremely time-stretched, pitch shifted version of the event with all but the most dominant partials filtered out of the sound. Other versions of this tone eventually overlap at manually coordinated harmonic pitch intervals. The event is played back using several different time-stretch factors and synchronous granulations during the course of the piece. Timing becomes more ordered and apparent until rhythmic granulations and tones lead to a final build-up, ending with another raw audification. The result is an exploration of timbral variety from a single seismic recording as the only sound generator.

4.2. *Haiti* (2012)

<https://soundcloud.com/seismicsounds/haiti-earthquake-12th-january>

Haiti uses seismic recordings of the 12th January, 2010 magnitude 7.0 Haiti event and explores variety of sounds produced from

²<http://www.iris.edu>

³http://ds.iris.edu/wilber3/find_event

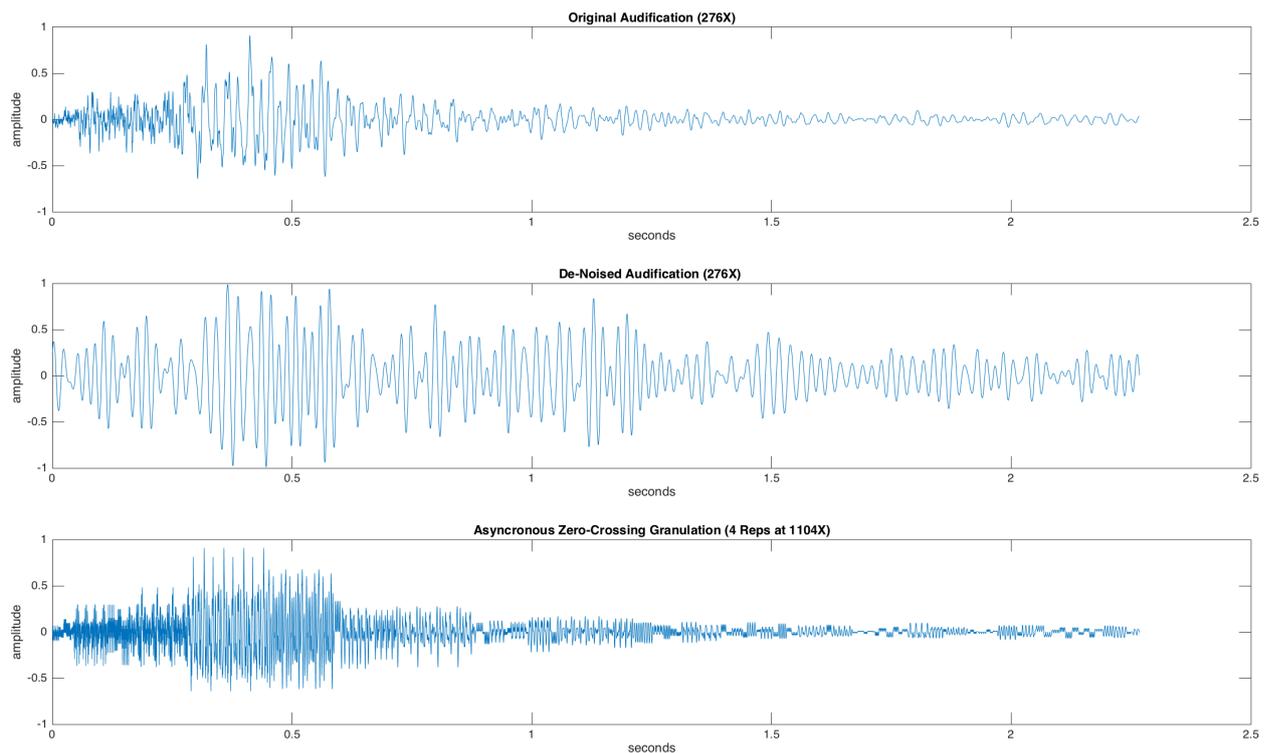


Figure 2: Audification Processing: De-Noising and Asynchronous Granulation

the same event recorded by the nearest 12 stations. The piece begins with granulated audifications of each station played in succession from the furthest to nearest station at the same playback rate. A brief recording of sensor noise from each station is played along with the granulations. Since each station has its own distinct sound, these noises represent signatures of each station as the listener moves nearer to the quake. A slow melody plays in the background that was generated by filtering out all but the most prominent single frequency from the spectra of each noise signature. A rhythmic sound in the background was generated from the impact of each station played back at high-speed in succession again from furthest to nearest. As this rhythm gradually increases in intensity a lower noise grows in the background, which is a time stretched recording of the impact played in reverse to further emphasize the backwards (far to near) build-up of the piece. At the climax the impact from the nearest station plays, followed by impacts from the other stations – this time increasing in distance. In the background the low growl of the time-stretched recording fades away. The piece becomes chaotic after the main event using asynchronous granulation based on zero-crossings.

4.3. *Shadow Zone Shadows* (2012)

<https://soundcloud.com/seismicsounds/shadow-zone-shadows>

Shadow Zone Shadows was an abstract sound study that spatialized seismic audifications according to the geographic location of their impact and simulated their spatial traversal through the earth. This piece was presented in a dark space with no visuals within the AlloSphere, allowing the listener to imagine being

placed at the center of the earth while experiencing a series of earthquakes occurring and moving around them over a 3D loudspeaker array. A variety of international seismic events were used ranging in magnitude from 5.5 (Los Angeles) to 9.1 (Sumatra). The piece used custom spatialization software that allowed the programming of spatial sound trajectories synchronized with specific points in the seismic data (start of event and reflections). Just as the seismic data is accelerated to become an audible audification, the spatial trajectories used were simulated at speeds much faster than seismic waves actually propagate through the earth.

The AlloSphere uses a spherical 54.1 channel loudspeaker configuration consisting of 3 rings of speakers (lower, ear-level, and upper). For the sound spatialization we mapped the latitude and longitude of seismic events to virtual sound source points on the surface of the loudspeaker sphere. A distance-based amplitude panning [7] algorithm was used with custom C++ software to pan sounds between loudspeaker locations.

4.4. *DOMUS* (2014-15)

<http://domus.urbanaction.org>.

DOMUS was an experimental architecture installation incorporating spatialized seismic sound and light within a hexayurt (Figure 3). 6 mid range speakers, 4 subwoofers, and a 360 degree LED pixel sphere chandelier displayed live seismic audifications and the *Christchurch* and *Haiti* compositions continuously from 10am to 10pm for the 7-month period of installation (October 2014-May 2015) at Materials and Applications, Los Angeles⁴.

⁴<http://www.emanate.org/past-exhibitions/domus>

The low-frequency seismic compositions resonated throughout the two story hexayurt DOMUS structure, emphasizing their seismic nature while demonstrating an “architecture of life” - a biophilic design model that responds and reflects the natural world [8]. The light chandelier by Rene Christian⁵ visualized the audifications by mapping sound frequency and amplitude to light hue and brightness respectively.



Figure 3: DOMUS Seismic Sound and Light Architecture

4.5. Sounds of Seismic (2012-16)

Sounds of Seismic (SOS)⁶ auditory display streaming system broadcasts continuous seismic sound generated from realtime collected global earthquake data. Influenced by John Cage’s *Variations VII* (1966), SOS is a “Musique Concrete” like audio composition in which the score is algorithmically generated by seismic waveform data. SOS can also be used as a listening tool for earth scientists to listen to a specific sensor on the Global Seismic Network (GSN). The conceptual framework of SOS is to create greater social awareness of natural ecological systems by generating multi-channel seismic sound electronically creating an infinite computational earth system soundscape.

SOS is an ongoing long term project built on our C++ Earthquake Sound Engine (ESE) and custom Python seismic data acquisition scripts by Stock Plum. Real-time miniSEED data is collected from Incorporated Research Institutions for Seismology (IRIS) and piped through ESE using the techniques outlined in Sections 2 and 3. We seek collaboration with public or private institutions providing streaming audio services and digital legacy design to present this infinite computational seismic audio sound performance.

5. SUMMARY

Through the desire to find musicality within seismic data we have discovered that time-domain granulation and frequency-domain filtering techniques are especially useful for deriving timbral variety between otherwise similar seismic audifications. In particular, a de-noising filter which removes all frequencies below a variable amplitude threshold is useful to produce unique tones and chords for seismic events. Asynchronous granulation based on zero crossings emphasizes transients (impacts) and produces unique stuttering rhythms for events. While the phase vocoder and synchronous granulation provide other means to time-stretch sounds, the required windowing and spectral interpolation will depart further

from the original sound characteristics in comparison with granulation based on zero crossings.

While many combinations of the aforementioned sound processing techniques are possible, we emphasize that the seismic data is ultimately the only source of sound *generation*. Because seismic data is physical wave data, using time-stretching and pitch-shifting allows us to magnify and focus on qualities already present in the data without mapping to arbitrary sound generators. We consider the ability to produce multiple timbres from a single data set without parameter mapping crucial to exploring the variety of musicality naturally present within seismic events. We have sought to create enough timbral variety necessary to produce every “instrumental” part of engaging pieces of electronic music via processing rather than mapping, so seismic data remains as the original, sole sound generator.

The ultimate goal of this research is to create a generative, dynamic audification-based musak which can highlight resilience and awareness of the natural world in which we inhabit. For inspiration we have looked back to the work of Haywards[1] and Dombois[9] and look forward to creating a music that is both a meaningful tool for geophysics monitoring as well as an engaging means of raising public seismic awareness. Ongoing work with SOS (Section 4.5) provides a platform for endless, live seismic data accompanied by real-time sound processing and, eventually, music generation.

6. ACKNOWLEDGMENTS

The authors would like to thank the Australian Council for the Arts Music Board and AlloSphere Research Group for supporting this research.

7. REFERENCES

- [1] C. Hayward, “Listening to the earth sing,” in *Auditory Display: Sonification, Audification, and Auditory Interfaces*. Addison-Wesley, 1994.
- [2] S. Speeth, “Seismometer sounds,” in *Journal of the Acoustical Society of America*, vol. 33, 1961, pp. 909–916.
- [3] F. Dombois and G. Eckel, *Audification*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin, Germany: Logos Verlag, 2011.
- [4] F. Grond and J. . Berger, *Parameter Mapping Sonification*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin, Germany: Logos Verlag, 2011.
- [5] C. Roads, “Granular synthesis,” in *Microsound*. MIT Press, 2001, ch. 3.
- [6] —, “Spectrum analysis: The phase vocoder,” in *The Computer Music Tutorial*. MIT Press, 1996, ch. 13, pp. 566–577.
- [7] T. Lossius, P. Baltazar, and T. de La Hogue, “DBAP-Distance-Based Amplitude Panning,” in *Proceedings of 2009 International Computer Music Conference, Montreal, Canada*, no. 1, 2009.
- [8] M. M. Stephen R Kellert, Judith Heerwagen, *Biophilic Design: The Theory, Science, and Practice of Bringing Buildings to Life*. Wiley, 2008.
- [9] F. Dombois, “Using audification in planetary seismology,” in *Proceedings of the International Conference on Auditory Display*, 2001.

⁵<http://renechristen.net>

⁶<http://www.sos.allshookup.org>

MALLO MARCH: A LIVE SONIFIED PERFORMANCE WITH USER INTERACTION*KatieAnna E. Wolf*

Department of Computer Science
 Princeton University
 Princeton, NJ, USA
 kewolf@princeton.edu

Reid Oda

Department of Computer Science
 Princeton University
 Princeton, NJ, USA
 roda@princeton.edu

ABSTRACT

In this extended abstract we present a new performance piece titled *MalLo March* that uses MalLo, a predictive percussion instrument, to allow for real-time sonification of live performers. The piece consists of two movements where in the first movement audience members will use a web application and headphones to listen to a sonification of MalLo instruments as they are played live on stage. During the second movement each audience member will use an interface in the web app to design their own sonification of the instruments to create a personalized version of the performance. We present an overview of the hardware and interaction design, highlighting various listening modes that provide audience members with different levels of control in designing the sonification of the live performers.

1. INTRODUCTION

In a classical concert performance, the audience plays a passive role by sitting and listening to live performers as they play acoustic instruments. With the advance of technology, new areas of musical performance have been developed that utilize the additional computational power to create live digital musical instruments. These instruments can be played by live performers, using various digital sensors (as in the EyeHarp, an eye-controlled instrument [1]), or by using live data streams as the “player” such that changes in the data trigger changes in the sound (as in *Leech*, a sonification of BitTorrent traffic [2]). In both contexts, the information provided by the performer or the data is sonified to produce a live real-time performance. When the information is provided by a live performer (rather than by data), the timing of the sonification is important as audience members expect the visual cues of the performer to synchronize with the sonified audio.

This synchronization is challenging because it takes time for information to be transmitted and for audio to be synthesized. This is particularly true in cases where the information needs to be transmitted long distances (i.e. around the world via the Internet), or when the audio synthesis is complex. Researchers have overcome these issues by making local audio processing as fast as possible and minimizing network transmission times [3]. However, there is a limit to the time that can be gained through these methods. A digital instrument called MalLo [4] was developed to

overcome the challenges of performing over long distances by predicting the strike of a percussion instrument before it occurs. The prediction is sent over the network so that the strike will be sonified at both the sending and receiving locations simultaneously. Rather than using MalLo to overcome long distances, in this work we propose to use MalLo as a way to overcome long processing times associated with complex audio synthesis between performers and a local audience.

Our piece *MalLo March* features multiple MalLo instruments played live on stage. During the performance, audience members use their own network-capable computing devices (laptops, smart phones, tablets, etc.) and headphones to design and listen to the live sonified performance. We intend to distribute in-ear headphones and headphone jack splitters so that everyone has the chance to participate, even if they do not have a device. The performance opens with *Movement 1* where audience members are asked to navigate to a URL using their web browser. Once the performers begin playing the MalLo instruments, audience members will be able to hear synthesized audio of the predicted strikes via their headphones and the web app hosted at the URL. During the first movement audience members will not have control over the design of the sonification. However, once *Movement 2* begins they will be able to interact with a user interface on the web app to change the sonification in various ways and manipulate the sounds of the MalLo instruments. By opening up the sonification design process to audience members they be able to explore the sonic possibilities of the performance.

In this extended abstract, we briefly describe the technical design of the performance by outlining the way the hardware of MalLo works and the types of interactions that will be available to audience members via the user interface. We also include a brief discussion on what we hope to gain from the performance.

2. HARDWARE DESIGN

In order to gain time with which to execute computational processes we use a predictive instrument called MalLo [4]. As shown in Figure 1, MalLo predicts note times and note velocities before they are played by capitalizing on the long distance traveled by percussion mallets. First, we track the head of a percussion mallet. Next, we fit a quadratic regression to the mallet’s path (taking advantage of the fact that its path is very predictable [5]). We compute the time at which the mallet will strike the surface, and we send this information to the receiver. The receiver takes in a continuous stream of predictions, each one more accurate than the previous. The receiver waits until the predicted time reaches a specified accuracy threshold and sonifies the note. During the period



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

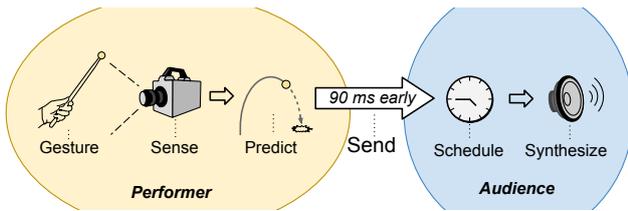


Figure 1: Our musical instrument tracks the head of a percussion mallet and predicts what time it will strike. This allows up to 90 ms of anticipation time which can compensate for time spent transmitting note information or processing of advanced synthesis algorithms.

between receiving the predictions and the final sonification, the receiver can be performing a variety of computations (e.g. complex audio synthesis algorithms).

Our system sends its timing predictions as absolute time messages, in Unix time (time since January 1, 1970). This allows for nearly unlimited scaling. The system clocks of the sender and receiver must be synchronized to a high degree. In previous work MalLo relied on GPS satellite signals to allow for time synchronization over long distances. However, for this performance we will use a variant of Precision Time Protocol (PTP) for time synchronization [6]. Also, because we will not have access to the system clock of every audience machine, we will implement it in our web app, on the client side.

Each audience listener will be required to listen via a web app that they access on a personal computing device. As shown in Figure 2, MalLo prediction messages are sent to each device where they are scheduled, and the audio is synthesized.

3. INTERACTION DESIGN

Audience members will navigate to the web app URL at the start of the piece. Using their own personal networked device and headphones, they will listen to the sonification of the predicted MalLo notes that are being sent over the Internet. During *Movement 2*, the audience members will see a user interface designed to give them control over the sonification of the performance. We imagine three types of interaction modes that allow for various levels of control of the sonification. In each of these modes, the incoming data to the sonification will be the live performance data sent from MalLo (note timing information and velocity), and the output will be the sound from the sonification algorithm chosen via the particular interaction mode.

Interaction Mode 1: Change the Channel In this mode, users are allowed to switch between preset sonifications similar to changing the channel on the radio and hearing the same song performed by different combinations of instruments. For each channel, the live data from the performers will be input into a different sonification to create a unique sonic performance.

Interaction Mode 2: Change the Instruments To give users slightly more control, in this mode users are able to choose a different preset sonification for each of the MalLo instruments. By doing so users are able to “Change the Channel” for each instrument.

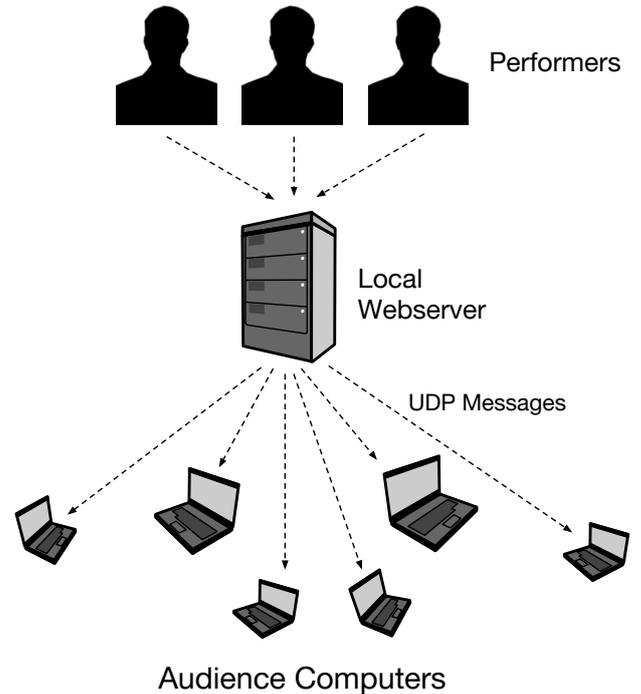


Figure 2: Note information is sent from a local webserver to each audience member’s web browser where. There the notes are scheduled and synthesized according to the parameters of the sonification.

Interaction Mode 3 - Create the Instruments At the lowest level of control, users are able to directly modify the synthesis parameters for each instrument. In particular they can alter pitch, duration, waveshape, samples etc.

4. DISCUSSION

MalLo March is a piece intended to combine live sonification of performers with end-user design of sonifications. With permission from each audience member, we hope to log information about how they used the interface during the performance and do a short survey at the end. In particular we are interested in observing how people interact with the various interaction modes, how successful they felt they were in creating sonifications, and what was gained or lost by adding the interaction in the second movement. Our composition is intended to be enjoyable for the audience as well as give us insight into the design of sonifications.

5. ACKNOWLEDGMENT

This material is based upon work partially supported by the NSF GRFP under Grant No. DGE 1148900. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The work is also partially supported by the Princeton University Project X Grant.

6. REFERENCES

- [1] Z. Vamvakousis and R. Ramirez, “Temporal control in the EyeHarp gaze-controlled musical interface,” in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2012.
- [2] C. McKinney and A. Renaud, “Leech: Bittorrent and music piracy sonification,” in *Proceedings of the Sound and Music Computing (SMC)*, 2011.
- [3] J.-P. Cáceres and C. Chafe, “Jacktrip: Under the hood of an engine for network audio,” *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, 2010.
- [4] Z. Jin, R. Oda, A. Finkelstein, and R. Fiebrink, “Mallo: A distributed, synchronized instrument for internet music performance,” in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2015.
- [5] S. Dahl, “Striking movements: A survey of motion analysis of percussionists,” *Acoustical Science and Technology*, vol. 32, 2011.
- [6] K. Correll, N. Barendt, and M. Branicky, “Design considerations for software only implementations of the IEEE 1588 precision time protocol,” in *Conference on IEEE*, 2005.

INSTALLATIONS

Flight Variant

Teresa Connors, Andrew Denton

The University of Waikato, Auckland University of Technology
New Zealand

tmconnor@waikato.ac.nz

andrew.denton@aut.ac.nz

“Thus the art in the time of hyperobjects explores the uncanniness of beings, the uniqueness of beings, the irony and interrelationships between beings, and the ironic secondariness of the intermeshing between beings.” [1]

Artistic Work Submission

Flight Variant is one of a series of ongoing audiovisual installation projects by Teresa Connors and Andrew Denton, which respond to the Anthropogenic climate and geological change. The work emerges from data collection processes that took place in Southern California in 2014 and 2015. These include high-speed and HD video jet streams recordings (see figure 1, 2) and audio recordings from and around the Los Angeles airports. The resulting installation is a generative work that is driven by an algorithm based on 2015 aviation statistical data. Additional components include flight data streamed from the Internet, sampled vocal clips from YouTube, TV, and the Radio, real-time convolution of acoustic instrument improvisation with field recordings.

Constructed in Max 7 (see figure 3), this installation layers a network of visual and aural content that produces an ever-evolving work. The core visual elements are a series of strangely articulated filmed jet streams that cut lines across a rich blue Californian sky.

30 years ago Bill McKibben imagined a world where the sounds of chainsaws would inhabit even the most isolated and inaccessible forests. Now we can look around and out, up and down, and in every micro and macro

space on the planet, and silence and human absence has all but disappeared. [2]

Flight Variant is a series of core samples - a database of human presence and movement across the sky, land, and airwaves. Similar to other ecologically-grounded creative practices, this installation explores the situated relationship of environment, material agency and creative process and as means to tease out an emerge co-creative methodology.

Thematic Statement

Flight Variant seeks to evoke a space of contemplation, uneasiness, and melancholy by engaging with the stratified signs of our collective impressions and impacts on our environment.

Web Links

Link to *Flight Variant* and other works:
www.divatproductions.com/ICAD2016.html

Link to Andrew Denton’s CV:
<https://aut.academia.edu/ADenton>

Link to Teresa Connors’s CV and web page:
<https://waikato.academia.edu/TeresaConnors>
www.divatproductions.com



Figure 1 Flight Variant—installation image. Photo Andrew Denton



Figure 2 Flight Variant—installation image. Photo Andrew Denton



Figure 3 Flight Variant Max patch in presentation mode

Technical Requirements

Flight Variant is a sound and video installation that requires a darkened gallery space. Ideally this space would consist of a flat white wall for large screen projection four metres wide by three metres high (this can be larger if possible). We would need to be able to mount the projector, computer, and speakers to the walls and ceiling, or by other methods depending on the nature of the site. The artists can provide the HD projection and computer technology for generating and presenting the installation. It would be helpful for the gallery to provide audio speakers, AC (electrical cabling), internet connection and suitable security measures for the equipment. This can be negotiated as to what is appropriate depending on the site.

References

- [1] Timothy Morton, *Dawn of the Hyperobjects*. <http://www.youtube.com/watch?v=zxpPJ16D1cY>. (accessed November 20, 2014).
- [2] Bill McKibben, *The End of Nature*, New York: Random House, 1989.

- [3] Damian Keller and Ariadna Capasso, “New Concepts and Techniques in Eco-composition,” *Organised Sound*, Vol.11, No. 01, (2006): 55-62.

Artist Biographies

Andrew Denton is a film and video artist who works with both digital and analogue media. He is currently undertaking a PhD, at Monash University, investigating ecological issues through affective moving image and sound. Andrew has presented his research at numerous international festivals, conferences and symposia, including: NZ International Film Festival (2015), Jihlava International Documentary Festival (2015), ASLEC 2014, TESS 2013 & 2014, Balance-Unbalance 2013, ISEA 2012, and SIGGRAPH Asia (2009). Andrew is Head of Department Postgraduate Studies at the School of Art and Design at AUT University, in Auckland, New Zealand. <https://aut.academia.edu/ADenton>

Teresa Connors is active as an acoustic/electroacoustic composer, opera singer, film scorer, and multimedia installation artist. She is currently completing a practice-based PhD, at Waikato University, which includes developing new techniques and methodologies for multimedia collaborations. Teresa holds a Master of Music degree (1st class honours) in composition from Waikato University and studied both composition and opera singing at Dalhousie University in Nova Scotia as well as the Banff Center for the Arts. Her creative works have received awards and support from the Canada Council for the Arts, British Columbia Arts Council, Bravo Fact and have been presented at international conferences, film festivals, and galleries including: NZ International Film Festival (2015), Jihlava International Documentary Festival (2015), ISEA (2015), Balance-Unbalance (2015, 2013), TIES (2014 & 2013), EMS (2014), Vancouver International Film Festival (2010 & 2009). www.divatproductions.com

Flight Variant

An audiovisual installation by Teresa Connors and Andrew Denton

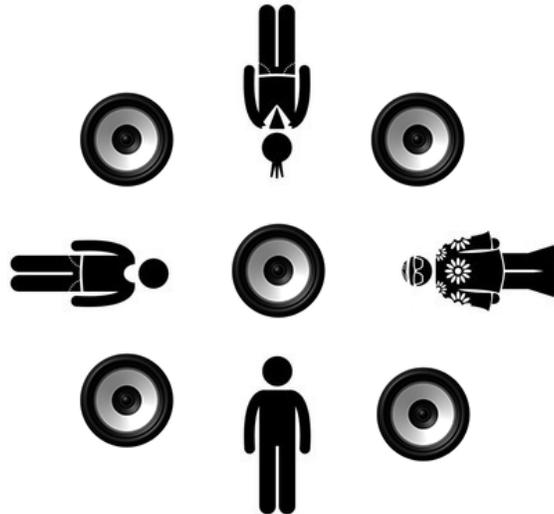
"The very feeling of wondering whether the catastrophe will begin soon is a symptom of its already having begun."

(Timothy Morton, *Hyperobjects : philosophy and ecology after the end of the world.*)



Flight Variant is one of a series of ongoing audiovisual installation projects by Teresa Connors and Andrew Denton, which respond to the Anthropogenic climate and geological change. The work emerges from data collection processes that took place in Southern California in 2014 and 2015. Constructed in Max 7 this installation layers a network of visual and aural content that affect each other simultaneously to produce an ever-evolving, iterative, work. The core visual elements are a series of strangely articulated filmed jet streams that cut lines across a rich blue Californian sky.

Terramomentum



Terramomentum (earth resonance) is an immersive, sub-sonic installation driven by seismic audification and musification. 4 blindfolded listeners lie on the floor as 5 subwoofers emit audifications and compositions derived from datasets of 3 major earthquakes: Haiti (2010), Christchurch (2011), and Kathmandu (2015). The installation repeats itself in an 8 minute cycle, allowing listeners to easily come and go as they please.

Seismic events (earthquakes) occur at sub-sonic frequencies below 20Hz. The emphasis of subwoofers and listeners on the floor allows for a strong haptic connection to the sounds - decreasing the amount of frequency shift necessary to perceive the original data sets. The installation is as much haptification as it is audification.

The authors can provide a computer, multi-channel audio interface, and amplifiers to be used for the duration of the installation. Ideally, the conference can provide:

- 5 subwoofers
- 5 x 5 meters of space in which to setup the installation

Example content:

- <https://soundcloud.com/seismicounds/christchurch-earthquake>
- <https://soundcloud.com/seismicounds/haiti-earthquake-12th-january>

Note: The authors have also submitted a paper detailing the seismic musification techniques to be used: Submission #21, Musification of Seismic Data.

Lux Mix

Lux Mix is an interactive sound installation that features modulated light as a medium for carrying audio signals. It needs to be installed in a quiet and relatively dimly lit area sheltered from strong artificial lights such as fluorescents. Seven small directional LED lights are arranged around two light detectors (about 20cm diameter hemispherical) mounted on a plain grey façade. The lights point forward into the space, not casting any light on the detectors. Near the entrance to the space is a plinth on which are placed an assortment of mirrors, mirror balls and other reflective objects with a sign describing how these are to be used:

Lux Mix is an interactive sound work.

Warning: This work involves flashing lights.

There are seven small lights and two white detectors.

To activate sounds, reflect the light back onto the detectors using any of the reflective objects in the space.

Touch, move and play with the reflective surfaces, but **please do not touch or adjust the lights or detectors** and do not bring other lights into the space.

When nobody is interacting with the work it is quiet, sinking into the background noise. When the lights are reflected back onto the detectors, the sounds encoded in those lights are heard through stereo speakers. Each light carries a different kind of sound, all created by non-repeating generative processes. The two detectors direct sounds to the left and right channels. Feedback of these audio signals back to the computer allows sounds to be generated differently depending on levels (and other analysed signal properties) received.

The lights flicker at audio rate, each modulated by a different audio signal. The audio signals are generated in real time by a computer. The two detectors turn audio rate light flickering back into audio signals, which are amplified and sent both directly to a pair of speakers and back to the computer for analysis.

The initial installation of Lux Mix was in the ANU School of Art Foyer Gallery, April 2016 as part of the Random9 exhibition "Light And Dark" (Random9 is an independent art group co-ordinated by Stephanie Parker). One light made magpie sounds if there was more light in the left detector or thrush sounds if there was more light in the right detector. Another light made a frequency sweep up (whoop whoop whoop) if there was more light in the left detector, and a sweep down (tchop tchop tchop) if there was more in the right, and it gradually got faster the more it was active. One light had voices - warmer words or colder words. There were lights for rhythms, melodies and drones, all synchronised and harmonising so you could mix them, and they would randomise when they were silent. The last light was feedback, so you could either make loud honking feedback sounds or you could use it in combination with the others to add a harsher quality.

For the ICAD installation the set of sounds would be reviewed and updated based on experience from the Foyer Gallery installation. For example, people tended to use one sound at a time and would find sounds more engaging if they were clearly able to control it. The sweep sound proved popular because it provided such understandable control. The bird sounds were also popular because they were pleasing sounds, perhaps also because they provided an unexpectedly natural element to the clean interplay of physical signals and processes. The more musical channels lost some impact because they were mostly not mixed with each other, so the relationships between the channels was less clear. Also the difference between the left and right detectors could have been clearer and more meaningful for the musical channels (melodies changed key, rhythms changed density).

The main goals of Lux Mix are:

- to invite play, experimentation, movement and music making

- to spur musing and questioning about the properties and possibilities of light, sound and generative audio.

The act of reflecting light onto a target requires some experimentation to find a light beam and see where it reflects. There is a challenge of dexterity to point the beam back to the detector, which is rewarded by the production of a sound. The continued requirement for dexterity in directing the beam to discover how the sound changes draws attentional focus into the task creating a real experience of playing an instrument in performance, and a conversation between the human operators and the generative processes they are driving.

This work operates on the same principle as “Sound Modulated Light” (2005) by Edwin van der Heide.

http://www.evdh.net/sound_modulated_light/

Sound Modulated Light is presented in a dark room with many modulated lights. Visitors each carry a battery powered box that detects light and amplifies it into headphones. Thus each visitor explores the available sounds by moving about the room and pointing their detector toward some combination of the lights. Lux Mix differs from Sound Modulated Light in a few significant ways. Firstly, to interact with Lux Mix, visitors choose from a range of reflective objects, each of which brings its own possibilities and challenges, so there is more to explore and discover in the physical interaction. Secondly, the hand held detectors in Sound Modulated Light do not offer the possibility of feedback, which is what makes Lux Mix more of a conversation than a broadcast. Another key difference is that the sounds in Lux Mix are played through speakers, not headphones, and it can be operated by more than one person at a time, making more of social, conversational medium.

Requirements:

Lux Mix needs to be installed in a quiet and relatively dimly lit area sheltered from strong artificial lights such as fluorescents, as most artificial lights flicker with a loud mains hum. Some leakage of undimmed incandescent lights or sunlight into the space can be less problematic, but the darker the space the better. The physical structure of the work can be adapted to different spaces. For its initial presentation in the ANU Art School Foyer Gallery it was installed in the plinth cupboard on a back plinth with a black backdrop and black fabric draped over the electronics behind the façade. I would expect to rebuild the façade and mounting to suit any new space, given a week or so to construct it. Only one power point is needed. It also requires a plinth about 1m tall to put the collection of reflective objects on.

Attachments:

Audio Files:

- Rhythm and Birds Demo.mp3
- Rhythm Demo.mp3
- Sweep Demo.mp3
- Vox Demo.mp3

Photographs:

- LuxMix_bedroom.jpg – photograph of the system at home before the first install.
- LuxMix_installed.jpg – selfie with the installed system
- LuxMix_diagram.jpg – rough sketches of the installation with approximate measurements.

TITLE OF CREATIVE WORK:

Native Modulations

DESCRIPTION / ABSTRACT:

Fascinated by the eye-catching 'bush graffiti' of the native Scribbly Gum, I felt an urge to decipher its mysterious language. With a background in audio production, I noticed similarities between these bark markings and the waveform shapes that are encountered when visualising sound. Taking detailed measurements, photos and video footage of the 'scribbles', I used these shapes to be primary modulators for an array of chosen sound sources.

The sounds you hear pulsing from the tree featured in the video are the result of synthesised tones being manipulated by the tree's markings. Each scribble has a distinctive shape and length that creates unique filtering and note duration. A base tuning of A4=432Hz is used, which is considered to be a harmonic intonation of nature and results in a more organic sound not usually found in everyday music. The outcome allows one to engage multiple senses in connecting with the deeper beauty of this iconic tree.

TECH REQUIREMENTS:

This video has been formatted with the intent to be displayed in full high definition on a screen which should be mounted at eye level and vertically i.e in portrait mode 9:16 at 1080 x 1920. Sound clarity is equally important so quality headphones or speakers will also need to be connected to the screen/display so the viewer can easily associate the sound with the imagery. The video has been edited to suit continuous looping/repeating playback.

AWARDS:

This piece received the Highly Commended Environmental Art Award and won the People Choice Award at the 2015 Sunshine Coast Art Prize New Media Category.

ABOUT THE ARTIST:

I am an emerging artist in nature-inspired audiovisual design - sound, images and technology. I focus my creative energies on encouraging and highlighting human connections with the natural world.

CONTACT:

Andrew Zylstra

Email: andrew@earthcollective.net

Phone: 0415 059 193

Research Through Design into Acoustic Sonification

Stephen Barrass

Faculty of Arts and Design
University of Canberra
Australia

stephen.barrass@canberra.edu.au

ABSTRACT

Acoustic Sonifications are physical objects shaped by digital datasets with the design to produce sounds that convey useful information about the dataset [1]. This installation allows hands-on explorations of three early experiments that introduce and demonstrate this concept. The installation includes iterations of each object that document the research process.

1. HRTF Bells

The first experiment, titled HRTF Bells, maps a Head Related Transfer Function (HRTF) dataset to the shape of a pair of Bells 3D printed in stainless steel. The differences in the left and right ear datasets are difficult to see, but can be heard immediately, when the bells are rung [2].



Figure 1. HRTF Bells

2. HYPERTENSION SINGING BOWL

The second experiment, titled Hypertension Singing Bowl, maps a year of blood pressure readings to the shape of a tibetan Singing Bowl that is 3D printed in stainless steel [3]. The need to represent 2 dimensional diastolic/diastolic datapoint led to a mapping that introduces tines with two levels of thickness. The bowl sings like a traditional singing bowl, but with a timbre that is changed by the dataset. Different datasets will produce bowls with different timbres.



Figure 2. Hypertension Singing Bowl

3. CHEMOTHERAPY SINGING BOWL

The third experiment, titled Chemo Singing Bowl, maps blood pressure taken over a year of chemotherapy to a Singing Bowl. This mapping aims to amplify the effect of the dataset on the acoustics of the bowl by mapping the diastolic/systolic pressure to control points for the curvature of 2D splines amped to the tines [4]. This bowl does not sing very well, possibly due to the curved shape of the tines. The subject who provided the data commented “it sounds as sick as I felt at the time”. In further iterations acoustic theory will be used to explore how to improve the singing effect.



Figure 3. Chemotherapy Singing Bowl

This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License.

The full terms of the License are available at
<http://creativecommons.org/licenses/by-nc/4.0/>

4. CONCLUSION

These three experiments lay a foundation for the development of a theory of Acoustic Sonification [3]. These practice-led experiments also demonstrate design research as a method for knowledge discovery and communication in the ICAD community [4].

5. REFERENCES

1. Barrass, S. (2012) *Digital Fabrication of Acoustic Sonifications*, Journal of the Audio Engineering Society, vol. 60, no. 9, pp. 709-715, September 2012.
2. Barrass, S. (2014) *Acoustic Sonification of Blood Pressure in the Form of a Singing Bowl*, in Proceedings of the Conference on Sonification in Health and Environmental Data, 12 September 2014, York University, UK.
3. Barrass, S. (2016) *Diagnosing Blood Pressure with Acoustic Sonification Singing Bowls*, International Journal of Human Computer Studies, Volume 85, January 2016, Pages 68–71, Science Direct, <http://dx.doi.org/10.1016/j.ijhcs.2015.08.007>
4. Barrass, S. (2016) *The Hypertension Singing Bowl : Establishing a Design Space for Acoustic Sonification*, in Johnson, A. (ed) Practice Led Research in New Interfaces for Musical Expression, Leonardo, Vol 49, No. 1, January 2016, MIT Press.

CONCERT

Transposed Dekany: a microtonal workshop and performance using the Satellite Gamelan app

Greg Schiemer email: greg@schiemer.com.au phone: 61 (0)+423 120 456

Abstract: My proposal is a workshop-rehearsal that leads to a performance by a consort of eighty iPhones. The workshop will focus on the features of a scale used in a microtonal composition called *Transposed Dekany*. Workshop participants will perform this using the Satellite Gamelan iPhone app. The app embodies both the musical score of *Transposed Dekany* and the software instruments used to perform it. The workshop will include a rehearsal that leads to a concert performance. Though the Satellite Gamelan app is designed to be easy to play and quick to learn, audience participation during a concert performance is beyond the scope of its design; anyone who would like to take part in a performance of *Transposed Dekany* is encouraged to join the workshop rehearsal where every player will be briefed on the musical expectations of the project.

Objectives: Participants will explore a microtonal space created using the Satellite Gamelan app. The app is based on a dekany, a 10-note scale devised by contemporary theorist and instrument-builder Erv Wilson. The app will be explained in terms of how this scale is derived from pure harmonics, what are its salient harmonic and melodic properties and what textural and acoustic by products players can expect when this scale is played simultaneously in different transpositions on different instruments.

Setup: To start the performance, every player make three selections using the app:

which family: the consort is divided into five families of instruments; each family enables a different scale transposition; players chose a family by selecting 1-of-5 coloured buttons (*Figure 1*).

which instruments: each family has sixteen members; each member choses a uniquely-assigned set of pitched instruments; these are selected using 1-of-16 buttons (*Figures 2a, 2b, 2c, 2d and 2e*).

start together: once these settings have been selected every player taps the centre button together (*Figures 3a, 3b, 3c, 3d and 3e*); this synchronises the clock that drives the app on every phone.

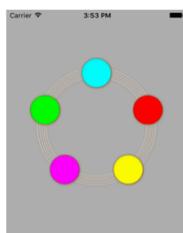


Figure 1



Figure 2a



Figure 2b



Figure 2c



Figure 2d

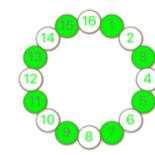


Figure 2e

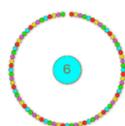


Figure 3a



Figure 3b

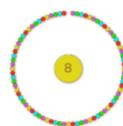


Figure 3c

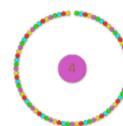


Figure 3d

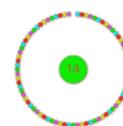


Figure 3e

Performance: Thereafter the app displays performance cues to which players respond in their own time:

when to play: throughout the performance the clock drives a sequence of 31 states that enable and disable each family of instruments; the sequence covers every combination of five families playing individually or in various combinations with other families; each state lasts 24 seconds; as the sequence advances, a clock updates the current state (*Figure 4*).

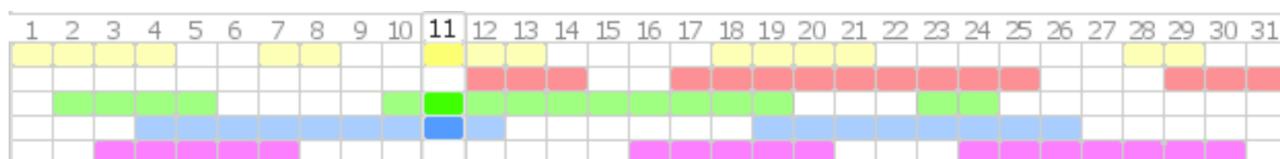
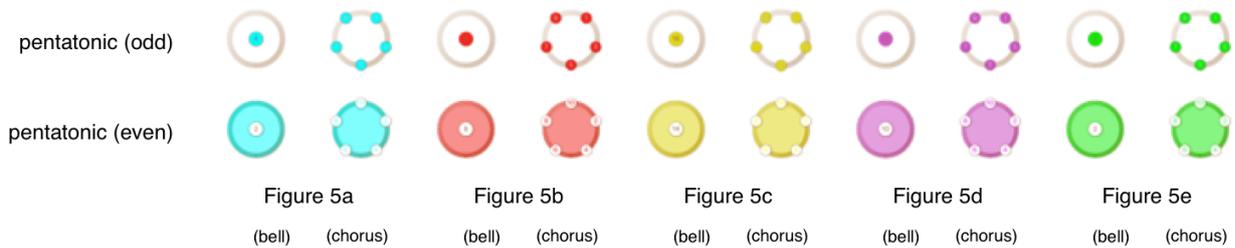


Figure 4

what to play: enabled states cue which instrument to play; bell tones are played by gently shaking the phone like a conventional handbell; chorus tones are played by tapping 1-of-5 points on the perimeter circle; as this particular flavour of dekany contains two 5-note scales that are recognisably pentatonic, separate cues have been provided for each pentatonic scale (*Figures 5a, 5b, 5c, 5d and 5e*).



Concept video: Workshop participants are encouraged to watch a concept video submitted for the Space Time Concerto competition in 2012 when the app was first used in a performance involving linked concert venues.

<https://www.youtube.com/watch?v=gfaZly6dhQA>

Satellite Gamelan app (currently iPhone only): Intending workshop participants are asked to download the Satellite Gamelan Version 1 prior to the workshop. It can be downloaded from iTunes free-of-charge.

<https://itunes.apple.com/app/satellite-gamelan-1/id578880973?mt=8>

A new version will be available closer to the start of the Conference and will remain free-of-charge until the Conference closes. Note: the new version will support a consort subdivided into five families as proposed in the concept video, unlike the app used in the 2012 performance which subdivided the consort into four families.

ICAD relevance: This proposal is relevant to ICAD in so far as it deals with an approach to mapping microtonal data. Much of this data - or tuning theory - originated in the minds of mathematicians and cartographers and has yet to find its place in the auditory world. The Satellite Gamelan app, though restricted to a single microtonal scale, offers a simple way to navigate musically uncharted terrain without the need to develop a highly nuanced performance practice associated with playing conventional instruments. In the process workshop participants will hopefully acquire a taste for the tuning variety that characterises much of the world's music.

Duration

- **workshop: 3 hours;** the workshop will be in two parts: part one will focus on using the app and understanding its harmonic features, while part two will focus on rehearsing for the concert.
- **Transposed Dekany: 12':24'';** expected stage setup time is approximately 10-20 seconds.

Technical and space requirements:

- **workshop:** I will need an assistant to coordinate assigning family and instrument numbers to each player and if necessary to assist any player that still needs help downloading the app.
- **Transposed Dekany:** the ideal venue will have high ceilings and reflective acoustics sympathetic to sound levels produced by a large consort of un-amplified instruments; no mains-powered concert amplification is required; every instrument can be set up by its player without technical support; however, players may need AC power outlets to charge phones prior to the concert.
- players enter the venue in single file holding an iPhone and taking up positions surrounding the audience (*Figure 5*). Once players are in place a lead player starts a silent 'new-years-eve' countdown from 'five'; the performance starts as all players tap the centre button together on 'zero' (*Figures 3a, 3b, 3c, 3d and 3e*).

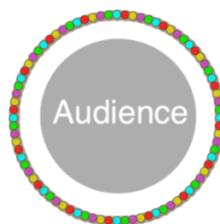


Figure 5

Enquiries: To register an expression of interest or make enquiries please visit www.gregschiemer.net/ICAD.

BioLogging Retrofit

The 'Under the Icecap' art and science collaboration aims to illuminate the fundamental connection between human activities and planetary dynamics, by creating an experimental installation and performance series that will visualise and sonify scientific and statistical datasets. In essence Under the IceCap renders complex environmental bio-logging data-sets collected by Southern Elephant Seals on their under-ice dives and open ocean transits with economic and climatic data, combining them to form 4D cartographic animations, sonifications and live performative and sculptural forms.

The byeline for the Institute of Marine and Antarctic Studies is Turning Nature into Knowledge. The Under the IceCap project supplies a second line Turning Knowledge into Culture encapsulating a powerful Art and Science synthesis and simultaneously raising the expectation but also the risk of the endeavour.



In performance at the Australasian Computer Music Conference 2015, UTS, Sydney.

The complex bio- logging data collected by Southern Elephant Seals on their dives in the Antarctic (and collated as Surface Wind Speed, Depth with Salinity, Depth with Temperature and Ocean Bottom with Bottom Density) are transcribed onto the punch paper music-box system. This a crude but effective Digital to Analogue sonification of data values. This simple prototype illustrates the potential to render tens of simultaneous data streams onto a pianola or disc-klavier for 'live' performance.

PROJECTS

- [BioLogging Retrofit](#)
- [BioPods_V2 The Nebuchadnezzar Suite](#)
- [NomansLanding](#)
- [BioPod](#)
- [Eine Kleine GeneMusik](#)
- [Songs from the UnderWorld_v3](#)
- [Architecture for Bees; Bees for Architecture](#)
- [Bristle](#)
- [Breathless; Take a Deep Breath](#)
- [Public Art concepts](#)
- [Ars Memoriae Mexicana](#)
- [Float like a Butterfly; Sting like a Bee](#)
- [Supereste ut Pugnatis \(Pugnatis\) ut Supereste](#)
- [Songs from the UnderWorld_V2](#)
- [WYSSA - All my love darling](#)
- [Songs from the UnderWorld](#)
- [BeeWork](#)
- [Milk and Honey](#)
- [Weeping Willow](#)
- [VoxAura](#)
- [CrayVox](#)
- [Bio_Logging and Under the IceCap](#)
- [Radiolarians](#)
- [New Host_V6](#)
- [Law of the Tongue](#)
- [VoxAEther](#)
- [Padme](#)
- [Adrift](#)
- [EcoLocated](#)
- [Zephyr II](#)
- [BioSonicS](#)
- [GhosTrain](#)
- [Run Silent Run Deep](#)
- [Sculpture in the Vineyards](#)
- [WA Marine Facility ~ Ondes](#)
- [AudioMaze](#)
- [Quint de Loup II](#)
- [The Wireless House](#)
- [Talking Stick](#)
- [Syren for Port Jackson](#)
- [Factory Spirit.](#)
- [Swarm](#)
- [What Survives](#)
- [Spinner](#)
- [Lotus](#)



[BioLogging Retrofit Performance](#) from [Nigel Helyer](#) on [Vimeo](#).

The primary aim is to produce creative work which is compelling and affective but is at the same time a work of scientific utility tapping into both sides of the brain! The key focus is the relationship of the environmental knowledge generated from Antarctic bio-logging data with the Anthropogenic changes in the biosphere. For ACMC bio-logging data has been transcribed using a punch tape system on a series of multi-note range music boxes as a short live performance.

You can find my paper *A Different Engine* in the [full proceedings](#) of the ACMC2015 conference.

TAG :

Page 1 of 1 pages

- [KellerRadioActive at IASKA](#)
- [Magnus Opus](#)
- [Preaching to the Converter](#)
- [Virtual Spirit](#)
- [Models](#)
- [AudioNomad + Syren](#)
- [Theorem](#)
- [Proposals](#)
- [LifeBoat](#)
- [An Unrequited Place](#)
- [Voyages from Eden to Utopia; Hercules](#)
- [The Transit of Venus](#)
- [Toll](#)
- [A Symphony for Other Cultures](#)
- [The Lament](#)
- [Sonic Landscapes](#)
- [Siren Song](#)
- [Seed](#)
- [Ship to Shore](#)
- [Oracle](#)
- [Re-Entry Vehicle; Natural Science in the Spirit World](#)
- [Naughty Apartment](#)
- [Mute in TalkTown; the sweet warm breath of science](#)
- [Meta-Diva](#)
- [Metamorphoses](#)
- [Leaven](#)
- [La Zona del Silencio](#)
- [Host](#)
- [Gyro-Diva](#)
- [Haiku](#)
- [Everything's Nice with American Rice](#)
- [Silent Forest](#)
- [Drift](#)
- [Dual-Nature \(Ebb and Flow\).](#)
- [Din](#)
- [Die Melodie der Welt; Bringing Home the Bacon](#)
- [Aura](#)
- [Vist the Drawings Gallery](#)
- [Radio Works: Order CD](#)
- [Sample Content Of Radio Works as MP3](#)
- [Big Bell Beta](#)
- [Din: Ding, Dang, Dong](#)
- [Chant](#)
- [Ariel](#)
- [GeneMusik - Sounds for Lower Life Forms](#)

TAG

- [sound sculpture](#)
- [public art](#)
- [environmental project](#)

Page 226

- [Installation](#)
- [social history](#)
- [arts and science](#)
- [interactive new media](#)
- [radiophonic](#)
- [Order CD's](#)
- [Models](#)
- [Sculpture](#)
- [Proposals](#)
- [Drawings](#)

| [62 Macgibbon Parade, Old Erowal Bay, NSW 2540 – AUSTRALIA](#)

| [+61 \(0\)4 19 49 34 95](#)

| [Contact](#) |

MALLO MARCH: A LIVE SONIFIED PERFORMANCE WITH USER INTERACTION*KatieAnna E. Wolf*

Department of Computer Science
 Princeton University
 Princeton, NJ, USA
 kewolf@princeton.edu

Reid Oda

Department of Computer Science
 Princeton University
 Princeton, NJ, USA
 roda@princeton.edu

ABSTRACT

In this extended abstract we present a new performance piece titled *MalLo March* that uses MalLo, a predictive percussion instrument, to allow for real-time sonification of live performers. The piece consists of two movements where in the first movement audience members will use a web application and headphones to listen to a sonification of MalLo instruments as they are played live on stage. During the second movement each audience member will use an interface in the web app to design their own sonification of the instruments to create a personalized version of the performance. We present an overview of the hardware and interaction design, highlighting various listening modes that provide audience members with different levels of control in designing the sonification of the live performers.

1. INTRODUCTION

In a classical concert performance, the audience plays a passive role by sitting and listening to live performers as they play acoustic instruments. With the advance of technology, new areas of musical performance have been developed that utilize the additional computational power to create live digital musical instruments. These instruments can be played by live performers, using various digital sensors (as in the EyeHarp, an eye-controlled instrument [1]), or by using live data streams as the “player” such that changes in the data trigger changes in the sound (as in *Leech*, a sonification of BitTorrent traffic [2]). In both contexts, the information provided by the performer or the data is sonified to produce a live real-time performance. When the information is provided by a live performer (rather than by data), the timing of the sonification is important as audience members expect the visual cues of the performer to synchronize with the sonified audio.

This synchronization is challenging because it takes time for information to be transmitted and for audio to be synthesized. This is particularly true in cases where the information needs to be transmitted long distances (i.e. around the world via the Internet), or when the audio synthesis is complex. Researchers have overcome these issues by making local audio processing as fast as possible and minimizing network transmission times [3]. However, there is a limit to the time that can be gained through these methods. A digital instrument called MalLo [4] was developed to

overcome the challenges of performing over long distances by predicting the strike of a percussion instrument before it occurs. The prediction is sent over the network so that the strike will be sonified at both the sending and receiving locations simultaneously. Rather than using MalLo to overcome long distances, in this work we propose to use MalLo as a way to overcome long processing times associated with complex audio synthesis between performers and a local audience.

Our piece *MalLo March* features multiple MalLo instruments played live on stage. During the performance, audience members use their own network-capable computing devices (laptops, smart phones, tablets, etc.) and headphones to design and listen to the live sonified performance. We intend to distribute in-ear headphones and headphone jack splitters so that everyone has the chance to participate, even if they do not have a device. The performance opens with *Movement 1* where audience members are asked to navigate to a URL using their web browser. Once the performers begin playing the MalLo instruments, audience members will be able to hear synthesized audio of the predicted strikes via their headphones and the web app hosted at the URL. During the first movement audience members will not have control over the design of the sonification. However, once *Movement 2* begins they will be able to interact with a user interface on the web app to change the sonification in various ways and manipulate the sounds of the MalLo instruments. By opening up the sonification design process to audience members they be able to explore the sonic possibilities of the performance.

In this extended abstract, we briefly describe the technical design of the performance by outlining the way the hardware of MalLo works and the types of interactions that will be available to audience members via the user interface. We also include a brief discussion on what we hope to gain from the performance.

2. HARDWARE DESIGN

In order to gain time with which to execute computational processes we use a predictive instrument called MalLo [4]. As shown in Figure 1, MalLo predicts note times and note velocities before they are played by capitalizing on the long distance traveled by percussion mallets. First, we track the head of a percussion mallet. Next, we fit a quadratic regression to the mallet’s path (taking advantage of the fact that its path is very predictable [5]). We compute the time at which the mallet will strike the surface, and we send this information to the receiver. The receiver takes in a continuous stream of predictions, each one more accurate than the previous. The receiver waits until the predicted time reaches a specified accuracy threshold and sonifies the note. During the period



This work is licensed under Creative Commons Attribution Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

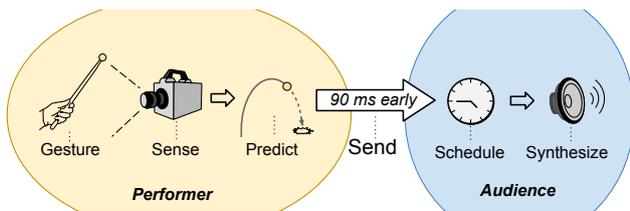


Figure 1: Our musical instrument tracks the head of a percussion mallet and predicts what time it will strike. This allows up to 90 ms of anticipation time which can compensate for time spent transmitting note information or processing of advanced synthesis algorithms.

between receiving the predictions and the final sonification, the receiver can be performing a variety of computations (e.g. complex audio synthesis algorithms).

Our system sends its timing predictions as absolute time messages, in Unix time (time since January 1, 1970). This allows for nearly unlimited scaling. The system clocks of the sender and receiver must be synchronized to a high degree. In previous work MalLo relied on GPS satellite signals to allow for time synchronization over long distances. However, for this performance we will use a variant of Precision Time Protocol (PTP) for time synchronization [6]. Also, because we will not have access to the system clock of every audience machine, we will implement it in our web app, on the client side.

Each audience listener will be required to listen via a web app that they access on a personal computing device. As shown in Figure 2, MalLo prediction messages are sent to each device where they are scheduled, and the audio is synthesized.

3. INTERACTION DESIGN

Audience members will navigate to the web app URL at the start of the piece. Using their own personal networked device and headphones, they will listen to the sonification of the predicted MalLo notes that are being sent over the Internet. During *Movement 2*, the audience members will see a user interface designed to give them control over the sonification of the performance. We imagine three types of interaction modes that allow for various levels of control of the sonification. In each of these modes, the incoming data to the sonification will be the live performance data sent from MalLo (note timing information and velocity), and the output will be the sound from the sonification algorithm chosen via the particular interaction mode.

Interaction Mode 1: Change the Channel In this mode, users are allowed to switch between preset sonifications similar to changing the channel on the radio and hearing the same song performed by different combinations of instruments. For each channel, the live data from the performers will be input into a different sonification to create a unique sonic performance.

Interaction Mode 2: Change the Instruments To give users slightly more control, in this mode users are able to choose a different preset sonification for each of the MalLo instruments. By doing so users are able to “Change the Channel” for each instrument.

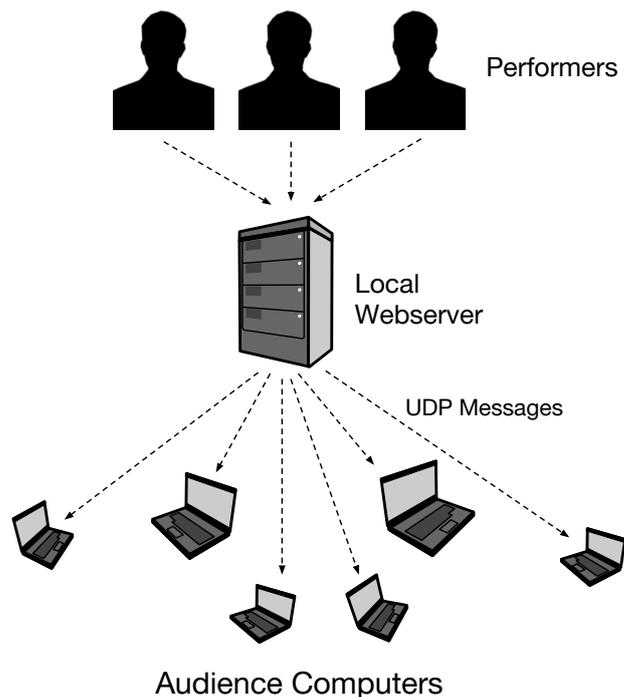


Figure 2: Note information is sent from a local webserver to each audience member’s web browser where. There the notes are scheduled and synthesized according to the parameters of the sonification.

Interaction Mode 3 - Create the Instruments At the lowest level of control, users are able to directly modify the synthesis parameters for each instrument. In particular they can alter pitch, duration, waveshape, samples etc.

4. DISCUSSION

MalLo March is a piece intended to combine live sonification of performers with end-user design of sonifications. With permission from each audience member, we hope to log information about how they used the interface during the performance and do a short survey at the end. In particular we are interested in observing how people interact with the various interaction modes, how successful they felt they were in creating sonifications, and what was gained or lost by adding the interaction in the second movement. Our composition is intended to be enjoyable for the audience as well as give us insight into the design of sonifications.

5. ACKNOWLEDGMENT

This material is based upon work partially supported by the NSF GRFP under Grant No. DGE 1148900. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The work is also partially supported by the Princeton University Project X Grant.

6. REFERENCES

- [1] Z. Vamvakousis and R. Ramirez, “Temporal control in the EyeHarp gaze-controlled musical interface,” in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2012.
- [2] C. McKinney and A. Renaud, “Leech: Bittorrent and music piracy sonification,” in *Proceedings of the Sound and Music Computing (SMC)*, 2011.
- [3] J.-P. Cáceres and C. Chafe, “Jacktrip: Under the hood of an engine for network audio,” *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, 2010.
- [4] Z. Jin, R. Oda, A. Finkelstein, and R. Fiebrink, “Mallo: A distributed, synchronized instrument for internet music performance,” in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2015.
- [5] S. Dahl, “Striking movements: A survey of motion analysis of percussionists,” *Acoustical Science and Technology*, vol. 32, 2011.
- [6] K. Correll, N. Barendt, and M. Branicky, “Design considerations for software only implementations of the IEEE 1588 precision time protocol,” in *Conference on IEEE*, 2005.

HEARING A GENE OF HEARING

Stephen Barrass

University of Canberra, Australia
stephen.barrass@canberra.edu.au



Hearing a Gene of Hearing is composed from the GJB2 gene that is important for human hearing. The gene provides instructions for making a protein that transports potassium ions in cells, which is important in transducing sound waves on the ear drum into electrical impulses in the cochlear. For this reason, mutations of the GJB2 gene can cause what is called “non-syndromic” hearing loss because it does not have any outward symptoms.

The composition of this sonification consists of a mapping from genetic components of the GJB2 gene to percussion notation for a Ubang clay drum. A subsequence of the GJB2 gene from 4 different species (Human, Orangutan, Dog, Mouse) was mapped to 4 instrument scores to be played in parallel on 4 Ubangs. The parallel performance of 4 gene sequences by 4 musicians explores the concept of collaborative sonification. The piece also explores data sonification as a medium for musical performance and aesthetic experience.



Hearing a Gene of Hearing premiered in the Concert Programme curated by Charles Martin for the International Conference on Auditory Display (ICAD), and was performed by in Canberra in July 2016. The piece was performed by the ANU Experimental Music Studio directed by Charles Martin and performed by Benjamin Drury (Human), Millie Watson (Orangutan), Ellen Falconer (Dog) and Ben Harb (Mouse), on 6 July 2006 at Lewellyn Hall in the ANU School of Music, Canberra.

Video - <https://stephenbarrass.com/2016/07/09/hearing-a-gene-of-hearing/>

The patterns that emerge are unusual because they do not have any regular or repetitive rhythm. This sonification of genetic patterns may help the listener understand more about the structure of DNA.

PHASERINGS FOR IPAD ENSEMBLE AND ENSEMBLE DIRECTOR AGENT

Charles P. Martin

Research School of Computer Science
The Australian National University
Canberra, ACT, Australia
charles.martin@anu.edu.au

ABSTRACT

PhaseRings for iPad Ensemble and Ensemble Director Agent is an improvised musical work exploring the use of dynamic touch-screen instruments, tracked by a gesture-classifying agent, to enhance the creativity of an ensemble of performers. The PhaseRings app has been designed specifically for ensembles to create expressive music with simple percussive gestures. Each performer can play with a small set of notes selected from a musical scale that is common to the whole group. The agent tracks the performers' gestures and reacts to moments of heightened gestural change by allowing the performers to access a new set of notes from a new scale. In this way, this performance links collaborative creativity to the dynamic interface of the individual touch-screen instruments.

1. DESCRIPTION OF THE WORK

In this performance, an ensemble of improvising iPad performers using the PhaseRings iPad app are tracked and directed by an Ensemble Director Agent running on a server. While artificial intelligence agents are often used as improvisation partners in computer music, they have only rarely been used to conduct or direct a performance. The agent software in this performance, Metatone Classifier, tracks performers by classifying their touches according to a vocabulary of percussive gestures. It encourages and rewards the group for creative interactions by updating their app interfaces with new notes and sounds when it detects increases in gestural variety and change.

Performance with a network of computer music interfaces has been explored as early as the late 1970s when The League of Automatic Music Composers connected their early personal computers as part of their compositions [1]. Weinberg has described such interconnections, where data is shared directly between interfaces, as a Local Performance Network [7]. An alternative model for networked computer was postulated by Pressing [6] who suggested that an intelligent agent could serve as a musical director, monitoring information from the performers, applying tests, and then issuing commands or interrupting processes in response.

In the present work, the concept of an Ensemble Director Agent has been implemented in the Metatone Classifier software which tracks improvised ensemble interactions during the performance. As the PhaseRings instrument runs on touch-screens, the



Figure 1: Screenshot of the PhaseRings app showing rings that trigger eight different notes and the GUI button to change setup (exposed by the Ensemble Director Agent).

morphology of the instrument can be updated during the performance in response to signals from the agent. So, this performance connects improvised collaborative creativity to the capabilities of the musical instrument. In the following sections the technical details of the PhaseRings app and Metatone Classifier agent will be described.

2. THE PHASERINGS APP

The PhaseRings app [3] is an annular interface for percussive ensemble performance. Each player is given a number of notes from a particular scale which are shown as rings on the screen (see Figure 1). These rings can be tapped for short sounds, or swirled to create different kinds of long sounds. These notes change throughout the performance in response to signals from the agent. Each performer's notes are taken from the same scale so that, while each player can explore a unique melodic space, their harmonic position is common to the group.

PhaseRings' interaction with the gesture tracking agent has been the subject of a series of HCI studies where multiple iterations of the interface were analysed [5]. The final interface follows a mixed-initiative model [2]; signals from the agent expose a button in the app GUI and the interface changes only if a member of the ensemble chooses to tap the button. This model allows the performers to retain ultimate control over interface changes while incorporating "suggestions" from the Ensemble Director Agent.



This work is licensed under Creative Commons Attribution Non-Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>



Figure 2: The ANU Experimental Music Studio performing with PhaseRings and Metatone Classifier at You Are Here 2015.

3. METATONE CLASSIFIER

Metatone Classifier is the Ensemble Director Agent used in this performance. This software tracks multiple touch-screen performers simultaneously and calculates measures of the overall ensemble behaviour. Each second, the performers' touches are classified into one of nine continuous gestures, then, the recent history of calculated gestures are compiled into transition matrices, similar to calculating a first-order Markov model. These transition matrices can be calculated for individual performers, or averaged over the whole ensemble, giving a high level representation of the group's behaviour.

In this performance, a metric called *flux* is applied to these ensemble transition matrices which measures how much performers change between different gestures. If the agent detects a sharp increase in flux, possibly related the start of a new musical section in the improvisation, it sends a signal to PhaseRings suggesting that the interface could be updated. Further details of the implementation of Metatone Classifier are available elsewhere [4].

4. REALISATION AT ICAD 2016

At ICAD 2016, this work will be realised by the ANU Experimental Music Studio, a flexible group of musicians with members studying musical performance, composition and other disciplines at the Australian National University. This group previously performed with PhaseRings and Metatone Classifier at the You Are Here 2015 festival in Canberra (see Figure 2) and members of the group have also participated in a series of studies and workshops to analyse and improve the design of these systems throughout 2014 and 2015.

5. REFERENCES

- [1] J. Bischoff, R. Gold, and J. Horton. Music for an interactive network of microcomputers. *Computer Music Journal*, 2(3):24–29, 1978. doi:10.2307/3679453.
- [2] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in*

Computing Systems, CHI '99, pages 159–166, New York, NY, USA, 1999. ACM. doi:10.1145/302979.303030.

- [3] C. Martin. PhaseRings v1.2.0, May 2016. doi:10.5281/zenodo.50860.
- [4] C. Martin, H. Gardner, and B. Swift. Tracking ensemble performance on touch-screens with gesture classification and transition matrices. In E. Berdahl and J. Allison, editors, *Proceedings of the International Conference on New Interfaces for Musical Expression*, NIME '15, pages 359–364, Baton Rouge, LA, USA, 2015. Louisiana State University. URL: http://www.nime.org/proceedings/2015/nime2015_242.pdf.
- [5] C. Martin, H. Gardner, B. Swift, and M. Martin. Intelligent agents and networked buttons improve free-improvised ensemble music-making on touch-screens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '16, New York, NY, USA, 2016. ACM. doi:10.1145/2858036.2858269.
- [6] J. Pressing. Cybernetic issues in interactive performance systems. *Computer Music Journal*, 14(1):12–25, 1990. doi:10.2307/3680113.
- [7] G. Weinberg. Local performance networks: musical interdependency through gestures and controllers. *Organised Sound*, 10(3):255–265, 2005. doi:10.1017/S1355771805000993.

ATOM TONE – live electronic concert using sonification of atomic data*Jiří Suchánek*

Janáček Academy of Music and Performing arts,
Ph.D. candidate of Composition and Theory of
Composition,
HF JAMU, Komenského nám. 6, 662 15 Brno,
Czech republic
jiri.suchanek@hf.jamu.cz

ABSTRACT

Project *Atom Tone* explores aesthetic possibilities of sonification of atomic data and using generated complex waveforms in live electronic music. The result is 30min concert. Sonification is done in max/msp patch that I've programmed for this project during Visegrad residency at A4 Bratislava 2015. Now I am continuing in development of the patch as a part of specific research during my Ph.D. studies at JAMU.

The sonification has two parts: synthesis and modulation. Synthesis uses atomic spectroscopic data as a source for additive synthesis technique – each oscillator is tuned to recalculated exact frequency of the atomic emission spectral line. Each element has unique list of spectral line frequencies (Figure 1). This atomic “fingerprint” is sonificated into the complex chord that is further modulated. Modulation is done only with numbers from Mendeleev periodic table related to selected element. Numbers can be routed to several parameters. This routing method is open to many possibilities.

The goal of the project is to discover possible new aesthetic qualities for the contemporary electronic music with this specific sonification technique.

1. INTRODUCTION

In this paper I will describe the basic concept of my max/msp sonification patch that I use for the live electronic concert. Also I will describe all optimal technical requirements for the concert realization. I will not talk too much about the atomic spectroscopy itself because it is well and deeply documented in lot of scientific texts, books or web pages [1]. I must mention that I am not physicist but musician/sound/media artist so still I have to learn lot of about the atoms. This project is interdisciplinary and I consult it with Institute of Theoretical Physics and Astrophysics MU [2] for relevant results.

2. SYNTHESIS – SONIFICATION OF SPECTROSCOPIC ATOMIC DATA

I synthesize the waveform of the one element. Sonification of molecules, or even reactions is something I am still working on and it looks like long term research. I use NIST spectroscopic database [3] as a source of light emission / absorption frequencies.

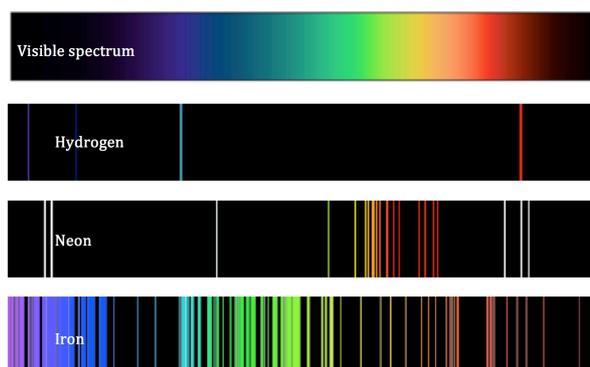


Figure 1: The visible light spectrum is displayed at the top and example of atomic spectra lines for three elements (hydrogen, neon, iron) are below.

Image © Neon spectrum: Deo Favente

I've formatted this database into the .txt files compatible to coll object in max/msp. Each element has from ten to thousands of exactly described lines. The parameters of each line that I selected for sonification is wavelength (nm) and relative energy [4].

The core of the sonification synthesis can be described simply: each line means one oscillator (that can be set to sine, saw, tri, tan, etc.). Frequency of the oscillator is counted from the line frequency simply by division (from Thz light range into the Hz audible range). The division number can be adjusted for the “best” musical result. The volume of each oscillator is logarithmic value of the relative energy of the line. I decided to use all measurable lines in vacuum. So I use the lines from visible light range and also in ultraviolet (Lyman series) and infrared (Paschen series) range (Figure 2).

I decided to select only the lines with higher relative energy to eliminate huge number of oscillators (I had quite problem with efficiency of poly~ in max/msp with more than hundreds of voices...). I plan to rewrite the code with Supercollider or try gen~ for better efficiency. But even with maximum of 100 oscillators the results are already clear and highly usable for the musical performance.

There are four energy states of each element in database. I use all of them, so every element can be heard in 4 different states – the differences are clearly audible.

With this method I can sonificate all elements described in the NIST database. That means elements with proton number from 1 (Hydrogenium) till 99 (Einsteinium). Elements with higher proton numbers are not included in the database – probably they exist so short time for needed measurements.

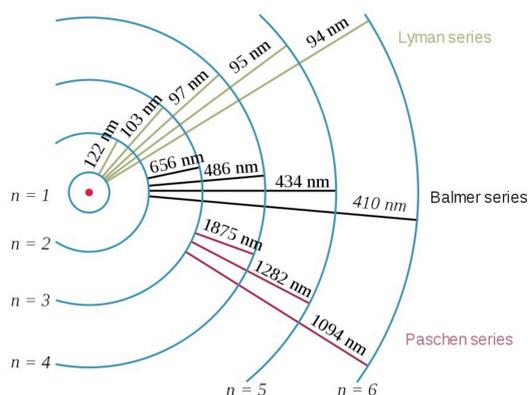


Figure 2: The relation between the lines wavelength, atom orbitals and energy state of the atom [5]

3. “MENDELEEV” MODULATION

Generated pure sound described above can be modified with several techniques. I use waveshaping, multiple frequency shifting, buffering, and then modifying with 2d.wave~, fir filtering. This processing is conceptually more open than the exactly described synthesis. The goal of this processing is to find the musically most suitable result and to have many options to play with.

The data used for the modulation/processing parameters are taken from Mendeleev periodic table. As a source data I use: atom number, atom weight, electronegativity, density, ionization energy, atomic radius, constant radius and period. The sources and parameter destinations are connected in matrix so it is very easy to change the routing and find the most interesting settings. I know this does not suite to the one of the sonification definition written by Thomas Hermann - especially in the context of the concert where I sometimes change the routing. *“The transformation is systematic. This means that there are precise definition provided of how the data (and optional interactions) cause the sound change.”* [6]

This is sort of collision between exact scientific sonification technique where listener has clear cue for analysis and artistic approach. For me the sonic result is more important than strictly respected method during the whole concert. I know my method very well but when music needs it I violate its rules. So for me the aesthetic decision stands above the purity of sonification method. I understand described sonification tool as great source for discovering new aesthetic territories for sound arts/music with kind of rich metaphoric and symbolic meanings.

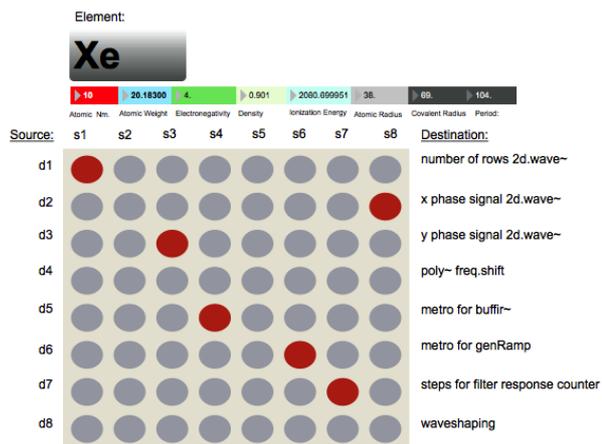


Figure 3: The view of routing matrix in the Atom Tone max/msp patch

4. ELEMENT MIXING AND PARAMETER CONTROLLING

Actual version of the Atom Tone max/msp patch offers synthesis of two elements at once. So I can fade from one element into another and that is basically the compositional concept of the performance: slowly evolving changes and fading from one element into another. I control the patch with mapped midi controller. I’ve sorted the sounds into the categories like bass, complex, high frequency, metallic, disturbing and I mix them usually in a contrast way. Now I am working on a patch where molecules and even chemical processes can be sonificated and synthesized. I can imagine concert as a chemical reaction.

5. CONCERT FORM

Concert takes approx. 30 minutes. It consists from prepared parts (beginning and the end) and improvised parts where I react on the mood of the audience and myself. One part of the concert is also video projection where actual sonificated atomic lines are projected. I play with laptop and midi controller. Sometimes I played with friend who processed my signal in analog modular synthesizer (Figure 4). I understand the concert as an exploration specific kind of aesthetic deeply hidden in a matter. That is what interests me the most – hidden music inside the matter. I feel something fundamental in this musical approach what keeps me continuing this research. I am trying to bring the artistic experience of the atoms – kind of impersonal reality that is here regardless on us but of course also inside us.



Figure 4: concert of *Atom Tone* at NEXT festival at Bratislava, 2015 [7]

6. TECH-RIDER

Audio:

OUT (from the soundcard) : 2x JACK TRS 6,3mm
stereo monitor (L+R) on the stage

Projection:

DVI cable to the projector - end placed on the table on stage
projection screen on the whole wall behind the stage - from the bottom of the stage.

Space:

high table (100cm from the ground) - size of the table
approx.: 1.2x0.7m
+ 230V 5x plugs
+ blue soft light on the table from the top
+ darkness + silence

time for soundcheck: 30min (including setup)

This is optimal technical setup. Of course the concert can be done also in somehow limited condition.

7. MY PERSONAL BACKGROUND AND MOTIVATION

I study Ph.D. (Composition and Theory of Composition) at Janáček Academy of Music and Performing Arts, Brno (JAMU). My Ph.D. thesis is "Sonification and aesthetics of data-mapping". I really would like to connect with many interesting people related to ICAD. I am sure the conference will have big impact on my thesis and the whole ICAD will be hugely inspiring for me. I hope that my concert could be also interesting for someone. I got the school funding for specific research related to described atomic data sonification used in music. Part of this support is also related to participation at ICAD2016 what is my dream plan. So I have covered the possible travel costs. I know it is not sure if I will be accepted by ICAD by I had to make this possible support in advance... Also I work as an assistant at Cabinet of Audiovisual Technology at Faculty of Fine Arts Brno University of Technology where I teach Audio Technology and History of Sound-Art. I really wish to come to ICAD2016 present my actual sonification music and meet interesting people.

8. LINKS WITH ONLINE MEDIA CONTENT

Here are links with the content related to this project:

Description of the project on my personal web page – including sound examples:

<http://www.jiri-suchanek.net/en/project/atom-tone/>

video documentary from premiere of the *Atom Tone* at NEXT festival:

<https://vimeo.com/149558954>

my personal web page with my other projects:

<http://www.jiri-suchanek.net/en/>

9. REFERENCES

- [1] <http://www.thespectroscopynet.eu/>
- [2] <http://astro.physics.muni.cz/en/>
- [3] http://physics.nist.gov/PhysRefData/ASD/lines_form.html
- [4] <http://physics.nist.gov/PhysRefData/ASD/Html/lineshelp.html#OUTRELINT>
- [5] http://www.thespectroscopynet.eu/?Physical_Background:Atomic_Emission:Line_Spectra
- [6] T. Hermann, "Taxonomy and definitions for sonification and auditory display", in *Proc. of the 14th Int. Conf. on Auditory Display*, Paris, France June 24 - 27, 2008, pp. ICAD08-2
- [7] <http://2015.nextfestival.sk/atom-tone/?lang=en>

WORKSHOPS

ICAD2016 WORKSHOPS

A Layer-Based Amplitude Panning (LBAP) Audio Spatialization Algorithm

Using Layer-Based Amplitude Panning (LBAP) Audio Spatialization Algorithm and D4 Max Library's Rapid Prototyping Tools for the Interactive 3D Audification and Sonification of Multidimensional Data, and in Interactive Music Scenarios in High Density Loudspeaker Arrays

<p>PRESENTER Assoc Prof Ivica Ico Bukvic, Virginia Tech email ico@vt.edu</p>	<p>PARTICIPANTS: 10-20 LOCATION: ANU School of Music REQUIREMENTS: Laptop with Max/MSP and D4 Max Library.</p>
<p>This workshop focuses on the new LBAP algorithm and the supporting D4 Max library's collection of tools for real-time 3D audification, sonification, and interactive performance in High Density Loudspeaker Array (HDLA) scenarios. It will cover theory and application of the said tools, including its unique features, including the Spatialization Mask, Motion Blur, Radius, and advanced shape painting and rendering. Participants will be given a hands-on opportunity to experiment with the software and learn how to create a scalable and transportable systems that can adapt to most speaker configurations. Of particular focus will be learning ways to import and integrate multidimensional data and render it spatially in real-time. So far, LBAP and D4 have been successfully tested with configurations of up to 137 loudspeakers and up to 1,011 concurrent 24-bit 48KHz audio streams with sub 22ms latency. The workshop may be of particular interest due to its focus on 3D spatial auditory displays using HDLA. It will be the first public workshop on a newly developed technology and is being offered in addition to a paper submitted to the conference.</p> <p>Given the focus on multichannel audio, aural components of the workshop will leverage existing audio infrastructure while also providing alternative visual tools for experiencing spatialized audio. While participation without obtaining the tools is possible and encouraged, for optimal experience participants should consider obtaining the software online at http://ico.bukvic.net/main/d4/ or sold on site—special pricing will be provided workshop participants. For additional info on the software and pricing, please contact the author. Considering the D4 library is specifically designed for the Max environment, participants who decide to obtain the D4 library will need a Max license and a laptop (Windows or Mac) with the up-to-date 64bit Java libraries installed.</p>	

Dare to Design Hearables

<p>PRESENTERS: Prof Simon Carlile, Starkey Hearing Technologies, simon_carlile@starkey.com</p> <p>A/Prof Stephen Barrass, University of Canberra, stephen.barrass@canberra.edu.au</p> <p>Jane Cockburn, Kairos Now jane@kairosnow.com.au</p>	<p>PARTICIPANTS: 10-20 LOCATION: ANU School of Music REQUIREMENTS: Imagination</p>
<p>This workshop will provide participants with the opportunity to collaborate in the exploration of Hearables, which are next generation augmented hearing aid technologies.</p> <p>The current early incarnation of hearables combines activity tracking technologies with audio playback or streaming using earbuds. These provide an opportunity to deliver new and convergent technology without the need to change consumer behaviour but potentially also enable new use-cases, technology convergence and services.</p> <p>The key questions include:(1) What can/could actually be achieved technically given the price point and regulatory and other frameworks (the platform)? (2) What needs or wants might these technologies actually satisfy (the product)? (3) Who would pay for this and how would they pay for it (the monetization)? (4) How would the business model be sustained in the face of competition, technical commoditization and market saturation (the business case)?</p> <p>Participants will learn about the state of the art in this area, and opportunities for industry collaboration and open source development of future products in this space. They will form groups and participate in an iterative design thinking process using personas, journey maps and reframing techniques. By the end of the workshop groups will develop a proposal that could be the basis for a grant or industry partnership.</p>	

Computer Knitted Data Scarf

PRESENTER Prof Angelina Russo University of Canberra angelina.russo@canberra.edu.au	PARTICIPANTS: 5-10 LOCATION: ANU School of Music REQUIREMENTS: Your own data-set
<p>Your ICAD registration includes a 'data beanie' knitted from possum fur and designed to keep your head warm in the frosty mid winter mornings in Canberra.</p> <p>This beanie is computer knitted from punch-cards that encode data logged from seals diving under the antarctic as part of Nigel Helyer's sonification concert piece <i>Biologging Retrofit</i> (http://www.sonicobjects.com/index.php/projects/more/biologging_retrofit/). During the ICAD concert, attendees will hear a sonification of the data that is knitted into their beanie!</p> <p>http://www.sonicobjects.com/index.php/projects/more/biologging_retrofit/</p> <p>The workshop will provide an overview of computer knitting processes, techniques and technologies. You will then have the opportunity to map your own dataset into a computer punchcard to control a knitting machine. For example one could map daily temperature from your home city as a pattern http://www.worldweatheronline.com/canberra-weather-averages/australian-capital-territory/au.aspx. Then choose colours to design your own personal Data Scarf to keep the mid winter chill at bay.</p> <p>Although each person is applying the same process, different design decisions will result in different scarves. You can map the data into a pattern in many different ways, and even whether to use colours to match your beanie. The mapping of the data onto the pattern and the choice of colour and contrast has a big effect on the final fashionability and desirability. We know because designing the data beanie has required many iterations to achieve something that we hope you will like to wear.</p> <p>If you like your scarf pattern you can order it, and it will be knitted during the conference.</p>	

Data Sonification using Python and Csound

PRESENTER Prof David Worrall Audio Arts and Acoustics Department Columbia College Chicago dworrall@colum.edu	PARTICIPANTS: up to 16 LOCATION: ANU School of Music Computer Lab REQUIREMENTS: Your own laptop preferred. Be prepared: Pre-workshop preparations.
<p>Python is a popular, easily learnt general-purpose programming language which can serve as a glue language to connect together many separate software components in a simple and flexible manner. Widely used in the scientific community, it can also be used as a high-level modular framework for controlling low-level operations implemented by subroutine libraries in other languages.</p> <p>Csound arguably has the widest, most mature collection of tools for sound synthesis and sound modification. There are few things related to audio-programming that you cannot do with Csound; it can be used in real-time to synthesise sound or process live audio or other control data (including MIDI and OSC) on the fly. It can be used to render sound on hand-held and other mobile devices or, when synthesis needs require it, sound can be rendered to file.</p> <p>The Python API (Application Programming Interface) to Csound is robust and available on all hardware platforms. The aim of this workshop is to provide a hand's on introduction to producing software data sonifications using a combination of these most powerful, open-ended, and extensible set of tools. If required, it will be divided into sessions on Python, on Csound individually, and then in combination.</p> <p>The workshop will begin with begin with a detailed description of a non-trivial application example.</p> <p>NB See the pre-workshop setup instructions (http://sonification.com.au/workshops/).</p>	

Neurofeedback and Contemplative Interaction

PRESENTER Dr George Poonkhin Khut UNSW Australia Art & Design	PARTICIPANTS: 5-10 LOCATION: National Portrait Gallery
<p>George Khut will provide an overview of his recent works with Alpha (brainwave) neurofeedback. Participants will have the opportunity to interact with the brainwave-controlled artwork in the gallery, take a 'behind-the-scenes' tour of the sound design and sonification strategies used in this work (using Cycling74's Max), and discuss questions raised by this work: how is this kind of interaction similar to, and different from, traditional contemplative practices such as trance, yoga and qigong meditation? How might a consideration of these issues inform the design of future health apps and services?</p>	

Hack Your ICAD Name Badge

PRESENTER Tim Barrass, Mozzi barrasstim@gmail.com	PARTICIPANTS: 20 LOCATION: ANU School of Music REQUIREMENTS: ICAD name badge, laptop, DOWNLOADS: Arduino 1.0.5, Mozzi library
<p>Update: Unfortunately the Mozzi Badge did not get finished in time for the conference. However we have some working prototypes, and some Mozzi learning kits, so the workshop can go on.</p> <p>The workshop will introduce the Mozzi programming library, and describe the process of hardware fabbing using Fritzing that has got to the stage of a MozziDuino hardware prototype.</p> <p>The MozziDuino is an Arduino Clone with onboard sound amplifier and speaker, a light sensor and an extremely sensitive electrostatic sensor.</p> <p>It has a USB port that allows you to program the Arduino to synthesise sounds using the Mozzi Synthesis library https://sensorium.github.io/Mozzi/. Have a look at the Gallery (https://sensorium.github.io/Mozzi/gallery/) to see how Mozzi has been used for art installations, museum exhibits, music performances, boutique synthesisers and custom special effects units The workshop will be led by Tim Barrass, the inventor of Mozzi, and the MozziDuino.</p> <p>Upon completion of the workshop you will</p> <ul style="list-style-type: none"> • Understand the capabilities of the MozziDuino wearable sonification synth. • Be able to program interactive sonifications of sensors with Mozzi. • Be able to add your own sensors and sounds to the MozziDuino. 	

Biologging Retrofit

PRESENTER Dr Greg Schiemer greg@schiemer.com.au	PARTICIPANTS: 16 (minimum), 80 (maximum) LOCATION: ANU School of Music REQUIREMENTS: iPhone, Satellite Gamelan App, willingness to Perform in the ICAD Concert.
<p>A workshop-rehearsal that leads to a performance by a consort of eighty iPhones. Participants will gain experience of using a distributed mobile platform for interactive collaborative exploration of sonic materials.</p> <p>The workshop will be in two parts: part one will focus on using the app and understanding its harmonic features, while part two will focus on rehearsing for the concert. Participants will explore a microtonal space created using the Satellite Gamelan app. The app will be explained in terms of how this scale is derived from pure harmonics, what are its salient harmonic and melodic properties and what textural and acoustic by products players can expect when this scale is played simultaneously in different transpositions on different instruments.</p> <p>Participants are asked to download the latest version of the Satellite Gamelan prior to the workshop from iTunes free-of-charge.</p> <p>Participants are encouraged to watch a concept video submitted for the Space Time Concerto competition in 2012 when the app was first used in a performance involving linked concert venues. View video (https://www.youtube.com/watch?v=gfaZly6dhQA).</p>	

The original Call for Workshop Proposals (<http://www.icad.org/icad2016/workshopsCall.shtml>) is available for reference.