SONIFICATION OF SOUND: TOOLS FOR TEACHING ACOUSTICS AND AUDIO

Densil Cabrera and Sam Ferguson

Faculty of Architecture, Design and Planning, University of Sydney
NSW 2006, Australia
densil@usyd.edu.au

ABSTRACT

This paper describes a collection of examples of how the teaching of acoustics and technical audio may be aided through sonification. Examples are fixed demonstrations, small programs in Max/MSP, and sonifications as part of a sound analysis program (PsySound3) developed by the authors. Examples include auditory graphs of frequency-dependent parameters used in architectural acoustics, sonifications of room modal distributions, sonifications of room impulse responses, simplification of input sound using spectral moments and the Hilbert transform, interactive sonification of head-related transfer functions and vowel formants, and sonifications of sound analysis parameters, including the output of psychoacoustical models.

[Keywords: Education, Sonification]

1. INTRODUCTION

In teaching technical audio and acoustics at a graduate level for a number of years, the authors have found that teaching is most effective when (i) students are presented with concepts in different modes (e.g., through theory, equations, experiments, software simulations, illustrations, animations, auralizations, etc) and (ii) students interact with the modes of presentation themselves. While there are many sources for most of these modes, the authors were struck by the fact that sonification has received little attention in teaching audio and acoustics. In this context, 'sonification of sound' may be a useful concept, whereby data about sound are rendered audible to facilitate students' understanding of phenomena.

Auralization may be regarded as a special type of sonification. Auralization aims to create a virtual acoustic reality for the listener to experience (e.g., the sound of a source within a computer modeled room). However sonification is a much broader concept than this, and can involve presentation of data to a listener in many other ways – such as auditory graphs, the implementation of equations of acoustic theory to produce auditory stimuli, and transformations of sound recordings or real time sound input so that relevant features are more easily heard.

Representing sound with sound can have the advantage of conveying the meaning of the data straightforwardly to the listener, especially in cases where the phenomena modeled are relevant to human perception. For example, representing octave band reverberation time with a tone centered on the nominal frequency of the band and possessing the duration of the reverberation time conveys an experience of the represented time and frequency. Conventional representation, using numbers in a table or bars on a chart, involves an abstraction of the phenomenon that distances the meaning from the representation. Of course, such abstraction has advantages, but these can be complemented by the advantages of the sonification.

Outlines of this project have been presented previously [1, 2] prior to public dissemination of the sonifications. The purpose of this paper is to present examples that are now freely accessible via www.arch.usyd.edu.au/~densil/sos. It also serves as an invitation for others in the auditory display community to contribute examples.

Implementations of the sonifications described in this paper are through fixed demonstrations, Max/MSP patches, and as part of a sound analysis program called PsySound3. Max/MSP is a patching signal processing language, a runtime version of which is freely available from www.cycling74.com. PsySound3 is a program by the authors, which is freely available from www.psysound.org.

The following sections describe the sonifications.

2. SPECTRAL DATA USED IN ARCHITECTURAL ACOUSTICS

Reverberation time (T) is the time taken for a room soundfield to decay by 60 dB. As predicted by Sabine theory, reverberation time is given in equation 1, where $\ln(10^6)$ represents 60 dB of exponential decay, 4 is derived from the statistical distribution of energy flow in a diffuse sound field in terms of its interaction with a surface, V is room volume in cubic meters, c is the speed of sound (which may be taken as 344 ms^{-1}), α is the random incidence absorption coefficient of a surface, and S is its surface area in square meters. The sum of the product of α and S for all surfaces represents the total sound absorption in a room.

$$T = \frac{\ln(10^6) \times 4V}{c\sum \alpha S} \approx \frac{0.16V}{\sum \alpha S} \quad (1)$$

While there are more sophisticated and accurate fomulations for reverberation time, Sabine's simple equation is still in wide use, and is an important tool in acoustics education in introducing the most important contributors to reverberation time to students. The absorption coefficient that forms part of this equation is clearly an important predictor of room acoustical performance, and is used extensively for design in architectural acoustics. An absorption coefficient is a number between 0 and 1 (at least theoretically), representing the proportion of incident sound energy that is not reflected by a surface. Absorption coefficients are frequency-dependent, and are usually published in octave bands from 125 Hz

to 4 kHz. The purpose of the sonification described here is to allow students to hear absorption coefficient data in a meaningful way, and to relate it to reverberation time.

Rather than sonifying the absorption coefficient directly, we prefer to sonify 1-α (the reflection coefficient, β) because the amount of sound experienced in the sonification then corresponds to the proportion of sound left after a reflection. Our sonifications of absorption data are done using octave bands of noise (125 Hz – 4 kHz) which together form pink noise within the band limits. The level of each octave band is controlled by 10log(β) – meaning that the sound representing an absorption coefficient of 0.9 would be 9.5 dB weaker than that representing an absorption coefficient of 0.1. The six octave bands of noise are played together using this level control. The duration of each band is equal to the reflection coefficient in seconds. This redundant encoding in the auditory graph gives an indication of the effect of a material on the reverberation time of a room (if the room were predominantly surfaced by that material). The duration does not present the actual reverberation time, but gives a good impression of the spectral weighting of reverberation.

The sonification is implemented in Max/MSP, and the control interface is shown in Figure 1.

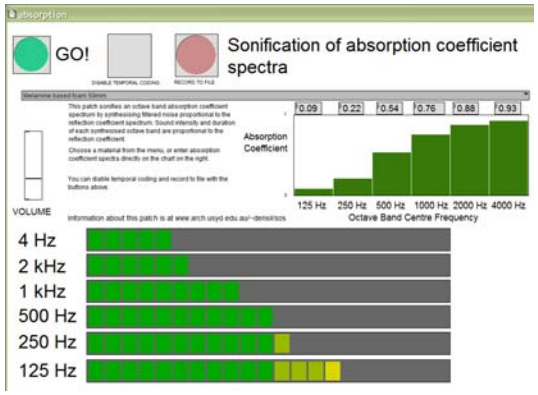

Figure 1. *Control and visual display interface for the absorption coefficient sonification. Absorption spectra for materials are selected from the drop-down menu (180 materials are listed). Alternatively, the absorption coefficient chart can be edited directly. The audio output is visually displayed in the lower part of the interface.*

The selection of materials based on their absorptive properties is a context where auralization is an effective communication tool for non-experts: an acoustical designer will allow their client to listen to the effects of various surface materials in a computer model of a room. By contrast this absorption coefficient sonification focuses on the acoustic properties of individual materials, and would be too abstract for non-specialists. However, it is effective in acoustics education because it provides a simple aural link between the abstract data (which the students are learning about) and the sound of rooms.

3. ROOM MODAL DISTRIBUTION

Room modes develop from periodic reflection patterns in rooms, and may be seen as discrete peaks in the transfer function between two positions in a room. In the high frequency range, modes are densely spaced, and exhibit behavior that is best understood statistically. In the low frequency range the effects of particular modes can be strong, and the frequency that divides these two frequency ranges is known as the Schroeder frequency (which approximates to 2000 times the square root of reverberation time divided by room volume). Since the effect of room modes in the low frequency range can be very strong, there has been much discussion about how room response in this range can be optimized [3]. Part of this discussion has been what the proportions of rectangular rooms should be for an even distribution of low frequency modes.

For a rectangular room, room modal distribution in frequency and space domains can be modeled using simple equations [4]. The frequency of any room mode is predicted by equation 2, where n_x, n_y and n_z are integers greater than or equal to zero, L_x, L_y and L_z are the length, width and height of the room, and c is the speed of sound (344 ms^{-1}).

$$f_{n_x,n_y,n_z} = \frac{c}{2}\sqrt{\left(\frac{n_x}{L_x}\right)^2 + \left(\frac{n_y}{L_y}\right)^2 + \left(\frac{n_z}{L_z}\right)^2} \qquad (2)$$

Axial, or one-dimensional, modes have only one non-zero n value, and have the longest mean free path (meaning that they tend to decay most slowly, and have the narrowest resonance peak in the frequency domain). The lowest frequency mode is axial, and axial modes are relatively common in the low frequency range, and their modal density remains constant throughout the frequency range. A harmonic series is produced by the axial modes of each room dimension. Tangential (two-dimensional) and oblique (three-dimensional) modes increase in modal density with frequency (following a square and cube relation to frequency respectively).

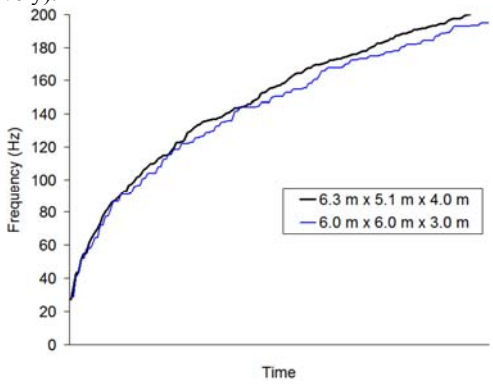

Figure 2. *Visual representation of the data used in the sequential sonification of room modes.*

We have sonified this equation in more than one way. Firstly, it has been sonified sequentially as a fixed example. This is done by calculating the room modes, and sorting them from low to high frequency. These are sonified as complex tones (so that the low frequency modes are audible as virtual pitches), and played as an ascending 'scale' or melody, with 100 ms per mode. In our demonstration, we contrast a room 6.3 m x 5.1 m x 4.0 m with one that is 6.0 m x 6.0 m x 3.0 m. This provides a clear aural demonstration of how room proportions can influence modal distribution: in the first instance, the melody is relatively smooth,

while in the second instance there are many sustained pitches with substantial steps to the subsequent pitch. The result is illustrated in Figure 2.

Our second sonification involves an interactive simultaneous synthesis of low order room modes, implemented in Max/MSP. In 1946, Bolt [5] produced a method of selecting good rectangular room proportions (meaning avoiding gaps and clusters in room mode frequency distribution) using a two dimensional chart with a footprint outlining the area of good proportions. This is used as the control surface for the sonification (Figure 3) because it is very well known in this field. The horizontal and vertical axes represent width and length as ratios to height. The user can move the cursor within this space, and hear either low order axial modes, or low order axial and tangential modes, with 100 Hz taken as the base of the ratios (i.e the first vertical axial mode is assigned a frequency of 100 Hz). Listening to axial modes only, the sound is a triad of complex tones (because the modes form three harmonic series), and so students can play with the interface to explore chords (as a trivial example, major and minor triads can be created by selecting appropriate room proportions). There is an inverse coding of consonance in this auditory display – dissonance is associated with regular mode spacing (favorable room proportions), while consonance is associated with mode clustering (unfavorable proportions). Amplitudes of the tones are all the same so that beating effects between modes are maximized in the sonification. When tangential modes are used, the sound becomes much thicker. The sonification does not currently synthesize oblique modes, but this may be added in the future.

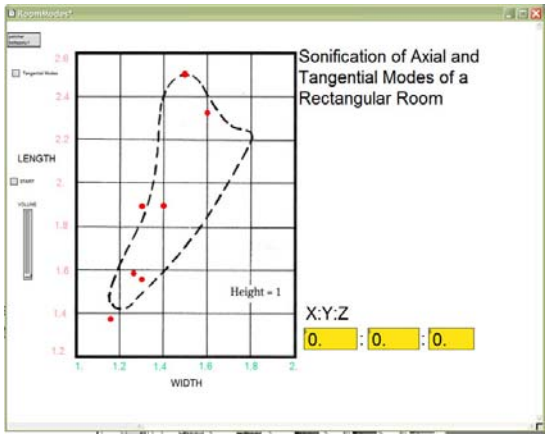

Figure 3. *Control and visual display interface for the rectangular room mode sonification. By moving the cursor around the chart, modes for the specified room proportions are sonified. Tangential modes can be switched on or off.*

Another way in which room modes could be sonified is by considering their spatial distribution. While we have not specifically implemented this as a sonification, this was developed as sonification for artistic purposes by one of the authors several years ago (Figure 4). The mode-related transfer function from a source to a receiver position in a room is a product of six cosines for each mode (involving the x,y,z coordinates of both source and receiver) moderated by the mode damping factor. This models the spatial patterns of pressure nodes (experienced as silence) and antinodes (maximally audible sound) of a rectangular room. In that implementation, a binaural mode space is generated for the listener, taking the listener on a random walk through a rectangular room 6.35 m x 5.1 m x 4.0 m (with modes synthesized up to 1 kHz). A sample of the result is available from the author's website.

Figure 4. *The pseudo-binaural playback system used for 'Interior' when exhibited in the Tin Sheds art gallery in Sydney, Australia.*

4. IMAGE-SOURCE MODEL

Image-source modeling is another simple way of predicting the behavior of sound in enclosures. Rather than modeling the modal behavior of sound waves, this method models specular reflections of sound rays (such behavior tends to be limited to high frequencies, and situations where surfaces are smooth). While the technique can be applied to rooms of any shape, its application to rectangular rooms is very simple and easily generalized, and so is useful for teaching purposes. The concept can be appreciated by imagining a rectangular room with surfaces consisting entirely of mirrors, containing a source and receiver. From the receiver's perspective, images of the room can be seen extending to infinity in all directions, and each image room contains an image source. This is illustrated in two dimensions in Figure 5.

If a corner of the real room is taken as the origin of Cartesian space, and (x_0,y_0,z_0) as the real source coordinates, then the x coordinates of all image-sources are calculated by equation 3, and the y and z coordinates are found similarly.

$$x_{n_x} = L_x\left(n_x + n_x \bmod 2\right) + x_0\left(-1\right)^{n_x} \tag{3}$$

Here, L_x is the room length in the x dimension, and n_x is an integer designating the image room coordinate (n_x=0 is the real room coordinate), as indicated by the numbers on Figure 5.

The distance, r, between source (or image-source) and receiver (x_1, y_1, z_1) is given by Pythagoras' theorem in three dimensions (equation 4):

$$r = \sqrt{\left(x_{n_x} - x_1\right)^2 + \left(y_{n_x} - y_1\right)^2 + \left(z_{n_x} - z_1\right)^2} \tag{4}$$

The received intensity (or squared amplitude), I, of sound from a source or image-source is related to the source power P, the absorption coefficient α (with an exponent based on the number of times the sound ray interacts with a room surface) and the distance traveled by the sound ray, r, as expressed in equation 5.

$$I = P\frac{(1-\alpha)^{|n_x|+|n_y|+|n_z|}}{4\pi r^2} \quad (5)$$

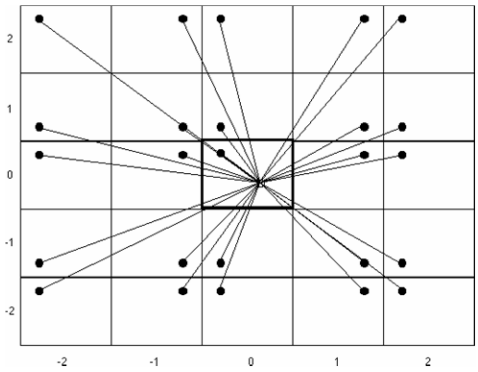

Figure 5. *Illustration of the image-source concept for rectangular rooms (in two dimensions), where each dot represents an image-source or source, with rays drawn to the receiver in the real room.*

Like the rectangular room modes theory, we teach these concepts partly through a spreadsheet implementation (which allows students to experiment with room dimensions, source-receiver positions and the surface absorption coefficient, yielding charts of the impulse response) and through measurements in real rooms. However, the spreadsheet approach requires considerable imaginative effort, and physical measurements have considerable logistical aspects which mean that we can only do them once or twice in a course. Therefore a student, Luis Miranda Jofre, has recently implemented this theory as a simple sonification, using Max/MSP. The advantage of this over conventional room modeling software (such as Odeon, Catt-acoustic or Ease) is that the model is dedicated to the demonstration of these simple principles of geometric acoustics, and also the software is freely available. A key aspect of making this sonification more than a crude auralization is the inclusion of a user-controlled scaling factor controlling the speed of sound (alternatively this could be thought of as controlling room size for a constant sound speed). This allows the user to scale the impulse response between a rhythm (slow speed of sound) and a timbre (fast speed of sound), so that time-domain and frequency-domain effects can be explored interactively.

This implementation is made crudely binaural by modeling two receivers 17 cm apart, which results in inter-aural time differences. Interaural level differences for a spherical head at 1 kHz are used. Binaural sonification allows the user to examine factors that affect the spatial distribution of a binaural impulse response, including symmetry (which can strongly affect inter-aural cross correlation or IACC). By scaling the speed of sound, the effect of IACC can be heard as a spatial rhythm or as auditory source width. While this could be refined through the implementation of full head-related transfer functions, that level of detail is unnecessary to demonstrate the effect.

The sonification implements all reflections from (n_x, n_y, n_z) of (-2,-2,-2) to (2,2,2), and so includes all first and second order reflections, as well as some higher order reflections, or 125 impulse arrivals per ear. This is simply implemented through a level control and a delay line based on the above equations. The user controls the source and receiver positions, room dimensions and absorption coefficient.

5. MEASURED ROOM IMPULSE RESPONSES

Room impulse responses are a very important way of characterizing the acoustical characteristics of a room, at least in terms of how sound travels from one position to another within the room. The impulse response is the time response of a system (such as a room) to a single 'click' (although this is usually measured indirectly, for example through a swept sinusoid, which is deconvolved to yield an impulse response with a high signal to noise ratio). Impulse responses are used for determining room acoustical parameters (such as reverberation time, early decay time, clarity index, inter-aural cross-correlation coefficient and speech transmission index), and for auralization (by convolving the impulse response with an arbitrary dry recording).

We have identified a collection of simple techniques that may be used to sonify measured room impulse responses. We recommend applying these to binaural recordings, so that spatial characteristics of room responses can be discerned. These are prepared as fixed demonstrations, and can be performed easily by anyone who is familiar with conventional sound editing software.

5.1. Time Stretching, Reversal, and Mirroring

The fine temporal detail in an impulse response is difficult to hear by playing it. The information is presented so quickly that most information is inaudible, partly because of auditory temporal masking. Even in a large room, usually the pattern of distinct reflections is within the first 100 ms of the direct sound. A very simple solution is to slow down the recording, which transforms the early reflection pattern into a clearly discernable sequence of pulses. A scaling factor of 16 means that 100 ms is heard over 1.6 seconds, and 10 kHz is heard as 625 Hz.

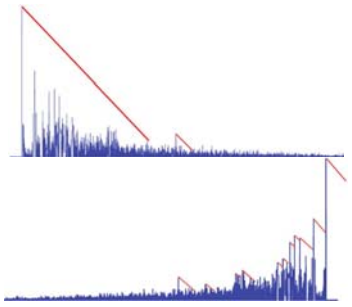

Figure 6. *The concept of temporal masking asymmetry as it applies to a room impulse response and its reversal. Red lines roughly indicate forward masking (backward masking effects tend to be much weaker).*

Another approach to this is to time-reverse the impulse response. One advantage of this is that temporal masking is asymmetric, with very little pre-masking, and much more post-masking.

Therefore (as illustrated in Figure 6), the effect of post-masking can be greatly diminished through time-reversal. The early reflection sequence often sounds rough when reversed (due to the audibility of multiple peaks) when it was smooth in its original form. Following the reversed impulse response immediately with the original ('mirroring'), gives the advantages of both presentation modes (enhanced audibility of early reflections in the reversed response, and the experience of room's sound decay through the original).

5.2. Spectral Weighting

An impulse, which is the signal at the heart of an impulse response, has a 'white' spectral power distribution. This means that half its power is in the top octave, three-quarters of its power in the top two octaves, and so on – leaving very little power in the low frequency range. Considering that room acoustics analysis is normally done in octave or one-third octave bands, the impulse response can be thought of as having a heavy high frequency bias. Other considerations are that the spectral energy distribution of normal sound sources (such as speech or music) is nothing like white, and often is closer to logarithmic. Therefore there is a strong argument that impulse responses should be spectrally weighted if they are to be assessed by the ear, possibly based on a logarithmic (or pink) energy distribution.

5.3. Reverberation Time

It is possible to listen to reverberation simply by playing a room impulse response, but difficult to discern details of the reverberation time from that. An auditory graph of octave band reverberation time can be constructed by synthesizing a pure tone at each octave band center frequency, which is assigned a duration equal to the reverberation time. We play these octave-related tones (125 Hz – 4 kHz) with simultaneous onsets, so a listener must hear their offsets to discern the reverberation times represented. Audibility of these offsets is enhanced by putting a sudden increase in the level of each tone just before the tone ceases.

The key to the success of this auditory display is that time is represented by time. The auditory graph can be played together with the impulse response, so that a listener can easily relate one to the other. On first hearing this, listeners may be surprised by the length of a reverberation time, because a 60 dB decay is usually inaudible in a room impulse response (especially considering its 'white' spectral weighting).

In many reverberation-sensitive rooms, somewhat longer reverberation times are desired in the low frequency range than in the high frequency range. This auditory graph provides an immediate experience of the extent to which such a criterion is met, since it would be represented by a pattern of falling pitches (as each tone builds up and ceases). Deviations from such criteria are clearly audible as a more chaotic pitch pattern.

5.4. Auto-convolution and Auto-correlation

The fine pattern of peaks in a room impulse response is usually all but impossible to hear because clear tones are usually not present (small reverberant rooms are an exception). Auto-convolution is equivalent to squaring the complex spectrum, and so is a simple way to exaggerate the contrast in a spectrum. Auto-correlation has an identical effect on the magnitude spectrum, but results in an even (or symmetric) function, like a very long linear phase filter. An auto-convolution (or auto-correlation) sequence can be performed, whereby each generation is processed with itself, yielding spectral powers of 2, 4, 8, 16 and so on. Eventually any sound processed in this way degenerates to a single pure tone (corresponding to the highest peak of the original spectrum). The most interesting sound (wherein a maximum number of peaks are audible) is somewhere between the original and the fully degenerated version. Such a sequence is illustrated by Figure 7. Temporal smearing accompanies this process.

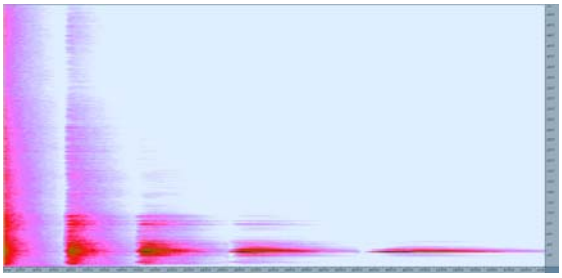

Figure 7. *Spectrographic view of a room impulse response auto-convolution sequence. The original impulse response is on the left, followed by progressive auto-convolutions. The frequency scale (vertical) is 0-5 kHz (linear), and the time scale duration is 15 s.*

6. SPECTRAL MOMENTS

Spectral moments are a way of simplifying a spectral magnitude distribution to a small number of values. The first moment (or centroid) can be thought of as representing the 'center of gravity' of a spectrum, and is essentially the mean frequency. The spread of the spectrum can be represented by its standard deviation (the square root of variance, which is the second moment). The third moment is skew. The fourth moment (kurtosis) compares the peakiness of the spectrum to that of a normal distribution, although we have not used the fourth moment in our sonification.

The calculation of spectral moments is given below for the first three moments (centroid, standard deviation and skew), where f_n is the frequency of each spectral component, and a_n is its magnitude, which is optionally squared in the calculation. Squared amplitude is proportional to the power, and in a physical sense is the better choice (because phase is not relevant to this calculation). However unsquared amplitude is closer to loudness perception, and so may be more appropriate for some applications.

$$C = \frac{\sum f_n a_n^{(2)}}{\sum a_n^{(2)}} \tag{6}$$

$$SD = \sqrt{\frac{\sum (f_n - C)^2}{\sum a_n^{(2)}}} \tag{7}$$

$$Skew = \sqrt{\frac{\sum\left(\left(f_n - C\right)/SD\right)^3}{\sum a_n^{(2)}}} \qquad (8)$$

In addition to the question of whether amplitude is squared or not, the frequency component distribution and the frequency units are important in this type of analysis. The most common choice for analysis is a linear distribution and linear units (Hz). However in our Max/MSP implementation, we use a logarithmic distribution (implemented through octave-spaced filters) and logarithmic frequency units. This is achieved by taking the logarithm (base 2) of the frequency prior to the calculation, yielding octave units. Then the result is taken as the exponent of 2 to convert back to Hertz for sonification and visualization. This approach appears to give more useful results (i.e. values that work well in sonification) than using a linear distribution and linear units, especially for the higher moments. While octave-spaced filters do not provide high precision, we have found that these works surprisingly well (e.g., in the centroid matching the frequency of a pure tone), and in any case, this sonification is devised for demonstration purposes rather than precise analysis. A more precise approach is available through PsySound3, outlined towards the end of this paper.

Centroid is sonified as a pure tone that has a frequency equal to the centroid. The amplitude of this tone is controlled by the measured amplitude of the input wave (i.e., the term that is also used in the denominator of the spectral moments). This not only makes the sonification easier to listen to, but makes sense because the centroid of silence is meaningless, and the centroid of high level sound would dominate over low level sound of equal duration in a long term spectral centroid calculation.

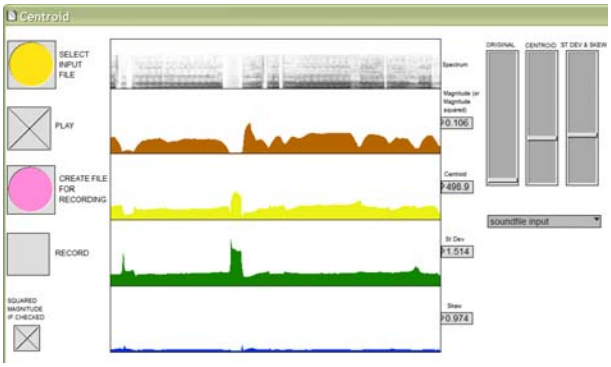

Figure 8. *Control and visual display interface for the spectral moments sonification.*

Standard deviation and skew are sonified as a noise band. The center frequency of this band is derived from the difference between the centroid and the skew. Hence, if the skew is positive, the center frequency is greater than the centroid. The width of the noise band (between upper and lower cutoff frequencies) is one standard deviation. All of these calculations are done with logarithmic frequency units, as indicated above.

The interface for this sonification is shown in Figure 8. The user can select the type of input (real-time or sound file) and can use three volume controls for the sonification: one for the original input, one for the centroid tone, and one for the noise band representing standard deviation and skew. These values are also displayed visually as a moving chart.

This sonification is very effective in bringing the concept of spectral centroid to life. The sonification of standard deviation and skew are interesting, but more obscure to the ear. Sound recordings that produce informative results can include room impulse responses, speech recordings, music recordings, musical instrument tones (for timbre analysis) and many other types of sound including real time improvisation.

7. HILBERT TRANSFORM

A similar style of sonification is offered through a simple Hilbert transform Max/MSP patch that we have developed. A Hilbert transform can be performed simply by phase shifting all components (in the frequency domain) by $-\pi/2$ and returning back to the time domain. For a finite length sampled waveform, extremes of the spectrum (around 0 Hz and the Nyquist frequency) are not used, because the Hilbert transform is ineffective for these. In its usual application, the original time series (no phase shift) is taken as real, while its Hilbert transform is taken as imaginary. The resulting magnitude of this complex waveform represents instantaneous amplitude, while the rate of change of the resulting phase represents instantaneous angular frequency. We only use the instantaneous amplitude (Hilbert envelope) from the transform in our patch.

Our patch allows the user to select a carrier signal (440 Hz pure tone, pink noise, the time-varying centroid of an input waveform, or no carrier). The Hilbert envelope, squared Hilbert envelope (which increases the contrast), or differenced Hilbert envelope (which produces non-zero results when the envelope changes) can be selected to amplitude modulate the carrier. Smoothing can be applied to the envelope.

One application of this patch is a rhythm-to-pitch transform. This is done by extracting the envelope (without carrier) from a relatively long duration recording, and speeding this up 100 times (using a sound editing program). This scaling factor is appropriate because it shifts the fluctuation frequency range of maximum human sensitivity (around 4 Hz) to the range of maximum pitch sensitivity [6]. Periodicity in rhythm is then heard as complex tones, the timbre of which is derived from the rhythmic structure.

Like the spectral moment sonification, the purpose of this tool is to draw out a simple abstraction from the sound, as well as to demonstrate a signal processing technique. Envelope extraction can be used to reduce an arbitrary sound to what could be considered to be its simplest attribute. In the case of the Hilbert transform, the process is simple but the concept is subtle, so it is very helpful to give students an audio demonstration, complemented by theoretical, numeric and graphical teaching material.

8. HEAD RELATED TRANSFER FUNCTIONS

In spatial hearing, the head related transfer function (hrtf) is the ratio of the complex frequency response from a source to an ear to that from the same source to a point that would be in the center of the head on the inter-aural axis (with the head absent). While binaural difference cues form the basis for localization between left and right, spectral cues found in the hrtfs are used to identify the polar angle (e.g., front-back and above-below) of the sound source. Reproduction of appropriate binaural difference and

spectral cues can produce a convincing externalized and accurately localized auditory image. However, a major issue in this field is that each person's direction-dependent set of hrtfs is distinctive, and substituting one person's for another's tends to produce vague and inaccurate localization [7]. This is mainly because the physical form of the external ear (especially the pinna) varies substantially between individuals (both in terms of shape and size). The main spectral features for hrtfs are above 2 kHz, and some important features are at very high frequencies (around 8 kHz). As shown by Iida *et al.* [8], these features can be simplified effectively using parametric filter functions, and we have used this concept in our sonification.

For the sonification, we use a parametric filter with an interface allowing the user to tune two notches and two peaks, adapted from the parametric hrtf models of Iida *et al.* [8]. In this model, the tuning of Peak 2 (a quarter-octave peak between 7 and 9 kHz) mainly affects image elevation. On the other hand, Peak 1 (a broader peak between 2-5 kHz) does not vary much with source position for a given individual, and so is thought to provide a stable reference against which other spectral variation is assessed (especially the two main notches). Notch 1 varies between 5 and 10 kHz, and Notch 2 between 8 and 11 kHz. The aim of the user might be to explore their own hrtfs by trying to generate an image at a particular polar angle (e.g., 0 degrees) by manipulating these. We supplement this by allowing the user to change the interaural time difference (±1 ms) and broadband interaural level difference (±10 dB), which together control the lateral angle of the auditory image.

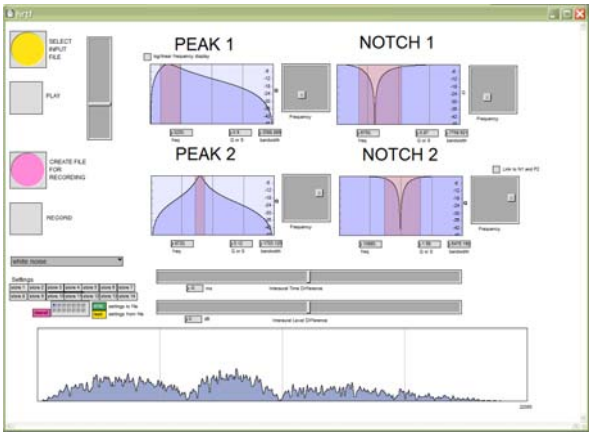

Figure 9. *Control and visual display interface for the parametric hrtf sonification. Each filter is controlled in two dimensions (center frequency and Q), and the output spectrum is displayed visually.*

9. FORMANTS

Vocal formants are resonances formed in the vocal tract which are used to produce and distinguish vowel and vowel-like sounds. The sound of vowels can be approximated through the tuning of just two formants, the first of which is related to how open the mouth is, and the second of which is related to the position of a constriction within the mouth formed by the tongue. The formants act as filters on the complex tone produced by the vocal folds, and the frequencies of each formant and of the complex tone can be independently controlled. Vowel formant frequencies vary between regional accents. Figure 10 shows approximate formant values for male Australian speech, or at least a subset thereof [9].

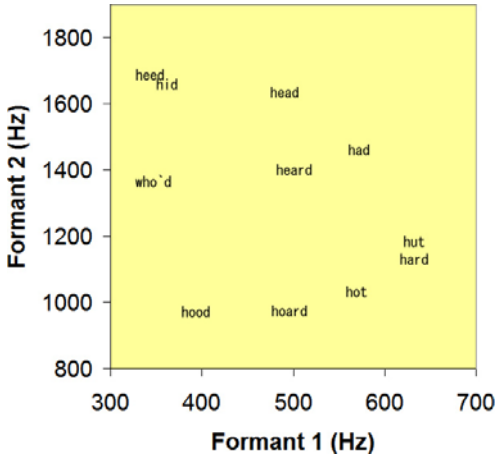

Figure 10. *Frequencies of the two major vowel formants for male Australian speech.*

The implementation of the formant sonification is similar to the parametric hrtf sonification. However, only two filters are needed, and a complex tone (with some vibrato) is provided as a source since this is similar to the source of the vocal system. The user's primary control is simply the frequency of the two formants, which can be explored using the mouse in the two-dimensional control space on the right side of Figure 11. The *Q* (or peakiness) of the two formant filters can also be adjusted.

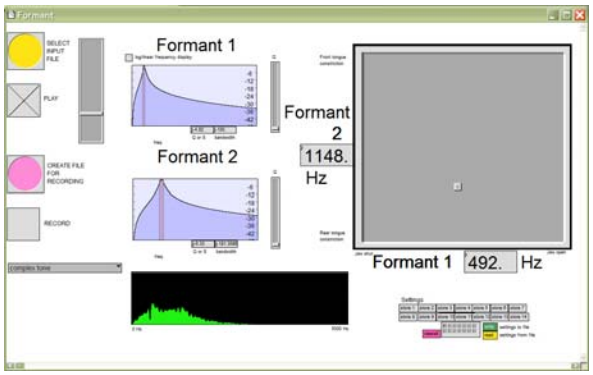

Figure 11. *Control and visual display interface for the speech formant sonification.*

The purpose of this very simple sonification is to give students an accessible tool to explore the concept of vowel formants. The two-dimensional controller allows students to attempt to create dipthong sounds, as well as steady-state vowels. After experimenting with the complex tone input, and perhaps storing preset formant combinations, students can use the tool to filter arbitrary sounds (using sound-files as the input), and so it could even have some use in the students' audio production projects.

10. PSYSOUND3

PsySound3 is software developed by the authors that implements acoustical and psychoacoustical algorithms for the analysis of sound recordings [10]. Algorithms currently implemented include (but are not limited to) a sound level meter module, generic digital signal processing (FFT, cepstrum, Hilbert transform, and auto-correlation), dynamic loudness, sharpness, roughness, loudness fluctuation and virtual pitch analysis. While the main purpose of this program is to provide free and easy access to sophisticated analysis of audio recordings with detailed results given as numeric output, the program also can sonify the analysis results (as well as providing visualizations).

Available sonifications depend on the data structure. The main three data structures are (i) time-series streams, (ii) spectrum and (iii) time-spectrum. By 'spectrum' we do not necessarily mean magnitude as a function of frequency, but instead use the term to refer to a data structure like magnitude spectrum (other examples of this data structure are the gamnitude cepstrum and the specific loudness pattern).

10.1. Time-Series

There can be a very large variety of single time-series output from the program, such as: sound pressure level, short term spectral and cepstral moments, Hilbert envelope, instantaneous frequency, auto-correlation function, loudness, sharpness, roughness, loudness fluctuation, total pitch strength, pitch multiplicity (the number of audible pitches), dissonance, inter-aural cross correlation peak value and lag time, and many more. PsySound3 provides a generic auditory graphing utility, allowing any of these to be mapped to the frequency of a tone, and/or the gain and panning of this signal. The mapping functions relating the data to the auditory graph parameters are user-controlled.

10.2. Spectrum

Examples of spectral data types include: the 1/3-octave band spectrum, octave band spectrum, fast Fourier transform spectrum, cepstrum, specific loudness pattern, specific roughness pattern, pitch pattern, and inter-aural cross-correlation function. Spectral data types can be sonified directly through re-synthesis. This usually involves a simplification of the sound, relative to the original, and involves a transformation of the sound when the spectrum comes from a psychoacoustical algorithm (such as the specific loudness pattern or pitch pattern). The point of this is to illustrate how the algorithm extracts or emphasizes various features from the original sound. However, it is difficult to hear much detail in this style of sonification – it is most useful in giving a general impression. Alternatively, spectral data may be sonified sequentially, using the same parameters as are available for time-series sonifications.

10.3. Time-Series Spectrum

This data type adds time as a dimension to the spectral data type, and is exemplified by the spectrogram, cepstrogram, specific loudness pattern over time, auto-correlogram and so on. At the time of writing, sonification of this data type remains to be implemented.

11. CONCLUSIONS

There is great potential to extend sonification further in the area of acoustic and audio analysis for education, and we hope that the examples presented in this paper will form just a small part of a more sophisticated collection. Sonification could be used for sound analysis in research as well as education, and PsySound3 provides a new platform for research-oriented analysis. We hope to continue to develop tools in this area, and invite others in the auditory display community to develop further sonification tools for education and research in sound.

12. ACKNOWLEDGMENTS

Parts of this project involve contributions from Robert Maria and Luis Miranda Jofre, who are masters students in audio and acoustics at the University of Sydney.

13. REFERENCES

[1] D. Cabrera, S. Ferguson and R. Maria, "Using sonification for teaching audio and acoustics," in *Proc. 1st Australasian Acoustical Societies' Conf.*, Christchurch, New Zealand, November 2006, pp. 383-390.

[2] D. Cabrera and S. Ferguson, "Auditory display of audio," in *Proc. 120th Audio Eng. Soc. Conv.*, Paris, France, 2006.

[3] F. E. Toole, "Loudspeakers and rooms for sound reproduction – a scientific review," *J. Audio. Eng. Soc.*, vol. 54, pp. 451-476, 2006.

[4] H. Kutruff, *Room Acoustics*, 4th edition, Spon, London, Great Britain, 2000.

[5] R. H. Bolt, "Note on normal frequency statistics for rectangular rooms," *J. Acoust. Soc. Am.*, vol. 18, pp. 130-133, 1946.

[6] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer-Verlag, Heidelberg, Germany, 1990.

[7] H. Møller, M. F. Sørensen, C. B., Jensen, and D. Hammershøi, "Binaural technique: do we need individual recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451-469, 1996.

[8] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues," *Applied Acoustics*, vol. 68, pp. 835-850, 2007.

[9] J. Epps, A. Dowd, J. Smith and J. Wolfe, "Real time measurements of the vocal tract resonances during speech," in *Proc. ESCA Eurospeech97*, Rhodes, Greece, pp. 721-724, 1997.

[10] D. Cabrera, S. Ferguson and E. Schubert, "'PsySound3': Software for acoustical and psychoacoustical analysis of sound recordings," in *Proc 13th International Conference on Auditory Display*, Montreal, Canada, June 2007.