Using sonification for teaching acoustics and audio

Densil Cabrera, Sam Ferguson and Robert Maria

Faculty of Architecture, University of Sydney, NSW 2006, Australia

ABSTRACT

In this paper we develop examples of how the understanding of acoustic and audio phenomena can be enhanced through sonification, especially with a view to application in education. The term sonification refers to the process of converting data into non-speech audio, and is distinct from auralization in that the process does not aim to simulate an actual or imagined sound environment. Measurements of audio and acoustical systems are most commonly represented numerically and graphically, and these two methods each have distinct advantages. However, display of such data using sound not only conveys important information, but also may provide an experience of important aspects of the phenomenon under consideration. When used in an education context, this method of data display should improve listening skills. We demonstrate various data transformations that allow a sonification of acoustical measurements or phenomena to bring out features of interest. We also demonstrate more abstract sonifications (auditory graphs) that can be usefully applied to this context.

INTRODUCTION

An understanding of acoustics and audio can come in many ways: for example through mathematical theory, verbal description, diagrams, animations, physical measurements, computer-based simulations and, of course, listening. Taking any single approach appears to be much less effective for learning than a combination of approaches. In this paper we outline some possibilities for a fairly uncommon approach to teaching acoustics and audio: sonification. 'Sonification' refers to the conveying of information to people through non-speech sound, and is a term that is widely used in the field of auditory display (as a counterpart to 'visualisation'). Auralisation, which is commonly used in architectural acoustics, is a different concept – it aims to present to a listener the sound that would be heard in a modelled environment. While that might be classed as a subset of sonification, in this paper we concentrate on various more abstract forms of data representation using sound.

We have previously presented some examples of sonification of audio system measurements (Cabrera and Ferguson 2006), and the present paper develops these examples further, and introduces new examples related to architectural acoustics and audio signal analysis. In the present paper, we describe sonifications related to architectural acoustics coefficients, room acoustical parameters, moments and percentiles of acoustic signals, room impulse responses, psychoacoustical parameters, and signal manipulations based on digital signal processing. The sonifications described in this paper can be categorised as follows: (i) they present auditory graphs of acoustic data; (ii) they convert statistical reductions of the acoustic signal into perceptually relevant sound; or (iii) they emphasise important details of an acoustic signal through digital signal processing. We have developed these partly to assist in teaching in the graduate program in audio and acoustics at the University of Sydney. However, a second reason for their development is that we teach sonification for undergraduate and post-graduate digital media courses, and so these sonifications also provide examples for that area of teaching.

ARCHITECTURAL ACOUSTICS COEFFICIENTS

We begin by describing the sonification of coefficients through auditory graphing. An auditory graph is an aural analogue of a visual chart – essentially a collection or series of values are presented to the ear, using sound parameters rather than visual parameters such as the height of bars on a bar chart. While one application of auditory graphing is in spreadsheet software for the visually impaired, in this case the rationale for auditory graphing is that the meaning of the phenomenon represented (which is the proportion of sound reflected or transmitted) is conveyed to the ears directly, rather than indirectly through numbers or via a visual chart.

Architectural acoustics makes extensive use of absorption and transmission coefficients. In room acoustics, absorption coefficients (α) are routinely used for design. However, we prefer to sonify $1-\alpha$ (the reflection coefficient, β) because the amount of sound experienced in the sonification then represents 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. In one implementation, the level of each octave band is controlled by $10\log(\beta)$ – 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. Pink noise within the octave band limits may be played immediately before the sonification of the absorption spectrum to provide an aural reference. A second implementation provides redundant encoding of the absorption data, by assigning a duration for each octave band equal to the reflection coefficient, in seconds. Hence an absorption coefficient of 0.1 yields a 0.9 s duration, whereas an absorption coefficient of 0.9 yields a 0.1 s duration. While the second implementation is more abstract, it presents the data in a more robust way because the sensitivity of the auditory system across the frequency range is not simple, especially considering simultaneous masking effects. One limitation is that for small reflection coefficients ($\beta<0.2$), the short duration of the synthesised signal (<200 ms) is likely to affect its loudness due to the integration time of the auditory system.

We sonify transmission coefficients similarly, except that data are presented in 1/3-octave bands, and the temporal representation is not implemented because of the very large range of values encountered. There is less need to have redundant temporal representation for the same reason – the range of levels is much larger than for absorption coefficient sonification, so the information is conveyed quite well without duration coding. The transmission coefficient is displayed directly – with 1/3-octave band level controlled by the sound reduction index multiplied by -1 (of course, a substantial gain offset is used to translate this to sound pressure level). In this sonification, the user has the option of weighting the spectrum (instead of using pink noise) either based on the R_w curve, the C spectrum adaptation term (A-weighted pink noise), or the C_{tr} spectrum adaptation term (which emphasises lower frequencies).

Auralisation of coefficient spectra may be done by filtering speech rather than noise signals.

ROOM ACOUSTICS

Reverberation time

Like architectural acoustics coefficients, reverberation time is a good candidate for auditory graphing. 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 simply by synthesising a pure tone at each octave band centre frequency, which is given 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. We have enhanced the audibility of these offsets 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. Usually we play the auditory graph 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 time stated in a reverberation time, because a 60 or 70 dB decay in the mid and low frequency range is not usually audible 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.

Room modal distribution

For small rooms, many have argued that a clustering of room modes in the low frequency range can lower the acoustic quality of the room. This is particularly evident for rooms that are to be used for audio systems (eg sound studios) because a flat frequency response is desired for the system within the room. Various criteria have been proposed for assessing low frequency modal distribution, especially for rectangular rooms (which many rooms approximate), and the most famous of these criteria is the blob diagram of Bolt (1946). Whether or not such criteria are of critical importance is a complex issue, discussed recently by Toole (2006). Nevertheless, it is important that students of room acoustics understand concepts that have been influential historically. In this section we consider two sonifications of room modal distribution in rectangular rooms, which allow modal distribution to be assessed by ear.

In the first sonification we synthesise a sequence of complex tones, the fundamentals of which correspond to the room mode frequencies. Although the low fundamentals might not be audible on their own (considering that they might be in the 20-50 Hz frequency range), the complex tone spectrum renders them audible as virtual pitches. The sonification consists of an ascending sequence of these complex tones, where each is played for 100 ms. It is very easy to hear the difference between evenly and unevenly distributed room modes, with the latter represented by a melody that lingers on certain pitches with sudden jumps.

We have also implemented a more abstract and interactive sonification of room modal distribution, in terms of room proportions. Inspired by Bolt's room proportions diagram, the sonification produces a triad of harmonic tones with fundamental frequencies proportional to the three rectangular room dimensions. The harmonics of each tone correspond to the frequencies of each axial mode. While tangential and oblique modes are not included in the sonification, the axial modes are in some ways the most important because of their long mean free paths compared to tangential and oblique modes (leading to resonances characterised by low damping or high Q, notwithstanding effects of uneven absorption distribution). This sonification is implemented interactively – the user enters the room proportions and a corresponding triad is produced. The degree to which the triad is dissonant provides a crude representation of the extent to which room modes are evenly distributed. A consonant triad represents uneven mode distribution, which is generally regarded as undesirable. This sonification is attractive because of its simplicity, but may be limited because of the likely influence of musical training in interpreting the sound. Furthermore, associating dissonance with the desirable result, and consonance with the undesirable result, is an inverse coding. These issues may be addressed in future refinements of this sonification.

ROOM IMPULSE RESPONSES

Simply playing a room impulse response might be regarded as a nascent form of sonification (the term 'audification' is sometimes used to refer to the process of sonifying time-series data with very little or no transformation). So in this section we consider how a room impulse response can be treated further for sonification, so that meaningful features are more apparent to the ear. Issues include temporal masking effects, spectral weighting, and audibility of spectral detail. While the processes described here can be applied to single channel room impulse responses, more satisfying results (in terms of sonification) are achieved through treating binaural room impulse responses, which can be listened to on headphones, giving a better impression of the acoustic space represented.

Fine temporal features

The early reflection sequence in a room impulse response is of substantial interest in understanding room acoustical quality. However, playing an untreated room impulse response presents this sequence so quickly that it is difficult to catch much by ear. Sonification is better achieved by slowing down the playback of the waveform, and/or reversing it. Slowing the waveform by a factor of 16 produces a sound that is slow enough for the ear to follow, with important frequency content still within the audible frequency range (eg 8 kHz is shifted to 500 Hz). There are other possibilities for slowing sound waves without changing pitch (eg via the Hilbert trans-

form or short time Fourier transform), but we have found that the simpler approach of merely slowing the waveform provides a compelling sonification.

Time-reversing a room impulse response is also very helpful for hearing fine temporal features. This is partly because of temporal masking asymmetry – meaning that the direct sound and prominent echoes tend to provide substantial masking for features succeeding them, but not so much for features preceding them. Since a room impulse response follows a roughly exponential decay, reversing it provides a large benefit in lowering the masked threshold in relation to the peaks (which represent acoustic reflections in the room). Even if no speed reduction is applied, the difference is striking, with early reflections that were scarcely audible in the forward play forming a clearly audible rough texture when the room impulse response is played in reverse. Another reason for the effectiveness of time reversal is that sounds that increase in loudness have a greater perceptual salience than those that decrease – an effect sometimes referred to as 'looming' (Neuhoff, 1998). This effect is probably because of a learnt association with an approaching, versus a receding, sound source (in everyday life, an approaching sound source could signal danger, or at least require a response). Mirroring the time-reversed and original impulse responses (producing a <> envelope) effectively combines the advantages of both modes of presentation: aural analysis of the reflection sequence in reverse, and hearing the sound of the decay forward presentation.

Further aural analysis is facilitated through fragmentation of the room impulse response. Clarity index is an energy ratio measure used in room acoustics, comparing the first 50 or 80 ms of an impulse response with the remainder. We sonify this concept by inserting a 300 ms silence between the early and late portions of the impulse response. We also optionally split the impulse response into three sections – the direct sound (which might be taken as the 0-6 ms period from the first arrival), the early reflections (6-50 ms) and the late reverberation (after 50 ms). In education, this at least provides a demonstration of key sections of an impulse response (including the notion of clarity index), and may also be helpful in hearing their relevant features.

The modulation transfer function (MTF), which is used for the determination of speech transmission index (STI), can form the basis of a room impulse response sonification. A sequence of spectrally weighted clicks accelerating gradually from 0.63 Hz to 12.5 Hz may be convolved with a room impulse response so as to sonify the MTF concept, allowing it to be assessed by ear. Clicks are chosen rather than the sinusoidally modulated noise that forms the basis of true MTF measurements because the silences between clicks are longer, providing greater contrast for the ear. Using this approach, it is not practical to sonify the full 98-value MTF matrix used for STI calculation (14 discrete modulation frequencies and 7 carrier signals consisting of ½-octave bands of noise centred on octave-related frequencies from 125 Hz to 8 kHz) because this would take too long. Instead we simply use broadband spectrally weighted clicks, meaning that differences between MTFs for different carrier frequencies is not presented by the sonification. The benefits of this sonification method include: (i) giving an aural explanation and experience of modulation transfer function, which is a key concept in STI and speech intelligibility assessment more generally; (ii) providing a simple way of listening to the idea behind clarity index, even though the procedure for calculating these is somewhat different; (iii) making periodicities in an IR's temporal envelope easy to hear.

Coarse spectral features

A problem with direct auditory display of a room impulse response, is that room acoustical measurements are done in octave or 1/3-octave bands (and hence use a 'pink' information distribution), whereas the impulse response has a 'white' energy distribution. Furthermore, the general purpose of room impulse response measurements is usually for the assessment of speech intelligibility or music quality, and neither speech nor musical sound has a white energy distribution. The effect of this is that the high frequency range has undue prominence. A -3 dB per octave filter can give an impulse response a spectral distribution that would have been caused by a 'pink' sound source. Furthermore, using appropriate filtering can allow the listener to understand better the part of the impulse response that is relevant for their application. For instance, for those interested in the speech spectrum, as is the case in many auditorium applications, a filter with a shape similar to a representative speech spectrum (such as that described by IEC Standard 60268:16) may allow a more appropriate understanding of the impulse response. Of course, another useful method for conveying the effect of a room impulse response on speech is to auralise it, by convolving it with a speech sample.

Fine spectral features

Hearing the fine spectral features of a room impulse response can be difficult because temporal features dominate perception when an impulse response is simply played. A simple solution to this is to sonify the steady state response of the room by convolving the impulse response with steady state noise (eg pink noise) or a swept sinusoid. This makes the magnitude transfer function easier to hear, especially when a number of impulse responses are being compared with each other. The swept sinusoid presents the spectrum in series, rather than in parallel, allowing for easier comprehension but probably requiring a longer duration signal than noise. Further emphasis can be given to the fine spectral features by raising the spectrum to some power, which can also be achieved through auto-convolution. A single auto-convolution is equivalent to squaring the complex spectrum, and successive auto-convolutions are equivalent to raising the spectrum to higher powers. The effect is that fine spectral features are exaggerated, making peaks in the transfer function much easier to identify by ear. We sonify room impulse responses in this way by implementing an auto-convolution sequence of the impulse response, where the result of each auto-convolution is itself auto-convolved, corresponding to spectral powers of 2, 4, 8, 16, etc. Ultimately the result will be a pure tone (corresponding to the greatest peak) but the sound in the stages between the original impulse response and the pure tone state can be very revealing to the ear.

A similar sequence for enhancing spectral contrast is obtained through auto-correlation. In fact the only difference between auto-correlation and auto-convolution is a time reversal of one of the input functions. The key difference in the resulting auto-correlation functions is that they are symmetric in time (referred to as 'even' functions), meaning that apart from the first order auto-correlation, there is no difference between the processes of auto-correlation and auto-convolution. The same spectral effect, of raising the magnitude spectrum to a power, corresponding to the auto-correlation order, occurs with this sequence (however, a difference is that the phase of the spectrum is linearised). The result is a temporal envelope that increases to a peak and then decreases, like a very long linear phase filter.

While peaks in magnitude spectra may be heard relatively easily, for the most part, spectral dips are difficult to discern

by ear. One way of making these audible is to simply 'invert' the magnitude spectrum. We do this by mathematical inversion ($1/x$) of magnitude values, followed by rescaling to the original total (rms) magnitude. The resulting inverted magnitude spectrum may also be enhanced for sonification, if required, by raising it to some power, as described previously.

Acoustic Quality Test (AQT)

Audio system measurement is an important part of the graduate program in audio and acoustics at the University of Sydney. One technique that we are currently teaching is the Acoustic Quality Test (AQT), which was proposed by Farina *et al.* (2001) for the assessment of audio systems in car interiors and small rooms. This measures the frequency/time response of a system to 200 ms sine tones (or wavelets), including the system's decay in the subsequent 33 ms or 66 ms (depending on the analysis context – 33 ms is appropriate for very small rooms like car cabins, and 66 ms for somewhat larger rooms). Reasons for this type of measurement include the auditory system's loudness integration time (of about 200 ms), the fact that most program material for audio systems (speech and music) is far from steady state, and that the direct and very early reflections are most influential for the perceptually relevant frequency response. The AQT frequency response can deviate from the steady state frequency response, especially when reflections and resonances are strong. The dynamic transient capability of a system, as defined by AQT, is the level decay in the 33 ms (or 66 ms) following a 200 ms excitation. Due to temporal masking, the maximum perceptible level difference over the 33 ms decay is about 20 dB.

The AQT spectrum, including its dynamic capability, are sonified simply through an ascending stepped sine sweep that alternates between the AQT peak level and decay level for successive frequencies. Each tone has a 100 ms duration (i.e. 100 ms for the peak tone, then 100 ms for the decay level tone, followed by the same for the next frequency). Either the 33 ms or 66 ms decay level can be chosen for this sonification.

Non-linear distortion

A method for simultaneously measuring the linear impulse response and non-linear harmonic distortion products of weakly distorting systems is given by Farina (2000). A sinusoidal sweep with a logarithmic frequency distribution is used as the excitation signal. The impulse response is obtained by cross-correlating the received signal with the original, with a +6 dB/octave compensation for the pink spectral distribution of the excitation signal. The resulting impulse response is preceded by a descending series of pseudo-impulse responses of harmonic distortion products (descending from high order harmonics, through to the 2nd harmonic – the true impulse response is, of course, the first harmonic).

We sonify this by reversing the entire sequence (so that the true impulse response is heard first, but reversed, followed by the harmonic distortion responses in ascending order). Generally, we boost the distortion products by 10 dB, because even very small amounts of distortion may be of significance for some audio systems. Since harmonic distortion depends, in a non-linear manner, on the amplitude of the system's input, we would normally have three or four measurements of a given system for different input levels. For a student, it is useful to hear both the original test swept tone (where distortion is heard as timbre) and the decomposed signal as described here (where the distortion spectrum is translated to time).

SPECTRAL MOMENTS

The spectral centroid is a single frequency representing the 'centre of gravity' (essentially the mean frequency or first moment) of a power spectrum. As such, it can be sonified as a pure tone at that frequency. For time-varying signals, the spectral centroid will vary, making an interesting application for auditory graphing. Applications can include room impulse responses, speech recordings, music recordings, musical instrument tones (for timbre analysis) and many other types of sound. For sonification, we control the level of the synthesised tone representing the centroid using the level of the source spectrum, 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. We play the source recording and spectral centroid together (in separate audio channels), so that a connection is made between the two in listening.

The centroid calculation for a discrete spectrum is given below, where f_n is the frequency of each spectral component, and a_n is its magnitude.

$$C = \frac{\sum_{1}^{\max n} f_n a_n^2}{\sum_{1}^{\max n} a_n^2} \text{ Hz} \qquad (1)$$

Alternatively this (and other spectral moments) can be calculated for the unsquared magnitude spectrum, yielding a different result. Our implementation allows either approach to be taken, so that students can explore and hear this difference, along with other approaches outlined later.

Spectral width (or spread around the centroid) is a statistic that can sometimes be of interest in simplified spectral analysis, although it can be calculated in several ways. Spectral width is taken as the variance (the second moment) of the magnitude spectrum, and is sonified as band limited noise covering this range, with steep pass-band limits at the frequencies one square root of the standard deviation apart (being the mathematical definition of variance). Skewness is the third moment of a distribution, and is sonified by shifting the noise band up or down in frequency corresponding to the skew. We have not attempted to sonify the fourth moment (kurtosis).

Different results for spectral centroid and spectral width are obtained depending on the frequency units and weighting used. Possibilities include using a linear or logarithmic spectral component distribution, or using auditory filter units (Barks or Erbs). Results also depend on the 'magnitude' units used (pressure, power, level, or psychoacoustical units such as excitation, excitation level, or specific loudness). However, level units (i.e. decibels) are less meaningful in these calculations than ratio scale units (for which zero means no sound). In psychoacoustics, models of sharpness are based on the centroid of the specific loudness pattern.

As measures of a spectrum, spectral centroid and spectral width are gross simplifications, which can be either advantageous or disadvantageous, depending on the application.

STATISTICAL SOUND LEVELS

Percentiles provide a succinct numeric summary of the distributions of time varying levels. However, it may sometimes be difficult for students to perceive a connection between the sound that they hear and the percentile levels (which are of-

ten referred to as statistical levels). Therefore sonification of these levels can helpful. Our approach to this has been to determine the percentile levels in 1/3-octave bands (using 'fast' integration), and to use these to control the level of simultaneously sounded 1/3-octave bands of noise. In implementation we use deciles (i.e. every 10^{th} percentile), presented in the 11 element sequence L_{min}, L_{90}, L_{80} … L_{20}, L_{10} and L_{max}, with 0.5 s duration for each (except for the median, L_{50}, which has a 1 s duration so that it can be identified easily). A 50 ms fade-in and fade-out is used to separate the levels, which is helpful when there are only small differences between them.

Since the decile levels in each 1/3-octave band do not necessarily reflect similar distributions, the spectral characteristics of the sonified signal may change substantially as the sequence moves from L_{min} to L_{max}. This provides a substantially richer appreciation of the sound than is available through A-weighted broadband levels.

Moments and percentiles are two approaches to the same problem (i.e. representing the statistical distribution of a dataset) and the application of one approach to spectrum and the other to time is simply due to convention. We may develop a more flexible sonification in future allowing students to explore sound recordings by selecting the analysis method for a given dimension.

PSYCHOACOUSTICAL PARAMETERS

The concept of loudness level can be helpful in relating the psychological attribute of loudness to the familiar decibel unit. Loudness level, in phons, is the sound pressure level of a 1 kHz pure tone that has the same loudness as the sound being assessed. One possibility for a sonification to help understand loudness metrics might be to simply synthesise the 1 kHz tone that corresponds to the loudness of the arbitrary sound under consideration, so that a listener can compare them aurally, as well as comparing the loudness of the tones corresponding to a number of assessed sounds. However, pure tones are less than ideal signals for sonification, especially if the loudness of the tone is its key parameter, because audio system and room acoustical transfer functions (and indeed the auditory system) may have fine irregularities that are impractical to control or compensate for. Nevertheless, we have drawn on this general concept in our partially implemented sonification of the combined psychoacoustical parameters of pitch height, loudness, sharpness, roughness and fluctuation strength: the idea is to synthesise a parametrically defined signal that has the same psychoacoustical scale values as the arbitrary sound being analysed. Rather than using a pure tone, we synthesise a harmonic tone, allowing sharpness and pitch to be varied independently (pitch is primarily controlled through fundamental frequency, and sharpness is primarily controlled through the energy of the harmonics, including compensation for fundamental frequency). Loudness is primarily controlled through gain. We also use frequency modulation depth as the primary control for roughness (fixed modulation frequency of 70 Hz), and amplitude modulation depth as the primary control for fluctuation strength (fixed modulation frequency of 4 Hz). We refer to 'primary control' because all of the psychoacoustical parameters will be affected, at least to a small extent, by all of the physical signal parameters, and demonstrating this point is one of the aims of the sonification.

The underlying concepts and predictive algorithms for these psychoacoustical parameters are summarised by Wicker and Fast (1999), and models for these parameters are commonly implemented in sound quality measurement software. However, while it is possible to measure them using such software, the algorithms are not reversible. Therefore, our synthesis method uses a large matrix of modulated complex tone signal parameter sets for which psychoacoustical parameter values have already been measured. Given a set of psychoacoustical parameter values, an appropriate candidate can be selected from this matrix for synthesis. More details on the methods used for this type of auditory display are given by Ferguson *et al.* (2006). This sonification method is rather limited, because it is only possible for sounds having all of their psychoacoustical scale values within the matrix. There are practical limits to the matrix (eg very high sound pressure levels are not used) as well as inherent limits (eg a high pitched tone cannot have low sharpness). In practice our sonification is only effective for moderate values of the psychoacoustical parameters. As a demonstration for educational purposes, such limitations are not a serious problem.

For analysis of the arbitrary signal, the pitch algorithm of Terhardt *et al.* (1982) is used, which predicts virtual pitches using a template-matching procedure, and also predicts pitch shifts due to level and masking. We simply select the strongest pitch of the analysis. However, the pitch algorithm can be disabled, which is appropriate for unpitched sounds – in which case a default fundamental frequency of 220 Hz is synthesised.

We use the time-varying loudness algorithm of Glasberg and Moore (2002), and hope also to implement that of Chalupper and Fastl (2002) for comparison. For time-varying signals, we assume that the overall loudness is represented by the 90^{th} percentile (N_{10}), as suggested by Zwicker and Fastl (1999). Sharpness is calculated according to the algorithm of Zwicker and Fastl (1999) – which is a weighted centroid of the specific loudness pattern.

This sonification is only partially implemented at the time of writing, with fluctuation strength and roughness still to be implemented (these algorithms are the most complex of the psychoacoustical parameters). We will use the roughness algorithm of Daniel and Weber (1997), and the fluctuation strength algorithm of Chalupper (Chalupper and Fastl 2002, and personal communication).

The audio system characteristics for this sonification are very important for its success. At least, the amplitude response of the system as a function of frequency must be flat over a fairly wide range (eg 63 Hz – 8 kHz) and the gain must be in rough calibration, and preferably in accurate calibration.

COMPLEX SPECTRUM

The magnitude frequency response of a system may be sonified simply through a swept sinusoidal signal. As discussed earlier, spectral features may be emphasised by raising the spectrum to some power. This simple approach to spectral representation can be extended to include the complex spectrum, which consists of magnitude and phase, or alternatively real and imaginary vectors.

The difficulty with representing phase is that the auditory system is quite insensitive to the phase of individual tones, so that phase cannot be used to sonify phase (of course, phase differences can be sonified through the combination of pure tones, but that confounds magnitude with phase). We use a spatial transform to sonify phase, because the location of the auditory image can be varied relatively independently of the signal level (which is used to represent magnitude). In the simplest approach, wrapped phase is represented from left to right (corresponding to 0 to 2π) either in headphone reproduction or a stereophonic loudspeaker system. Level based panning is used (rather than time-based panning) because it

provides more precise image localisation, especially in the high frequency range. One problem with this is that there is a sudden jump in image location as the wrapped phase shifts from 2π to 0. More ambitiously, phase could be better encoded by mapping it directly to angles around the listener in the horizontal plane, although this is technically demanding to implement. Since panning from front to back is problematic, several loudspeakers (eg 8) encircling the listener might be required for a convincing loudspeaker implementation. However, the localization of pure tones is difficult for some frequencies, and indeed for all frequencies in reverberant rooms, so such an implementation may be somewhat impractical. Headphone implementation using head-tracking might be more successful because of the exclusion of room acoustical effects. In either system, exploring the listening environment with small head movements would substantially improve localisation. Another approach to improving localisation is to represent the magnitude spectrum using a harmonic tone sweep, possibly with some additional high frequency noise, so that pinna-related spectral cues can be used for front-back discrimination. In such a display, the combination of magnitude and phase might be thought of as a fully spatial auditory display, because magnitude is associated with source distance due to the dispersion of the sound waves from sources (eg a point source). In fact, the inverse square law (-6 dB per doubling of distance) corresponds simply to the spectrum (where -6 dB represents a halving in magnitude). If, instead, the magnitude spectrum is squared, then the inverse square law maps simply to the power spectrum.

A very simple transformation of frequency response is to read it (and so sonify it) as a time series. The result is something that might be compared to the cepstrum – the listener hears periodicities in the frequency response transformed into audible tones (i.e. this sonification uses the auditory system for frequency analysis). Such periodicities reflect the presence of harmonically related components. In our implementation of this, we sonify the real response on one channel, and the imaginary response on the other. To do this usefully, a rather large window length must be used for the FFT (otherwise the playback duration is very brief). However, if the spectrum is mirrored beyond the Nyquist frequency, then the playback duration may be extended indefinitely. If the sampling rate of the sonification is the same as the original wave's sampling rate, then the resulting frequencies match the actual fundamental frequencies of harmonic series in the original wave.

Effects of phase transformations can provide interesting demonstrations when sonified – including the Hilbert transform discussed in the next section. More simply a FFT can be taken, and the phase manipulated by setting it to a single value (eg 0 radians) or to a random value prior to resynthesis. For FFTs with long window lengths (eg several seconds) this provides a simple demonstration of the importance of phase in structuring a wave – a recording of speech may become completely unintelligible using this approach, despite having the same power spectrum.

APPLICATIONS OF THE HILBERT TRANSFORM

The Hilbert transform offers some interesting possibilities for waveform analysis, and indeed sonification. A Hilbert transform is performed simply by phase shifting all components (in the frequency domain) by -π/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. Combining these in a frequency- and amplitude-modulated sinusoid can reconstruct the original signal, notwithstanding some exceptions.

One application of this transform is envelope extraction. The instantaneous amplitude defines the envelope function. A simple sonification of a Hilbert envelope function is achieved by multiplying it with a carrier signal, which could be steady state noise, or indeed a steady state spectrum derived from the input signal. In another sonification using the Hilbert transform-derived amplitude envelope, we aim to emphasise change in a waveform (and so de-emphasise its steady state parts). To do this we take the derivative of the amplitude envelope prior to resynthesis (in practice, differentiation is achieved through differencing the sampled waveform). The effect of this is to silence parts of the waveform possessing constant amplitude, bringing to the fore parts in which the envelope is changing rapidly. This effect can be strengthened by raising the derivative to some power (although steady state portions are silenced regardless of the exponent used). Another approach that we have taken to sonifying the envelope is to take the logarithm of the envelope function and differentiate that. Taking the logarithm of the envelope function is particularly useful for waveforms exhibiting exponential decay (eg recordings made in reverberant conditions) – after differentiating, exponential decays are silenced, leaving other features. In these examples the resulting waveform is normalised prior to playing for sonification.

Periodicity in the envelope function can be an important acoustical feature, which might be heard as rhythm. Hence, a rhythm-to-pitch transform is possible through speeding up the amplitude envelope function, and playing it as an audio waveform (following resampling and normalisation). We tend to use a factor of 100 (or sometimes 128) for this, since these are easy ratios for a listener to understand, and they place the fluctuation frequencies that are of greatest perceptual salience (around 4 Hz) into the frequency range of greatest pitch sensitivity (Zwicker and Fastl 1999). The result of this transformation is that periodicity within the original envelope forms harmonically-related tones, meaning that regular rhythms form strong and consonant pitches, whereas an irregular envelope forms a more complex noise-like sound.

The converse of sonifying the amplitude envelope independently of the instantaneous frequency is to sonify instantaneous frequency independently of the envelope. This is done most simply with a constant amplitude envelope. Among the other manipulations that we have experimented with are combining the instantaneous amplitudes of one waveform with the instantaneous frequencies of another; and resynthesis for which instantaneous amplitude and instantaneous frequency are exchanged. Such manipulations allow a listener to hear features of the analysed waveform in different ways. A related example of Hilbert transform application is described by Smith *et al.* (2002), who adapted this technique for research into the perception of speech's envelope versus its fine spectral features.

Two refinements of the abovementioned procedures can sometimes be helpful in enhancing the meaningfulness of analysis and sonification. One is to smooth (or low-pass filter) the instantaneous amplitude and/or frequency functions. Another is to take a multi-band approach to envelope analysis and resynthesis, perhaps based on the auditory filter bandwidths (as was done by Smith *et al.*).

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 centre 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 (eg front-back and above-below) of the sound source. Reproduction of appropriate binaural difference and spectral cues can produce a convincing externalised and accurately localised 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 localisation. 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).

We illustrate this concept through sonification over headphones, firstly by presenting a series of white noise samples filtered through various people's hrtfs for 0 degrees azimuth and polar angle. The result of hearing these is a series of shifting image locations in the median plane, which can also be heard as spectral changes. Listening to the same sonification through loudspeakers makes the spectral changes much more obvious (and the spatial changes much weaker). Another useful way to listen to these hrtfs is to down sample the filters by one or two octaves – this makes the peaks and notches more easily identifiable by ear, removing any vestige of localisation from the sensation.

In a second sonification of this concept, 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*. (2006). 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 (eg 0 degrees) by manipulating these. We supplement this by allowing the user to change the inter-aural time difference (±1 ms) and broadband inter-aural level difference (±10 dB).

These sonifications may be somewhat limited using normal headphones (circum-aural or supra-aural), because the external ear is used in the reproduction. To some extent this can be ameliorated by the application of a generic inverse filter for the selected headphones.

AUDIO AND AUDITORY CONSIDERATIONS

Sonification is conveyed to a listener using an audio system, such as a computer with headphones or loudspeakers. One obvious limitation of such systems is that their response will not be perfect. The phase response is usually not a problem in terms of introducing auditory artefacts, but the amplitude response as a function of frequency can vary appreciably between playback systems (loudspeakers and headphones). Furthermore, the time response of loudspeaker systems in rooms would normally be smeared by early reflections and reverberation. The gain of playback systems is normally not controlled well, which has significance at least for psychoacoustical model sonifications – for which playback should be at least in rough calibration. Other limitations include the quality of spatial rendering, and the presence of non-linear distortion. Depending on the purpose of the sonification, these limitations can be managed through sensible choice of the reproduction system, and possibly through compensating for the response of a system (eg through inverse filtering).

Another issue with sonification is the response of the auditory system. An argument could be made that it is necessary to compensate for this – for example by making spectral adjustments in relation to equal loudness contours. At present, we are not making any such adjustments for our sonification examples, except for calibrating the playback gain for psychoacoustical sonification.

Sonification is only effective between the masked threshold (due, for example, to background noise) and the maximum comfortable (and distortion-free and safe) playback level. In some listening environments, this would give as little as 20 dB range between minimum and maximum effective levels. This contrasts with the subject of some of the sonifications described here, for which important features may have several tens of decibels contrast. Hence the presentation context may affect the effectiveness of such sonifications, and it may be that further treatment is appropriate for adverse listening contexts (which can include classroom teaching).

IMPLEMENTATION

At the time of writing, these sonifications are in various stages of implementation. In some cases they are fixed demonstrations, developed by hand. In other cases we have implemented them as small computer programs, and we are currently working to expand this implementation. In many cases, the software implementation is done through Max/MSP, which is a graphical programming environment designed primarily for real time digital audio processing with easy development of intuitive human-computer interfaces. Max/MSP is widely used for teaching in the field of audio, although not so much in the field of acoustics. For students it is much more approachable than Matlab as an analysis tool.

We plan to make these sonifications and associated documentation freely available in the next 6 to 12 months.

CONCLUSIONS

This paper describes some possibilities for the sonification of data related to technical audio and acoustics. Our work in this area is continuing, and we plan to make it broadly available as the work matures. Sonification provides a way of understanding audio and acoustics through experience, and so can complement other forms of learning in education. One advantage of sonification of sound phenomena is that the representation is experienced in the same sensory mode as the phenomenon – and so sonification should be helpful in developing an ear for audio and acoustics.

REFERENCES

Bolt, R. H. 1946, "Note on normal frequency statistics for rectangular rooms," *Journal of the Acoustical Society of America*, 18, 130-133

Cabrera, D. and Ferguson, S. 2006, "Auditory display of audio," *Proceedings of the 120th Audio Engineering Society Convention*, Paris, France

Chalupper, J. and Fastl, H. 2002, "Dynamic loudness model (DLM) for normal and hearing-impaired listeners," *Acta Acustica united with Acustica*, 88, 378-386(9)

Daniel P. and Weber, R. 1997, "Psychoacoustical roughness: implementation of an optimized model," *Acta Acustica united with Acustica*, 83, 113–123

Farina, A. 2000, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," *Proceedings of the 108th Audio Engineering Society Convention*, Paris, France

Farina, A., Cibelli, G. and Bellini, A. 2001, "AQT – a new objective measurement of the acoustical quality of sound reproduction in small compartments," *Proceedings of the 110th Audio Engineering Society Convention*, Amsterdam, The Netherlands

Ferguson, S., Cabrera, D., Beilharz, K. and Song, H. J. 2006, "Using psychoacoustical models for data sonification," *Proceedings of the 12th International Conference on Auditory Display*, London, UK

Glasberg, B. R. and Moore, B. C. J. 2002, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, 50, 331–342

Iida, K., Itoh, M., Itagaki, A. and Morimoto, M. 2006, "A novel head-related transfer function model based on spectral and interaural difference cues," *Proceedings of the 9th Western Pacific Acoustics Conference*, Seoul, Korea

International Electrotechnical Commission (IEC) 1998, Sound System Equipment – Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index, IEC:60268-16, Geneva, Switzerland

Neuhoff, J.G. 1998, "Perceptual bias for rising tones," *Nature*, 395, 123-124

Smith, Z. M., Delgutte, B. and Oxenham, A. J. 2002, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature* 416, 87-90

Terhardt, E. Stoll, G. and Seewan, M. 1982, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *Journal of the Acoustical Society of America*, 71, 679–688

Toole, F. E. 2006, "Loudspeakers and rooms for sound reproduction – a scientific review," *Journal of the Audio Engineering Society*, 54, 451-476

Zwicker, E. and Fastl, H. 1999, *Psychoacoustics: Facts and Models*, Springer, Berlin, Germany