DESIGNING SOUND: TOWARDS A SYSTEM FOR DESIGNING AUDIO INTERFACES USING TIMBRE SPACES

Craig Nicol, Stephen Brewster and Philip Gray

Glasgow Interactive Systems Group, Department of Computing Science, University of Glasgow
{can,stephen,pdg}@dcs.gla.ac.uk — http://www.dcs.gla.ac.uk/~can

ABSTRACT

The creation of audio interfaces is currently hampered by the difficulty of designing sounds for them. This paper presents a novel system for generating and manipulating non-speech sounds. The system is designed to generate Auditory Icons and Earcons through a common interface. Using a timbre space representation of the sound, it generates output via an FM synthesiser. The timbre space has been compiled in both Fourier and Constant Q Transform versions using Principal Components Analysis (PCA). The design of the system and initial evaluations of these two versions are discussed, showing that the Fourier analysis appears to produce higher quality results, contrary to initial expectations.

1. INTRODUCTION

Many authors, for example Gaver [1] and Mynatt [2], mention a lack of clear design tools for sounds or auditory interfaces. This paper presents ongoing work on a system to address this need.

Currently auditory interfaces are designed on an *adhoc* basis with no unifying tool that allows designers to work with both the primary paradigms in the field, i.e. Auditory Icons and Earcons. This slows progress in creating the interfaces and in understanding how to design sounds for humans.

The system we are developing will use a more natural interface than the current tools and enables sounds to be described in terms of the sources that produce those them rather than in terms of their wave properties. It is hoped that this system will help designers find useful sounds for their interfaces and from this a complete set of design guidelines for sonic interactions can be realized.

The interface will be designed using timbre spaces as in Hourdin *et al.* [3], based on perception experiments in humans by Grey [4]. Sounds will be analysed and loaded into this timbre space where they can be modified before being output via a synthesiser.

After a short discussion on why this is an important topic, Section 3 discusses human perception and auditory interfaces research. Section 4 gives an overview of the technology and design behind the system. We conclude with a brief summary of work completed and consider future directions.

2. OVERVIEW

As computer displays get smaller on devices such as mobile phones and personal digital assistants, audio interfaces will become even more important for providing information to users. Ambient audio can be used to enrich a user's information awareness. New methods of designing, prototyping and evaluating audio interfaces need to be developed that can be understood by designers with a background in Human-Computer Interaction and psychology.

Sound has always been an important part of interacting with the physical world. Compared with our rich sonic environment, the computer interface is a poor cousin. Gaver [5] noticed this and developed a system of Auditory Icons whose sounds were related to a real-world sound, and whose properties could be adjusted to reflect changes in the underlying environment. Many distinct audio interfaces have been designed since but despite numerous guidelines defining how each type of interface should be designed, there is no common tool for developing or evaluating these interfaces.

In the design of Audio Aura [2] for example, in response to a problem the authors uncovered in existing interfaces, the following guidelines were followed to prevent an "alarm response":

> "[Background sounds are designed to avoid] sharp attacks, high volume levels, and substantial frequency content in the same range as the human voice (200 - 2,000 Hz)."

In contrast, the computer music community has defined many methods for creating and manipulating sounds and has defined some common platforms to unify them. The most basic musical interface is the MIDI standard, used by Earcons [6], which defines a series of commands that specify musical notes and operations on these such as sustain, pan and instrument used. The biggest problem with MIDI is that the output sound is not guaranteed. Although there has been some standardisation, the basic MIDI specification does not guarantee any particular sound or behaviour will be consistent between devices.

Another popular field within computer music concerns the transformations made available by the various Analysis-Synthesis (A/S) techniques (reviewed by Masri *et al.*[7]). It allows composers to manipulate sounds rather than the sources that produce them. Sounds can be sliced, stretched, reversed and changed into other sounds. A/S achieves this by presenting the sound in an analysed format which is different to that used for recording and storage.

This project is working towards combining the research in computer music and audio interface design, allowing interface designers access to complex acoustic and musical methods for developing sound through an interface defined in terms of human perception.

3. HUMAN HEARING AND MACHINE SOUND

This section discusses current research in the fields of audio perception and sound production. A comparison of different synthesis techniques is shown in Table 1.

Technique	**Description**	**Advantages**	**Disadvantages**
Sampled Sounds	Play a recording through the sound card Most used in consumer audio interfaces	Low CPU Load Easy to record new sounds	Hard to edit sounds High disk usage
Additive Synthesis	Sum of many unique sine waves Used by Gaver to create Auditory Icons	Infinitely complex sounds Used by Hourdin *et al.*	Long computation time Low quality if real-time
Frequency Modulation (FM) Synthesis	One sine wave frequency altered by another Developed by Chowning [8]	Complex sound from few waves Many sound cards support it	Hard to match real sounds Raw sound quality is poor

Table 1: Comparison of sound synthesis techniques

3.1. Perception of Sound

There are four components to the way we hear a single sound: the pitch, the loudness, the duration and the timbre. Where sounds are combined, the relative values of these components are important as is the temporal pattern of the sounds (e.g. Earcons, Section 3.2.2).

The timbre of the sound is what differentiates two sounds whose pitch, loudness and duration are equal. Unlike these other measures, the timbre is a result of interactions of frequencies in the sound. The objective measurement of timbre has been a long-running problem in acoustics, but it has only been with recent advances in signal analysis that timbre can be fully investigated (e.g. by [4] and [3]).

3.2. Interfaces and Design

Our system is to be developed for a desktop environment, for situations where the expressiveness and flexibility of the sound development is far more important than accurately recreating a sound from the real world. An example given by Gaver is his Auditory Icon illustrating a file being copied [1]. In this icon, the expressiveness of the pouring sound he uses is more important than using a less expressive realistic photocopying sound.

3.2.1. Auditory Icons

Auditory Icons were devised by Gaver [1]. They are auditory representations motivated by real-world sounds. They are parameterized so they can reflect changes in the underlying environment.

The major problem with Auditory Icons is that the parameterization is difficult. Even where the mapping between a perceptual description of a sound and the system state it represents is obvious, it is rarely easy to modify the sound signal in the correct way as standard sound editors operate at the signal rather than the perceptual level. The icons Gaver uses have been developed by studying the sources that create the sound, which is a slow process.

3.2.2. Earcons

Earcons are short musical segments that are abstract representations of data. In contrast to Auditory Icons, Earcons [6] do not attempt to describe an event with a real-world sound. Hierarchical Earcons, as used in the experiments by Brewster *et al.* [9], attempt to assign some structure to Earcons by mapping different attributes to different levels of description of the interface. For example, menus are described by the timbre of the sound, and menu items are described by the rhythm of the sound.

Earcons allow a much richer space of sound than Auditory Icons since Auditory Icons are independent of each other and can only be parameterized with respect to simple object interactions, such as a scrape. Earcons can be parameterized with a wide range of musical features, allowing a single Earcon to present much more information than an Auditory Icon.

3.2.3. Combined approach

A single Earcon is a complex unit formed of many notes. Earcons treat timbre as a single dimension. Earcons use musical concepts such as pitch, rhythm, duration and tempo as further dimensions.

Auditory Icons treat timbre as the combination of many dimensions and rarely use pitch or tempo. A rare example is given by Gaver's bouncing objects where the temporal proximity of events reflects the springiness and original height of the dropped object.

The timbre-level control exhibited in Auditory Icons is a useful extension to the musical-level control found in Earcons. With a single system that can manipulate at both these levels, a much richer design space can be presented to developers.

3.3. Timbre Spaces

To effectively control sound, we need a representation that is flexible enough to allow designers a variety of ways to use it. A timbre space is one such representation, and has been chosen for our work because the studies detailed below suggest a link between human perception and timbre spaces. This implies that designers should find it easier to create a desired sound with a timbre space representation than via traditional synthesis algorithms.

Grey performed perceptual experiments on timbre [4] where he played a series of sounds to volunteers and asked them to rate how similar the sounds were. He built a three-dimensional *timbre space* of the sounds, such that similar sounds were closest together. A physical description for each axis was found. For example, he related the first axis to the spectral distribution of energy.

Hourdin *et al.* [3] demonstrated an automated way to generate a similar ten-dimensional space. They compared their space with that of Grey, showing a correlation between the axes of the two. In their automated analysis, a set of input sounds is analysed so that each sound is described by how its frequencies change over time. Taking each frequency as a separate dimension, and the sound as a path through this multi-dimensional space, the sound can be projected into a lower dimensional timbre space by finding correlations across the sounds and mapping the most important features across all sounds to a timbre space axis. This has the effect of reducing the number of dimensions in the space.

There are many possible timbre spaces and each one has its own strengths and weaknesses for different tasks. Each combination of analysis method and dimensionality reduction produces a different space. By comparing a wide range of spaces, evaluated in terms of quality of sounds they can produce, the time it takes them to produce the sounds and the amount of data each needs to store, a space will be chosen that will suit fast and easy editing

and good reproduction of sound. To best fit these spaces into the existing work, the input sounds for generating these spaces will be the same as those used by Hourdin *et al.*

3.3.1. Signal Analysis

Signal analysis is the first stage in producing a timbre space. It is the process by which an input signal is broken into its constituent frequencies. A good overview of the many different methods for doing this can be found in Roads [10].

The two main analysis methods used for this project are the Short-Time Fourier Transform (STFT) and the Constant Q Transform (CQT). In general, the CQT produces fewer frequency bins since its logarithmic frequency scale matches that of the human ear more closely. The STFT computes faster and has greater resolution across the frequency range as it has a linear frequency scale.

3.3.2. Dimensionality Reduction

Principal Components Analysis (PCA) is the simplest of the dimensionality reduction techniques. In PCA, the data are rotated such that the directions where the data have the most variance is aligned with the primary axis and the other axes are aligned similarly with the remaining variance in the data. The original dimensions we use are the frequency bins from the signal analysis. By taking the 8 directions across these bins with the highest variance, a simplified representation of the sound is created.

4. SYSTEM DETAILS

The complete system we are developing includes an analysis engine based on the timbre space work described above. New sounds presented to the system are converted into paths in this space. The system allows manipulation of these sounds via their path representation. When an output is required, the system maps the path from the Timbre Space onto an FM synthesiser.

The sound manipulation component of the system allows morphing between sounds in the space. To generate completely new sounds, the paths can be warped into any shape in the space. For example, if the "alarm response" described earlier was coded into the space, a sound could be made more alarming simply by adjusting its path closer to the "alarm region".

This section will now discuss the work and evaluations completed so far.

4.1. Conceptual model

In Figure 1, the overall design of the system can be seen. On one side there are the synthesisers to choose from and on the other are the object models that will be manipulated by the user. Connecting these two sets is the perceptual mapper that will convert an object level description into a timbre space representation and then map this directly into a synthesiser to produce the required sound.

At the object level, properties of an object can be described and manipulated. Each of these properties will relate to a path or transformation in timbre space. The objects can be defined either by building up a perceptual model of each property via its physical parameters, such as with the alarm response, or by taking a set of sample sounds representing the range of values the property will take and analysing them to define the paths and transformations automatically.

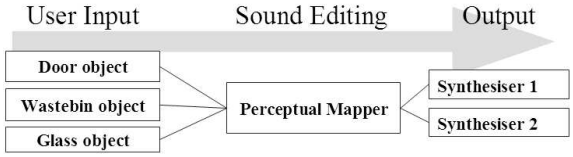

Figure 1: Conceptual Model of the design system.

An object model could easily be produced for Gaver's Auditory Icons using the latter approach. This allows direct comparison of this system with Auditory Icons.

The timbre space captures the perceptual information in the sound. This is the mediator between the object properties and the synthesis engine. It is here that the perceptual properties are realised and the transformations are transformed.

The synthesis engine is a self-contained module such that any synthesiser can be used. The choice of synthesiser will be made for reasons of speed, accuracy or coverage of perceptual space. It is expected that each synthesis engine will only cover a portion of the entire perceptual space. A Gaver-style synthesiser optimised for contact sounds will produce poor-quality representations of sustained sounds such as the flute, whereas an analogue synthesiser whose sounds are made of summed waves will have difficulty simulating the complex spectrum of a scraping sound which has many inharmonic frequencies across the spectrum.

4.2. Analysis of Timbre Spaces

Analysis has been performed on a range of musical and synthetic signals including output from an FM synthesiser and a selection of 27 sounds selected to match those in Hourdin *et al.*'s experiments.

A set of timbre spaces have been compared on their compilation speed, timbre space size and accuracy of sound reproduction. To analyse the spaces, the 27 sounds have been mapped into each timbre space and reconstructed from that space using additive synthesis without any changes to their paths. A selection of the resultant sounds are available on our Website.

The STFT and CQT algorithms have been tested. Boxcar, Gaussian, Kaiser 6 and 8 windows have been used. Window sizes of 2ms to 400ms have been used with overlaps of one to eight windows. For the CQT, three frequency resolutions have been used.

The CQT has proved to be quicker than the STFT and produces less output data for the same input. The spaces take up between 20Mb and 500Mb to store the original sounds as paths, with size increasing linearly with resolution. Generally, the accuracy of the generated sounds is improved greatly when Kaiser windows are applied as opposed to Gaussian windows or no windows, although this has only been analysed by the authors' perception.

The STFT will not compile at its highest resolution setting as the space is too large. The CQT output sounds at the highest frequency resolutions produce a lower quality sound than lower resolutions despite requiring more storage. At lower resolutions, the STFT based timbre spaces produce slightly better quality than the equivalent CQT for any time resolution. Again, this has only been tested by the authors' perception.

When the STFT is compiled with a boxcar window, the PCA compilation takes exponentially longer as the time resolution increases. In every other case, compilation time appears to grow linearly against time resolution. This suggests that there is much less

structure to the STFT output when no window is applied, which makes the PCA much less effective in this circumstance.

5. FUTURE PLANS

With this system complete, a range of experiments is possible. In particular, sounds can be developed according to both Auditory Icon and Earcon principles such that in any given interaction, the most important information will be mapped to the Auditory Icon portion of the sound (the note) and secondary information will be mapped to the Earcon properties of the sound (the pattern of notes).

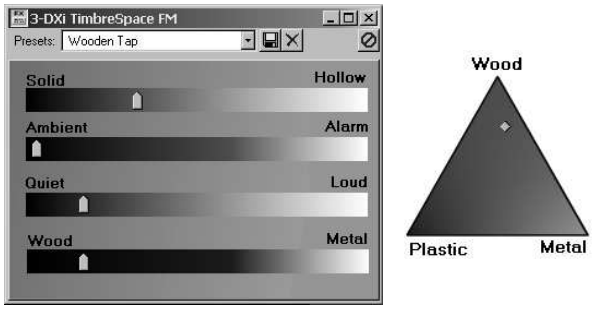

Figure 2: Possible designs for a GUI for audio design.

Our completed tool will be implemented within a MIDI environment in order to take advantage of existing work on Earcon design. It will accept MIDI signals to control pitch and amplitude and will add an interface to allow real-time editing of timbre.

The interface will allow selection of any of the pre-selected 27 timbre paths included in the timbre space as well as any others the user has added to the system. For each of these paths, a selection of transforms will be made available. These transforms will include morphing between two or more paths, looping the sound within a path, scaling the pitch of the path or any other user-defined transformation within the timbre space. These transformations include making the sound more 'alarming' by adding properties of the 'alarm response' to the sound, or by making the sound more like sounds previously identified as 'hollow'.

Two possible GUIs for interacting with this device are shown in Figure 2. The slider interface allows the user to pick an original sound and modify it according to a selection of transformations. Only some of these transformations will be visible at any one time as there are limitless numbers of such transformations.

The triangular interface allows the user to select three sounds from the space and morph between those sounds by moving the pointer within the triangle. In addition, the points of the triangle could also be used to represent perceptual transformations.

The strength of each transform, or the relative strengths of the timbres affected by the transform, can also be controlled by any MIDI controller. This allows the transform to be adjusted over time by any external input and allows the change in the timbre to be recorded in the same place as the change in the melody.

6. CONCLUSIONS

We have set out to enable designers more flexibility when developing sounds. The timbre space has been chosen to represent the sounds due to the perceptual basis afforded to it by the work of Grey. Preliminary results show that the Timbre Space is heavily reliant on the quality of the audio analysis stage. STFT looks to produce the best quality output at this moment, but experiments on other techniques are ongoing.

The system shows promising results in terms of sound quality using a low-complexity timbre space. With such a space, designers have the opportunity to explore Auditory Icons and Earcons within the same environment, opening up new platforms for experiments on how audio interfaces can be designed. The inclusion of perceptual transformations allows the timbre space to capture perceptual knowledge in a format that makes it easy for a designer to explore and exploit that knowledge. The designers can think in terms of what they want to hear or what response they want to elicit rather than in terms of how the sound is constructed.

7. ACKNOWLEDGEMENTS

This work is part of a PhD studentship funded by EPSRC.

8. REFERENCES

[1] W. W. Gaver, "Synthesizing auditory icons", in *ACM Interchi '93*. 1993, pp. 228–235, ACM.

[2] E. D. Mynatt, M. Back, R. Want, M. Baer, and B. Ellis, Jason, "Designing audio aura", in *CHI '98*. 1998, pp. 566–573, ACM.

[3] C. Hourdin, G. Charbonneau, and T. Moussa, "A multidimensional scaling analysis of musical instruments' time-varying spectra", *Computer Music Journal*, vol. 21, no. 2, pp. 40–55, 1997.

[4] J. Grey, "Timbre discrimination in musical patterns", *Journal of the Acoustical Society of America*, vol. 64, pp. 467–72, 1977.

[5] W. W. Gaver, "How do we hear in the world? explorations in ecological acoustics", *Ecological Psychology*, vol. 5, no. 4, pp. 285–313, 1993.

[6] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and icons: Their structure and common design principles", *Human Computer Interaction*, vol. 4, no. 1, pp. 11–44, 1989.

[7] P. Masri, A. Bateman, and C. N. Canagarajah, "A review of time-frequency representations, with application to sound/music analysis-resynthesis", *Organised Sound*, vol. 2, no. 3, pp. 193–205, 1997.

[8] J. Chowning, "The synthesis of complex audio spectra by means of frequency modulation", *Journal of the Audio Engineering Society*, vol. 21, no. 7, pp. 526–534, 1973.

[9] S. Brewster, P. Wright, and A. Edwards, "A detailed investigation into the effectiveness of earcons", in *First International Conference on Auditory Display*, G. Kramer, Ed., Santa Fe Institute, Santa Fe, NM, 1992, vol. Auditory display, sonification, audification and auditory interfaces., pp. 471–498, Addison-Wesley.

[10] C. Roads, *The Computer Music Tutorial*, Massachusetts Institute of Technology Press, 1996.