Acoustics 2008
Geelong, Victoria, Australia 24 to 26 November 2008

Acoustics and Sustainability:

How should acoustics adapt to meet future demands?

PsySound3: An integrated environment for the analysis of sound recordings

Densil Cabrera (1), Sam Ferguson (1) and Emery Schubert (2)

(1) Faculty of Architecture, Design and Planning, University of Sydney, NSW 2006, Australia
(2) Empirical Musicology Lab, School of Music and Music Education, University of New South Wales, NSW 2052, Australia

ABSTRACT

This paper presents possibilities offered by a computer program for analysing features of sound recordings, PsySound3. A wide variety of spectral and sound level analysis methods are implemented, together with models of loudness, roughness, pitch and binaural spatial analysis. In addition to providing access to these analysis methods, this analysis environment provides a context for easy comparison between analysis methods, which is very useful both for teaching and for the testing and development of models for research applications. The paper shows some of the potential for this by way of example. The software is structured so as to be easily extensible (using the Matlab programming environment), and many extensions are envisaged. Written by the authors and colleagues, PsySound3 is freely available via www.psysound.org.

INTRODUCTION

In the Australian Acoustical Society conference of 1999, the first author presented a paper on a computer program known as PsySound, which he wrote to provide analysis capability using several psychoacoustical and acoustical methods (Cabrera 1999).The present paper describes a newly written software environment for the analysis of sound recordings, providing similar but much more extensive capabilities. PsySound3 was written by the authors and colleagues to provide free and flexible access to a wide range of sound analysis methods with an emphasis on psychoacoustical algorithms, such as loudness related parameters, pitch parameters, auditory spatial parameters, parameters related to music perception, and other aspects of sound quality. In addition to implementations of psychoacoustical analysis methods, PsySound3 provides a range of conventional analysis methods, such as spectrum and cepstrum analysis, autocorrelation, Hilbert transform, and a sound level meter emulator. Integrating many analysis methods into a single software environment facilitates comparison between analysis methods for research or education purposes. This paper provides some examples of comparative analysis performed by the program. The software environment is extensible, and we envisage that more analysis methods will be contributed to the environment as the project develops.

PROGRAM STRUCTURE

PsySound3 is implemented using Matlab (with the Signal Processing toolbox). We envisage releasing compiled versions once the program reaches a more mature state, but in its present form using PsySound3 requires very little knowledge of Matlab because a user interface is provided by PsySound3. Easy extensibility of the program comes from its implementation in this commonly used programming environment (with code arranged in a modular hierarchical file structure), and extensions can be shared through participation in the development group, which uses a central code repository accessible via the internet.

The program is designed to analyse sound files in common formats such as Microsoft wav. Files may be calibrated or gain adjusted in several ways, and the program includes the facility for using recordings of microphone calibration signals. Large groups of files may be analysed (although the analysis may take quite some time, and so is best done on a dedicated computer). The program includes a set of audio analysers, which are written as independent modules, each taking advantage the program infrastructure (graphical user interface, calibration, data format, etc.). Since these analysers are probably the major point of interest to acousticians, this paper focuses on them.

In some cases existing analysers are based around pre-existing code, which has been 'wrapped' with some additional features and inserted into the program directory structure. Having many analysis methods driven by the one program has the advantage of being able to coordinate and, to an extent, automate analyses in which the outputs of multiple analysis methods are compared. The analysers mentioned in this section are examples of what can be done, but there is the potential for many other analysers to be added in the future.

Many analysers work by dividing the soundfile into successive (or overlapping) frames, from which spectral patterns or single-number parameters are derived. Having stepped through the entire sound file, the main analyser output formats are time-series objects (showing how a single parameter varies over time), spectrum objects (showing how a parameter is distributed across a non-time dimension such as frequency) and time-spectrum objects (such as a spectrogram). Some analysers can yield an output rate equal to the audio sampling rate of the file being analysed, potentially leading to a substantial increase in the size of the analysis compared to the file size of the input wave. However, the program allows the output of multiple analysers to be 'synchronised', meaning that high output rate data are downsampled, low output rate data are upsampled (although in practice that is rare), and the step size of discrete analysis frames is set to a given value. Synchronised output data allow for easy comparison between time series and may avoid excessive data density for a given application.

Presentation of results and post processing

The next section of this paper includes several illustrations that are edited versions of graphical output from PsySound3. Matlab's graphing functions are used by the program, allowing charts to be edited, exported, and edited further. The program also provides full numeric representation of analysis outputs, along with statistical reduction of the output data.

The concept of using listening for analysis purposes has been an interest of the authors for some time, and several possibilities are presented by Cabrera *et al.* (2006). Hence, a distinctive innovation introduced by PsySound3 is a set of tools for sonification of analysis results. The program implements a range of techniques collectively known as 'exploratory sound analysis' (Ferguson and Cabrera 2008). The concept of exploratory sound analysis is to represent the analysis parameter(s) using a reorganised version of the original sound recording. This is in contrast to the more conventional and abstract form of sonification called 'auditory graphing' where, for example, the frequency of a tone might be mapped to parameter values.

EXAMPLES OF ANALYSIS

This section of the paper presents a set of simple analysis examples that illustrate some of the potential value and associated issues in using analysis methods that are currently implemented in PsySound3.

Effect of bandwidth on loudness

The modelling of loudness is considerably more complex than calculating sound pressure level, and several methods can be used. In its present form, PsySound3 implements the ISO532B steady state loudness model, Chalupper and Fastl's dynamic loudness model (2002), and Moore, Glasberg and Baer's (1997) steady state loudness model. Loudness models model sensitivity as a function of frequency, bandwidth, and time (in the case of dynamic models) and they yield natural loudness units (sones) rather than decibels. The fact that these relationships are neither independent nor linear provides what is both an advantage and disadvantage of loudness modelling: while the result should be a more accurate representation of loudness than a sound pressure level measurement, a calculation done for a particular listening level cannot be simply reinterpreted for other listening levels (as a sound pressure level measurement could be). While weighted sound pressure level does model the variation of sensitivity across the frequency range, it does not model the loudness effect of bandwidth – i.e., that broad bandwidth stimuli tend to be louder than narrow bandwidth stimuli of the same sound pressure level. This effect, which is easily experienced in controlled listening, is illustrated in Figure 1 by a comparison between pink noise and 1 kHz pure tone stimuli (both having a sound pressure level of 60 dB) as analysed by three loudness models. The charts show the specific loudness pattern, which is the loudness attributable to auditory filters (represented by Bark or Erb units), which when integrated yields the overall loudness. This is a particularly striking example because the result is contrary to a comparison between the stimuli using A-weighting: the 1 kHz tone is still 60 dB(A) after the application of A-weighting, but the pink noise has an A-weighted sound pressure level of 54.7 dB(A) due to the attenuation of low frequency power by the A-weighting filter. The example also serves to illustrate one of the problems with psychoacoustical modelling, that is, since loudness is a subjective phenomenon observed statistically, various theories may be proposed to model it, each with its own assumptions and limitations. In the low frequency range, the Bark units span a substantially broader frequency range than Erb units, which is one reason why the specific loudness pattern of the pink noise differs between the models.

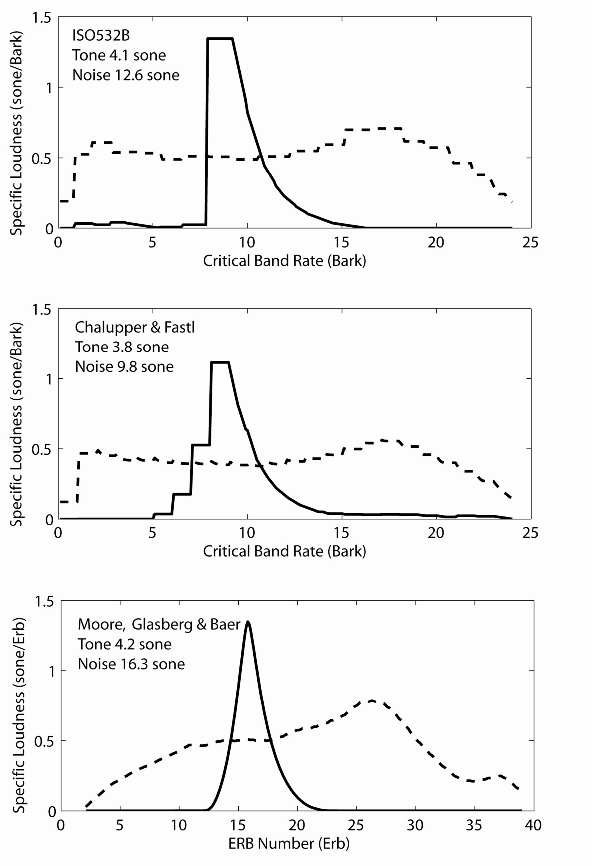

Figure 1. Loudness modelling of a 1 kHz pure tone (solid line) and steady state pink noise (dashed line) both with an unweighted sound pressure level of 60 dB. The three charts show the specific loudness patterns using three loudness models that are implemented in PsySound3.

Temporal integration

Another comparison of interest is the time response of measurements. Figure 2 shows three ways of measuring sound strength as a function of time, applied to a short speech phrase. The Hilbert transform can be used to derive instantaneous sound pressure level, which is arguably the shortest integration time possible, and much shorter the integration time of audition. The result of the Hilbert transform possesses a widely fluctuating fine structure with a quite well defined overall envelope revealing fluctuations in the voice level, as well as the exponential reverberation decay in the case of reverberant speech. The widely used sound pressure level using 'fast' integration (125 ms time constant in an exponential integrator) is also shown, which exhibits comparatively little variation during the speech phrase (especially in the case of reverberant speech), and a smooth exponential reverberation decay of about the same slope as the Hilbert-derived sound pressure level. Finally, the dynamic loudness model of Chalupper and Fastl (2002) yields a fairly widely varying result for time-varying loudness, but without the fine structure of the Hilbert-derived sound pressure level. The comparison between sound pressure level and loudness also highlights the difference between the units used (with loudness units a ratio scale of perceived loudness, while sound pressure level is a logarithmic scale of squared sound pressure which tends to compress the data). The reverberant decay does not form a straight line when loudness units are used (and the decay function will depend on the calibration level of the analysis).

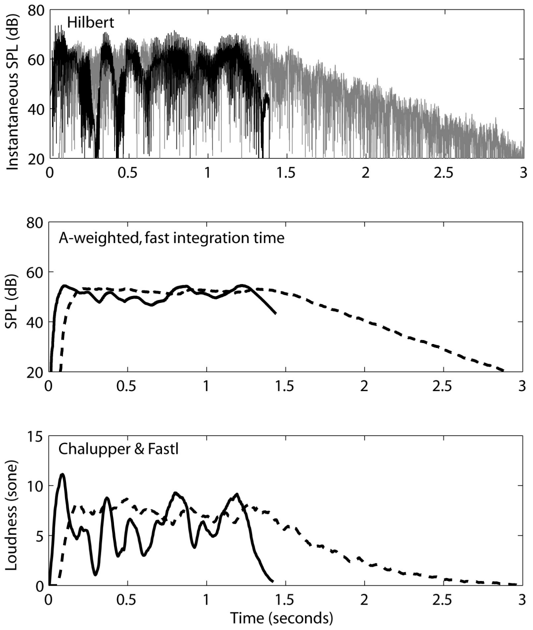

Figure 2. Comparative measures of sound strength as a function of time of an anechoic speech phrase "I'm speaking from over here" and the same recording convolved with the impulse response of a room (the traces that extend to 3 s). From top to bottom: the instantaneous sound pressure level derived from the Hilbert transform; A-weighted sound pressure level with fast temporal integration; and dynamic loudness based on the model of Chalupper and Fastl.

Although it may appear from Figure 2 that longer integration times are not likely to be useful for loudness modelling, in fact this depends on the context. Soulodre and Lavoie (2006) find that an integration time constant of the order of 3 s performs best when tracking subjective time-varying loudness responses, and our preliminary analysis using different data (music recordings) supports this conclusion. The reason for the effectiveness of the longer integration time is that the response parameter is confounds loudness perception and the processes involved in human response. By contrast, the much finer temporal resolution of a dynamic loudness model is not based on direct observation of time-varying loudness responses to continuous music or speech, but instead on low level testing of temporal masking and the relationship between duration and loudness.

Time-spectra

Spectrograms are a very common method for visualising how power spectrum changes as a function of time. PsySound3 uses the term 'time-spectrum' in a more general sense, referring to any spectrum-like data that vary with time. This includes, for example, frequency spectra derived from short term fast Fourier transforms or other filtering techniques, cepstra, short term auto-correlation functions, specific loudness patterns, specific roughness patterns and quantized pitch patterns.

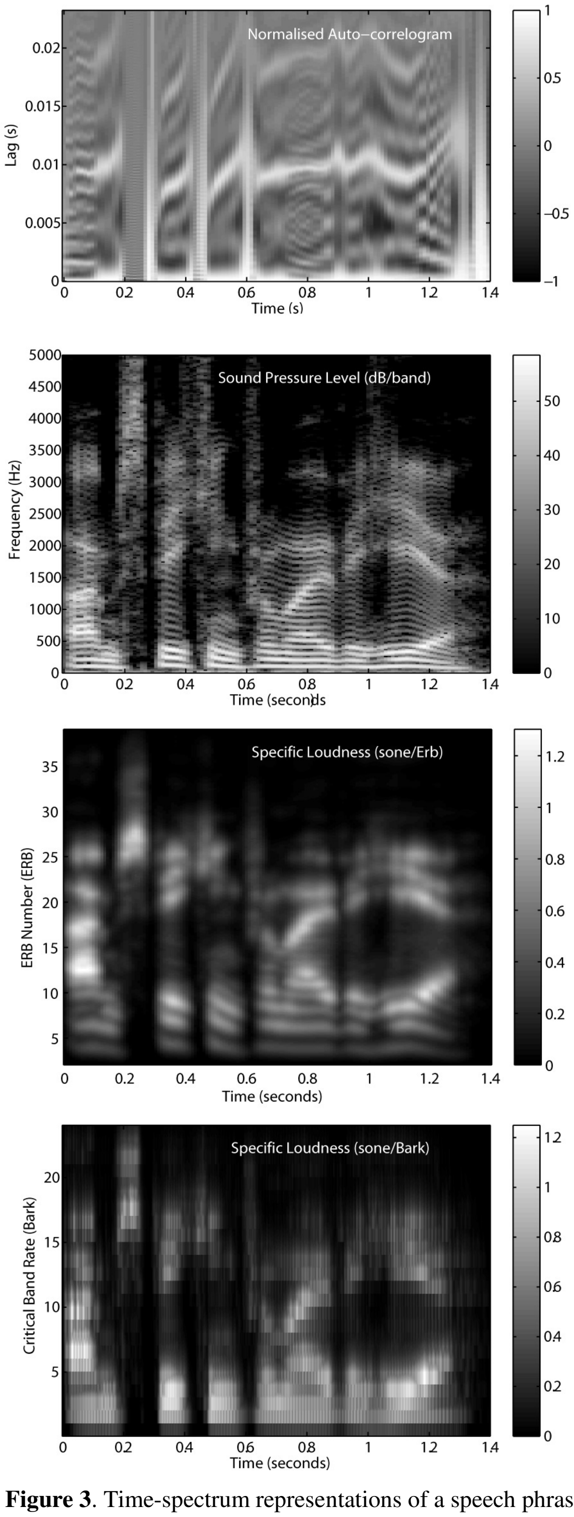

Figure 3. Time-spectrum representations of a speech phrase, "I'm speaking from over here". From top to bottom: normalised autocorrelogram; spectrogram; specific loudness pattern based on Moore et al.; and specific loudness pattern based on Chalupper and Fastl.

Figure 3 gives examples of time-spectrum outputs for the analysis of the same anechoic speech phrase as used in the previous example.

The second of the charts in Figure 3 is the familiar spectrogram, which clearly shows both the formant and harmonic structure of the speech, as well as features associated with consonants (for example high frequency content of the 's' at 0.2 s). The time-frequency resolution of spectrograms is a user-controlled parameter in the PsySound3 analyser, as would be usual in software of this type. The top chart (autocorrelogram) reveals periodicity in another way, with peaks due to low frequencies represented by large lag times. Hence the falling pitch at the end of the speech phrase is associated with an increasing lag time for the white peak. The lower two charts show time-varying specific loudness pattern, for repeated iterations of Moore, Glasberg and Baer's static model (second from bottom), and Chalupper and Fastl's dynamic model (bottom). The first of these models is implemented with a higher auditory filter resolution, with peaks formed by the individual lower harmonics and formants, while the coarser resolution of the second model does not show effects of individual harmonics (this is also partly due to the broader filter bandwidths of the Bark scale compared to the Erb scale). Note that Glasberg and Moore (2002) published a dynamic loudness model, but implementations for this in PsySound3 are yet to be optimised for practical use (due to the much greater computational load of the model).

Pitch height and sharpness

In this section we consider some parameters available to represent concepts related to frequency, pitch and sharpness. One aspect of pitch analysis is that, particularly in the context of music, we may be interested in extracting multiple simultaneous pitches, rather than only single time-series pitch parameters. The pitch model of Terhardt *et al.* (1982) provides a means of identifying a spectrum of pitch percepts, and respective saliences for the identified pitches. This model is based on a short term Fourier transform, followed by peak extraction and masking analysis. Two types of pitches are identified: spectral pitches – which are audible tones present in the spectrum (after the modelling of auditory masking); and virtual pitches – which are pitches inferred by the presence of harmonic series in the masked spectrum. Parncutt (1989) developed ways of further analysing the output of Terhardt's model for harmony analysis. Although Parncutt was concerned with analysis of a quantized frequency scale (twelve-tone equal temperament), PsySound2 implemented these models so that they could be applied to the analysis of arbitrary sound recordings, and this capability is retained in PsySound3. Nevertheless, since the envisaged application of this is analysis of twelve-tone equal temperament recordings of music, quantization is applied to the output (rather than input) of the models.

The top chart of Figure 4 illustrates one of the output data types from this implementation – a time-spectrum showing the pitch saliences of the twelve chroma (or pitch classes, with 1 denoting the musical note 'A') as a function of time, for a recording of a 3-note motif played on a shakuhachi. While chroma corresponding to the three notes are each clearly visible, it can also be seen that other chroma have significantly non-zero values, and indeed a vestige of the second note is sustained through the third note. Note that a chroma difference interval of 5 or 7 (for example, from 2-7 or 2-9 on the chart) corresponds to a perfect fourth or fifth respectively (relating to the harmonic ratios of 3:4 and 2:3).

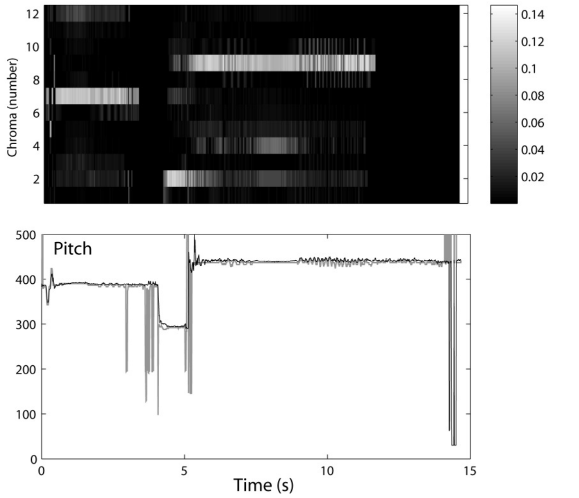

Figure 4. Representations of pitch, analysed from a 3-note phrase played on a shakuhachi. The upper chart shows the chroma pattern derived from a combination of Terhardt et al. and Parncutt's pitch models. The lower chart shows the estimate of a single time-varying pitch, based on the short-term auto-correlation function peak lag (grey) and SWIPE' (black).

Reducing a rich spectrum to a single 'frequency' or pitch height value can be done in many ways, some prioritising physical signal analysis, while others aim to represent perception. A simple way of tracking pitch is to find the lag of the highest non-zero peak in the short-term auto-correlation function of the signal, and read frequency as the inverse of the lag time. This approach is prone to several errors, and many researchers have developed more robust methods of pitch tracking. One of these methods, implemented in PsySound3, is SWIPE' (i.e. SWIPE with a prime symbol) (Camacho 2007). The lower chart of Figure 4 compares the frequency derived from short-term auto-correlation and the calculation from SWIPE'. Most obviously, the latter exhibits a substantially more stable pitch tracking, and there are also fine differences between the results of the two methods in terms of the exact pitch identified, and the tracking of vibrato.

The sharpness or brightness of sound is a characteristic of timbre rather than pitch, but is also a feature that can be derived from the frequency content of a spectrum, and sometimes is represented as a single frequency. A simple way of estimating brightness is to take the first moment of the power spectrum (known as spectral centroid) (Lichte 1941). In the case of a pure tone, centroid and pitch will be the same, but for complex sounds the centroid represents the balance of power across the frequency range. This simple approach can be compared to the concept of sharpness, which is modelled using a weighted centroid of the specific loudness pattern (Zwicker and Fastl 1999). Results for spectral centroid and sharpness analysis of the shakuhachi motif are shown in Figure 5. It can be seen that these share many features, but diverge at the end (the reverberant decay) because of the sensitivity of sharpness to the overall loudness (which is due to the sensitivity of masking curves to absolute level in the loudness model). Where there are common features, it can be seen that there are differences in scale between the two methods of modelling.

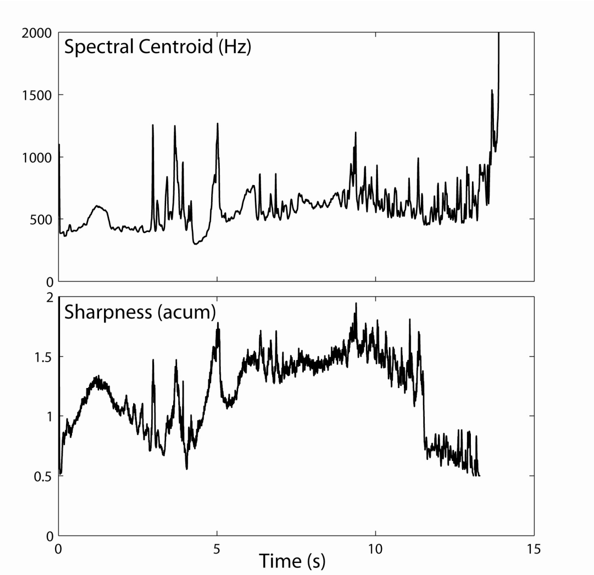

Figure 5. Time-varying representations of brightness or sharpness. The upper chart is the spectral centroid derived from the short-term Fourier transform, and the lower chart is sharpness, calculated from the output of Chalupper and Fastl's dynamic loudness model.

Other models of brightness of sharpness exist – in musical timbre analysis it is common to include the pitch value in the denominator, and so to provide an enumeration of brightness relative to the fundamental frequency. However Schubert and Wolfe (2006) found that that approach reduces the predictive power of the model in relation to subjective brightness judgments.

Pitch strength

Pitch strength can be understood and calculated in many different ways, and in its present form PsySound3 does not implement all available options. Perhaps the simplest representation of pitch strength is the height of the highest non-zero peak in the short-term auto-correlation function of the signal – a number between 0 and 1, where 0 indicates no pitch and 1 indicates a strong pitch. A similar but more refined pitch strength estimate is provided by SWIPE' (Camacho 2007), and these two estimates are shown for the shakuhachi sample in Figure 6. This reveals considerably greater sensitivity to degradations of pitch strength in the SWIPE' calculation than the simple auto-correlation calculation (which for the most part is close to its maximum value of 1).

The pitch modelling of Terhardt et al. (1982) and Parncutt (1989) also provides ways of representing various aspects of pitch strength, including the strength of spectral pitch components ('pure tonalness'), the strength of virtual pitches ('complex tonalness') and an estimate of the number of simultaneously audible pitches ('multiplicity'). Figure 6 shows the time-varying pure tonalness of the shakuhachi sample. While some features of the SWIPE' analysis are shared with the pure tonalness calculation, it is notable that pure tonalness falls to zero at the end of the note, rather than continuing through the reverberant decay. It can also be observed that vibrato reduces the pure tonalness (e.g., at the end of the first and last notes) showing that this type of pitch modelling is sensitive to the blurring of spectral peaks over the analysis window length.

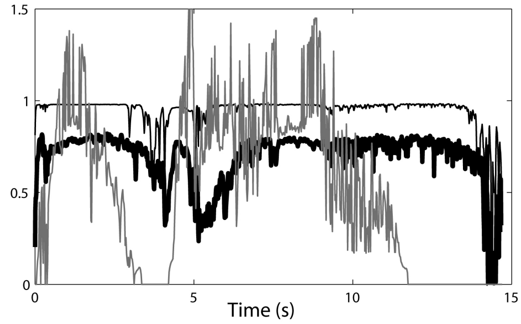

Figure 6. Time-varying pitch strength of a shakuhachi phrase, represented by the short-term autocorrelation function peak height (fine black), SWIPE' (heavy black), and pure tonalness (grey).

Binaural analysis

Most of the currently implemented analysers are for single channel input. For these, the program will analyse either one of the channels, or a sum or average of the two channels. However, the program infrastructure allows analysis of multichannel files, and in the field of psychoacoustics, binaural files are of particular interest. Currently the suite of analysers includes an interaural cross correlation analyser, based on the approach taken by Ando (1998). This estimates aspects of spatial hearing, including the lateralisation and width of auditory images, based on the time-varying short term interaural cross correlation function.

Figure 7 illustrates this type of analysis, for a recording of a speech phrase "I'm speaking from over here." with one repetition, played from a loudspeaker to a dummy head microphone in a room. Two examples are given – one with curtains covering two of the room walls, and the other without (in all other respects, the source signal and recording conditions are identical). These recordings correspond to stimuli described by Pop and Cabrera (2005), namely "room 2" with a 1.6 m source-receiver distance.

The analysis shows that although the auditory image tends to be centred (tau values close to 0 ms), the width of the auditory image (which is inversely related to IACC) increases once the room reverberation is contributing to the sound. Greater image width (lower IACC) occurs in the more reverberant room condition. Note that while this is an application of running interaural cross-correlation, a similar analysis is often applied to binaural room impulse responses to estimate auditory spatial parameters.

Binaural loudness modelling has advanced recently, with models proposed by Moore and Glasberg (2007) and Sivonen and Ellemeir (2008), and a recent study at the University of Sydney on binaural loudness in soundfields of various diffusivities (Miranda and Cabrera 2008) has seen further evaluation of these models and their implementation in the PsySound3 platform. We hope to add more binaural and multichannel analysers to PsySound3 as the project develops further.

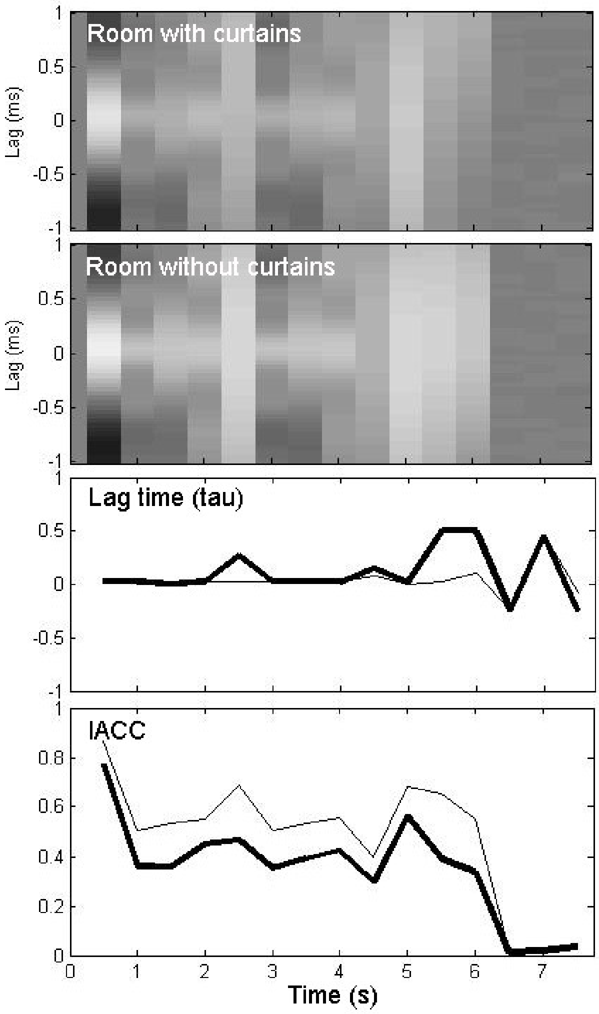

Figure 7. Binaural analysis of a once-repeated speech phrase ("I'm speaking from over here") in a room with and without curtains over two of the four walls. The speech starts at 0.5 s and the repetition of it at 2.5 s, and the sound after 4 s is essentially reverberant decay. The top two charts show the time-varying interaural cross correlation function (with black corresponding to -1 and white corresponding to 1). The lower two charts show the interaural lag time (an indicator of lateralisation) and the interaural cross correlation coefficient (an indicator of image width). The thin line is for the room with curtains.

CONCLUSION

PsySound3 is a developing project, driven by the research priorities and teaching needs of the authors and others in the development group. PsySound3 has found applications in research on music perception, auditorium acoustics (Lee and Cabrera 2008), developing psychoacoustical models (Miranda and Cabrera 2008) and auditory display (Ferguson and Cabrera 2008). Like its predecessor, PsySound3 is also likely to find applications in scientific studies of music perception (c.f. Schubert 2004). The development group is open, so the future direction of the project depends on the interests of active participants.

This paper has illustrated how the application of multiple analysis techniques can reveal diverse information both about the object of the analysis (the sound recording) and the analysis algorithm. In a sense, the 'correct' answer in psychoacoustical modelling is found by subjective testing, and models, which are constructed to emulate a limited set of subjective responses, are then applied to diverse sound recordings. When multiple models of the one percept are developed, they will yield somewhat different results for a given input, although we cannot quantify the error without, at the very least, knowledge of the model limitations, and preferably further subjective testing. Hence, it is hoped that providing a diversity of analysis methods will foster a healthy scepticism of psychoacoustical models, while also providing substantial analysis power for diverse research projects.

Computing capacity is one of the important limitations of PsySound3. Some analysis algorithms are quite slow, and have high memory demands. While the program runs on most modern computers, it runs better on high capacity computers. It should benefit from ongoing improvements in computer capacity, as well as efforts towards code optimisation.

ACKNOWLEDGMENTS

Many people have made significant contributions to this project, including Jens Brosbøl, Arturo Camacho, Josef Chalupper, Peter Daniel, Dik Hermes, Matt Flax, Shin-Ichi Sato, Alex Tarnopolsky and Ernst Terhardt. There are also many others who have contributed useful feedback and suggestions. This project was funded by the Australian Government through the Australian Research Council LIEF scheme. The authors would particularly like to thank Catherine Stevens (MARCS Auditory Laboratories, University of Western Sydney) for her support throughout this project.

REFERENCES

Ando, Y. 1998, *Architectural Acoustics*, Springer

Cabrera, D. 1999, "Psysound: A computer program for psychoacoustical analysis," *Aust. Acoust. Soc. Conf.*, Melbourne, Australia, 47-54

Cabrera, D., Ferguson, S. and Maria, R. 2006, "Using sonification for teaching acoustics and audio," *1st Australasian Acoust. Soc. Conf.*, Christchurch, New Zealand, 383-390

Cabrera, D., Ferguson, S. and Schubert, E. 2007, "PsySound3: Software for acoustical and psychoacoustical analysis of sound recordings," *13th Int. Conf. Auditory Display*, Montreal, Canada, 356-363

Camacho, A. 2007, *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*, PhD thesis, University of Florida

Chalupper, J. 2000, "Modellierung der Lautstärkeschwankung für Normal- und Schwerhörige," *DAGA 2000*, 254-255

Chalupper, J., and Fastl, H. 2002, "Dynamic Loudness Model (DLM) for Normal and Hearing-Impaired Listeners," *Acta Acustica united with Acustica* 88, 378-386

Ferguson, S. and Cabrera, D. 2008, "Exploratory sound analysis: sonifying data about sound," *14th Int. Conf. Auditory Display*, Paris, France

Glasberg, B.R. and Moore, B.C.J. 2002, "A model of loudness applicable to time-varying sounds," *J. Audio Eng. Soc.* 50, 331-342

Lee, D. and Cabrera, D. 2008, "Analysing room impulse responses with psychoacoustical algorithms: a preliminary study," *Aust. Acoust. Soc. Conf.*, Geelong, Australia

Lichte, W. 1941, "Attributes of complex tones," *J. Exp. Psych.* 28(6), 455-480.

Miranda, L. and Cabrera, D. 2008, "Evaluation of binaural loudness models with signals of different diffusivity," *Aust. Acoust. Soc. Conf.*, Geelong, Australia

Moore, B.C.J., Glasberg, B.R. and Baer, T. 1997, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc.* 45, 224-240

Moore, B.C.J., Glasberg, B.R. 2007, "Modeling binaural loudness", *J. Acoust. Soc. Am.* 121(3), 1604-1612

Parncutt, R. 1989, *Harmony: A Psychoacoustical Approach*, Springer

Pop, C. and Cabrera, D. 2005, "Auditory room size perception for real rooms," *Aust. Acoust. Soc. Conf.*, Busselton, Australia

Schubert, E. 2004, "Modeling perceived emotion with continuous musical features," *Music Perception* 21: 561-585

Schubert, E. and Wolfe, J. 2006, "Does timbral brightness scale with frequency and spectral centroid?" *Acta Acustica united with Acustica* 92(5), 820-825

Sivonen, V.P. and Ellermeir, W. 2008, "Binaural loudness for artificial-head measurements in directional sound fields," *J. Audio Eng. Soc.* 56(6), 452-461

Soulodre, G.A. and Lavoie, M.C. 2006, "Development and evaluation of short-term loudness meters," *121st Audio Eng. Soc. Conv.*, San Francisco, USA

Terhardt, E., Stoll, G. and Seewann, M. 1982, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.* 71, 679-688

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