

ICA and Related Models Applied to Audio Analysis and Separation

M D Plumbley, S A Abdallah, J P Bello, M E Davies, J Klingseisen, G Monti and
M B Sandler

Audio & Music Technology Lab, Department of Electronic Engineering
King's College London, Strand, London WC2R 2LS, UK
mark.plumbley@kcl.ac.uk

Abstract

Over the last decade, an increasing number of researchers have become interested in the problem of Computational Auditory Scene Analysis (CASA), the use of computers to recognize sound sources in a complex auditory environment. In this paper, we give an overview of some approaches we are using in this area, and in particular for automatic music transcription and separation of audio sources.

1 Introduction

While automatic interpretation of visual scenes has long been an interest of AI researchers, the equivalent problem of automatic interpretation of auditory scenes (i.e. sounds) has not attracted as much research (apart from the special application of automatic speech recognition).

However, over the last decade or so, an increasing number of people have become interested in this problem of *Computational Auditory Scene Analysis* (CASA), the use of computers to recognize sound sources in a complex auditory environment, typically music or other non-speech sounds. Several approaches based on AI techniques, such as the 'blackboard' approach, have been used to tackle this problem so far [18, 9, 17].

In contrast to the traditional speech recognition problem, this is concerned with the analysis of complex auditory scenes, made up of multiple sound sources. The name of the field was inspired by the seminal work of Bregman [5].

Particular aspects of this problem which we are interested in include Automatic Music Transcription (e.g. converting music audio sounds to MIDI), and Blind Source Separation (sometimes known as the 'Cocktail Party Problem').

2 Automatic music transcription

The goal of the automatic music transcription problem is to be able to recognize the different notes being played in an audio piece of music. In addition, it may also be desirable to be able to recognize other features such as the identity of instruments being played, loudness, stress or stress.

There are several possible applications for successful music transcription. As one example, the OM-RAS (Online Music Retrieval And Searching) project is aiming to construct a digital music library system. It may hold music in various forms, from images of printed musical scores, through MIDI files, to audio renditions from particular performances, and will need to be able to convert from any one of these formats to another. Automatic music transcription will allow the audio renditions to be converted to MIDI, and from there to some score-like representation if necessary.

Another application is in low bit-rate audio coding. The recent MPEG-4 Structured Audio (MP4-SA) standard offers the potential for future very low bit rate compression of audio [23]. High quality polyphonic music analysis is needed if the structured descriptions (currently authored by hand) are to be created automatically from audio signals.

Unfortunately, there are many variables that make automatic music transcription hard. There are a wide range of different instruments available, producing many different possible waveforms. Percussive instruments may not have a simple frequency content. Also, much western music is composed to create 'interesting' effects for the ear, through e.g. the use of notes in harmonic relation to each other, and these can also cause problems for music transcription systems. In the next sections, we will describe some methods of tackling these difficulties.

2.1 Monophonic music transcription

A special case of the automatic music transcription problem is *monophonic* music transcription. Here it is known that at most one note is present at any one time, e.g. from a solo trumpet. In contrast, *polyphonic* music transcription refers to the case where any number of notes may be present at any one time. (Note that there is possible confusion here, since ‘monophonic’ or ‘mono’ can also mean single-channel audio, in contrast to e.g. ‘stereo’ audio).

If a source is known to be monophonic, the problem is greatly simplified. Only a single fundamental frequency need be found to model the different frequencies found at a particular point in time [3].

2.2 Polyphonic music transcription

With polyphonic music transcription, particular problems are introduced. For example, consider two notes played at the same time, where their frequencies are an octave apart (e.g. one is at double the frequency of the other). Without knowing what each note ‘should’ sound like, it is very difficult to tell the difference between this situation, and one where a single note is being played that has harmonics equal to the sum of the two notes.

2.3 Blackboard model approaches

One popular approach to automatic music transcription is the use of the a *Blackboard* model, used for example by Mellinger [18] and Martin [17]. In this model, hypotheses are generated and placed into a database (the Blackboard) which is visible to a number of Knowledge Sources. Each Knowledge Source (KS) is composed of a precondition/action pair: if the precondition is satisfied (by hypotheses present on the blackboard), then the KS will fire, removing hypotheses it needs and replacing them by others. For example, one KS may know that frequency tracks at f , $2f$ and $3f$ is evidence for a note with fundamental frequency f . On firing, it would replace these ‘frequency tracks’ with a partial hypothesis for a note at f .

Knowledge sources can be quite complex. For example, Godsmark and Brown [12] described a blackboard model which integrates evidence from several grouping principles, such as temporal and frequency proximity, common onset and offset times, common frequency modulation, and similarity of timbre. Each principle is handled by a special hypothesis formation expert, with the parameters of the experts set to be consistent with known psychophysical measurements.

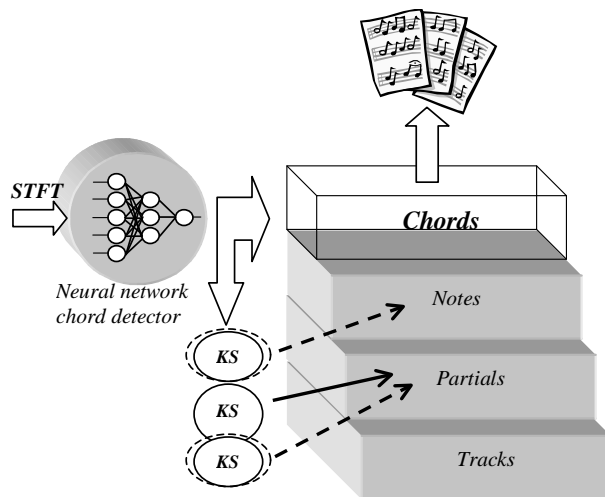


Figure 1: Blackboard model incorporating neural network chord recognizer.

We have also investigated the blackboard model, and have augmented this to include the use of a feed-forward MLP neural network as a knowledge source for chord recognition [4] (Fig. 1). Currently this is giving us encouraging results for chords of up to three notes.

2.4 Multiple-cause models

One of the issues to be addressed in polyphonic music transcription is the issue of *timbre modelling*. If we knew what the shapes of the notes were like, in the frequency domain, we could find a combination of notes that exactly matched the frequency content of the input. However, this is somewhat of a chicken-and-egg problem: until we know what the note shapes are for a particular piece, it seems that we cannot find the notes, and if we do not know the notes, we cannot find the note shapes. One possible solution to this is to use the *Multiple-Cause Model*.

The multiple-cause model [21] (Fig. 2) searches for representations of the underlying causes of the input data, together with amounts of each ‘cause’, which take account of the input data as closely as possible. This model is designed to cope with input data which is composed of several causes active at the same time. In contrast to many neural networks, this network does not operate in a simple feed-forward manner: rather the encoding layer and connections are adjusted until the encoding forms a good reconstruction of the observed data.

The multiple-cause model was originally proposed for analysing images into constituent causes. A clas-

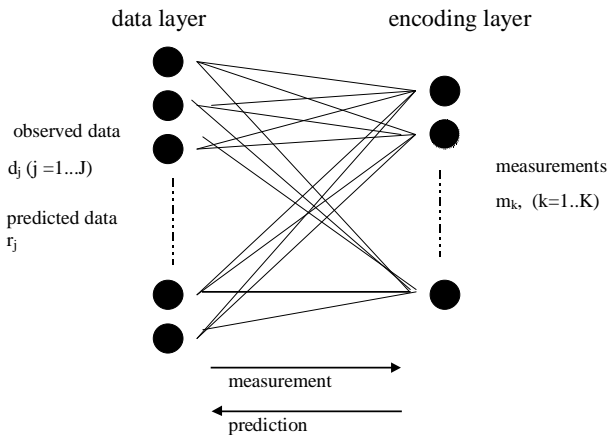


Figure 2: Multiple cause model architecture.

sic example is the ‘bars’ problem, where each image is composed of a number of horizontal and vertical black bars on a white background. Thus an image is ‘caused’ by several bars, maybe overlapping, and each pixel in the input image may be ‘caused’ by one or more bars being present in the image. The multiple cause model is trained by minimising an error function, such as negative log likelihood or mean squared error, by adjusting both the underlying measurements of each cause, and the patterns due to each cause.

The multiple cause model is underdetermined, due to interdependence between the measurements of each cause and the underlying patterns. Depending on the problem, some form of constraints may need to be imposed [13, 6]. For a musical example, we would expect positive amounts of each underlying note. On the assumption that notes and instruments were independent sources, each would then contributing approximately positive amount to the received spectrum.

2.4.1 Artificial spectra

As an initial step, we next applied the multiple-cause model to artificial spectra produced from synthesized sounds [15]. In the first instance, we trained the model on mixtures of spectra, downsampled to 30 bins, of a synthesized clarinet playing one of 8 notes (G3, C4, A3, D4, F4, G4, A4, E4). The training set was composed of linear additions of these basic spectra (not mixed in the time domain), with probability $p = 0.4$ of each spectrum.

About 800 presentations of training patterns were necessary for successful learning. This was also tested on notes from a violin and an alto recorder,

with similar results.

Separation of patterns composed from spectra of different synthesized instruments playing the same note was also attempted. For six instruments, about 600 presentations were needed to separate the patterns, and about 3000 presentations for 10 instruments. This appeared to take longer to learn than the experiments with different notes on the same instrument. This is probably due to the similarity between the patterns representing the same note, particularly the alignment of the fundamental and first few harmonics into the same bins in each pattern.

Combining these two approaches, patterns composed from the spectra of three instruments (Clarinet, Oboe, Trumpet) playing each of three notes were used for the training patterns. The spectra used to form the training patterns were downsampled into log-scaled bins, with a relative scaling of $2^{1/12}$ between the bins. With synthesized instruments, different notes on the same instrument would then appear simple as shifted along this log-frequency scale, with 1 bin shift equivalent to 1 semitone.

A multiple-cause model with $K = 9$ measurement units learned the patterns after about 700 presentations of the training patterns. After this, the patterns were post-processed to identify which patterns were relative shifts of each other. This correctly identified that 3 instruments were used, with relative semitone shifts of $(0, +2, +4)$, $(0, +1, +5)$ and $(0, +2, +3)$.

2.4.2 Real Sounds

We constructed an audio signal composed of a linear time-domain addition of pulses of notes played on different instruments from the University of Iowa musical instrument samples web page [19] (Figure 3). This resulted in an audio signal of 8.3s sampled at 44.1kHz, and fourier transformed with a window width of 4096 samples, yielding 90 spectra of about 0.09s duration each.

The algorithm found most of the nine underlying patterns after 300 presentations of the set of training patterns, equivalent to 20 hour’s learning using Matlab on a 350MHz Pentium II [15]. In Figure 4 we see that the sounds have been separated such that the input sound is represented by a small number of measurement units, with other units off. This indicates that the output units have found a *sparse coding* [10], i.e. a coding with many units ‘off’, even without penalty terms that have been used to encourage sparse output distributions (see e.g. [14, 20]).

There are some instruments that are not completely separated: in particular, it seems that sep-

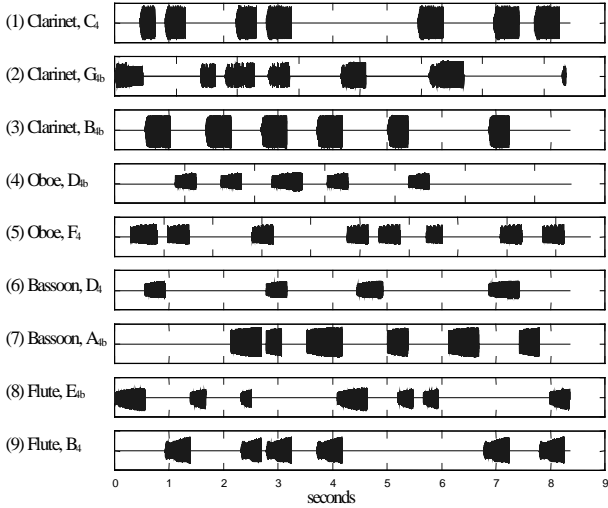


Figure 3: Waveforms of nine notes.

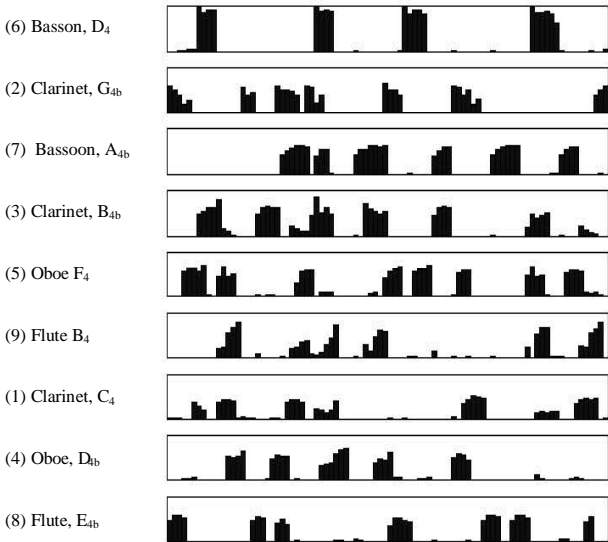


Figure 4: Activation of measurement units m_i

aration of flute and clarinet is difficult. In this example, Flute E_{4b} (input 8) and Clarinet G_{4b} (input 2) are poorly separated, with parts of each instrument found in the corresponding outputs (the label in Fig 4 indicates the closest instrument/note). Also, some of the Flute B_4 (input 9) is still mixed with the Clarinet B_{4b} (input 3), with the attack phase of the flute being ‘picked up’ by the Clarinet output. This difficulty may be due to the relatively pure waveforms, and therefore dominant fundamentals, that these instruments have, although more investigation is needed to confirm this.

2.5 ICA and sparse coding

Recently, Independent Component Analysis (ICA) [2] and the related technique of *sparse coding* using an overcomplete basis [11] have emerged as promising areas of research. The observation in the previous section that a transcription of a musical sound has a sparse representation suggests that sparse coding could be used directly in the search for a possible transcription.

We have applied this sparse coding approach to the analysis of polyphonic harpsichord music [1]. We use a ‘prior’ for note probability that contains a sharp peak at zero. This leads to a coding of the input spectra where most notes are ‘off’, and only a few notes represent the input.

We tested our algorithm on real sounds generated by a MIDI synthesizer and then resampled. We used Bach’s Partita in A minor for keyboard, BWV827, which was chosen since it mainly consists of two or three independent lines, with a few block chords. We used magnitude spectra rather than power spectra so that the noise be more approximately Gaussian, as assumed in our model [1].

The model learned 55 basic spectra, 49 of which were note spectra, and others appeared to represent transistions. When the output activations of these basis vectors are plotted, they show a good ‘piano-roll’ representation for the music contained in the input spectrogram (Fig. 5). To check, a simple re-synthesis was performed. After some trial and error adjustments, this resulted in a passable (but certainly imperfect!) rendition of the original piece.

3 Audio source separation

Most of the above automatic music transcription approaches use just a single audio channel. However, we are also interested in how to analyze sound scenes when more than one audio channel is available, e.g. from several microphones. This is referred to as the

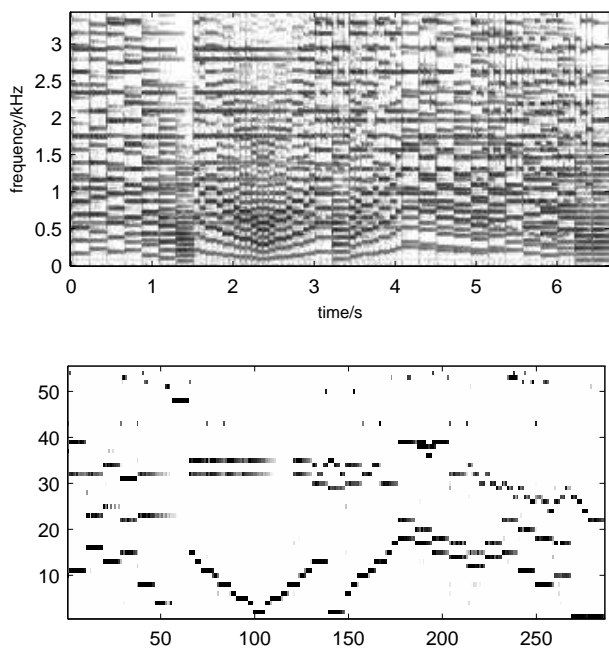


Figure 5: Input spectrogram (top) and output note activations (bottom). The output is visibly sparser than the input.

Blind Source Separation or Independent Component Analysis (ICA) problem [2].

The original ICA problem was formulated for N observations that are simple linear mixtures of N sources. However, in more realistic auditory environments, room acoustics can introduce delays, echoes and reverberation, and many authors have proposed algorithms to tackle this problem [7, 16, 22].

Many of these approaches can be viewed as attempting to perform ICA on different frequency bands, with some coupling between frequency components. The coupling is necessary to overcome the so-called *permutation problem*: ICA produces answers which are symmetrical with respect to any permutation of the sources. If this were allowed to happen at each frequency band independently, we may get e.g. frequency band f_a of source 1 presented on output 1, but frequency band f_b of the same source being presented instead on output 2.

Our approach to this problem [8] is to introduce an extra scale parameter β_k for each source k , modelling the tendency for the amplitudes of the various frequency bands of a particular source to vary together. We then use a *likelihood ratio jump* test to try out the possible permutations, and choose the permutation with maximum likelihood.

Comparisons of this algorithm with the Smaragdis [22] algorithm (this time on speech sources) suggest that for simple room acoustics, either approach will work well. However, if a room with non-trivial echo components around 5-10ms is used, then the Smaragdis algorithm failed to separate the sources, but our likelihood ratio jump method was successful (Fig. 6).

4 Conclusion

We have given a brief overview of some of our approaches to the analysis of complex auditory scenes. For automatic music transcription, we have outlined some approaches we have used to tackle this problem based on blackboard models, multiple-cause models, and sparse coding. For source separation, we have introduced the likelihood ratio jump test as a method for robustly separating audio sources which have been mixed in an acoustic environment with non-trivial echo.

Computer-based analysis of complex auditory scenes is an interesting and challenging area of research. With current popular interest in e.g. MP3 and delivery of music over the internet, we expect to see many more advances in this area.

Acknowledgements

SA is supported by an EPSRC Postgraduate Studentship. JPB is supported by the “Gran Marciscal de Ayachucho” Foundation in Venezuela and the British Council. JPB and GM are supported by the OMRAS project, jointly funded by JISC (UK) and NSF (USA). JK was supported by the EC, through the Socrates placement scheme.

References

- [1] S. A. Abdallah and M. D. Plumbley. Sparse coding of music signals. Submitted for publication, 2001.
- [2] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.
- [3] J. P. Bello, G. Monti, and M. B. Sandler. An implementation of automatic transcription of monophonic music with a blackboard system. In *Proceedings of the Irish Signals and Systems Conference (ISSC 2000)*, Dublin, Ireland, 2000.

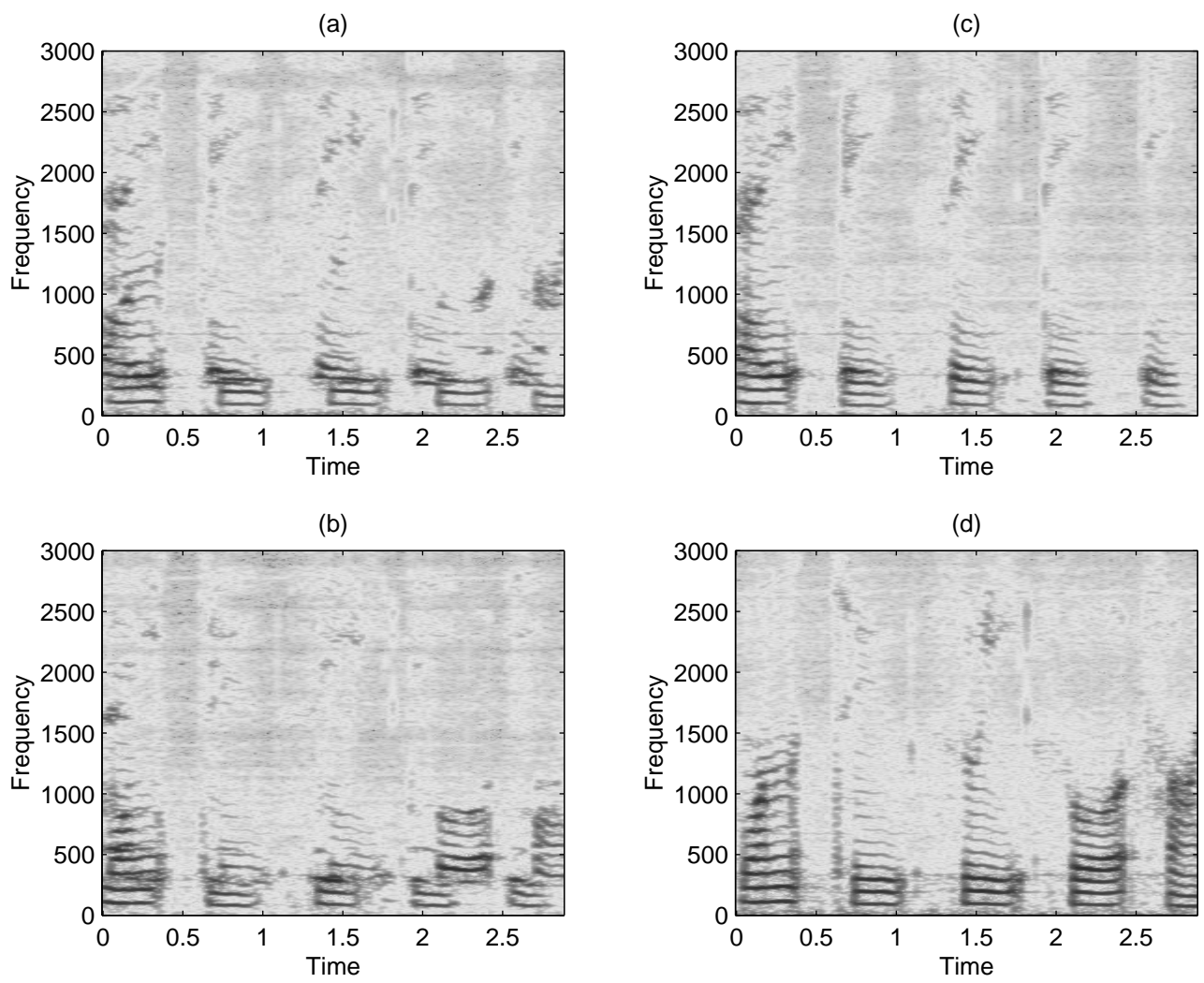


Figure 6: Spectrograms for the source estimates using the Smaragdis algorithm, (a) and (b), and the Likelihood Ratio Jump algorithm (c) and (d)

- [4] J. P. Bello and M. B. Sandler. Blackboard system and top-down processing for the transcription of simple polyphonic music. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*, Verona, Italy, December 7-9, 2000.
- [5] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA, 1990.
- [6] D. Charles and C. Fyfe. Modelling multiple-cause structure using rectification constraints. *Network: Computation in Neural Systems*, 9:167-182, 1998.
- [7] A. Cichocki, S. Amari, and J. Cao. Blind separation of delayed and convolved signals with self-adaptive learning rate. In *Proc. NOLTA '96*, 1996.
- [8] M. E. Davies. Audio source separation. Submitted for publication, 2001.
- [9] D. P. W. Ellis. *Prediction-Driven Computational Auditory Scene Analysis*. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, June 1996.
- [10] D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559-601, 1994.
- [11] D. J. Field and B. A. Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607-609, 1996.
- [12] D. Godsmark and G. J. Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27:351-366, 1999.
- [13] G. F. Harpur and R. W. Prager. Techniques for low entropy coding. Technical Report CUED/F-INFENG/TR. 197, Engineering Department, Cambridge University, UK, 1995.
- [14] G. F. Harpur and R. W. Prager. Development of low entropy coding in a recurrent network. *Network: Computation in Neural Systems*, 7:277-284, 1996.
- [15] J. Klingseisen and M. D. Plumbley. Towards musical instrument separation using multiple-cause neural networks. In *Proceedings of the International Workshop on Independent Component Analysis And Blind Signal Separation, 19-22 June 2000, Helsinki, Finland*, pages 447-452, 2000.
- [16] T. W. Lee, A. J. Bell, and R. H. Lambert. Blind separation of delayed and convolved sources. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 758-764. MIT Press, Cambridge, MA, 1997.
- [17] K. D. Martin. A blackboard system for automatic transcription of simple polyphonic music. Technical Report 385, MIT Media Lab, Perceptual Computing Section, July 1996. Available at <ftp://sound.media.mit.edu/pub/Papers/kdm-TR385.ps.gz>.
- [18] D. K. Mellinger. *Event Formation and Separation in Musical Sound*. PhD thesis, Center for Computer Research in Music and Acoustics, Stanford University, 1991. Also Dept of Music Report STAN-M-77.
- [19] University of Iowa. Musical instrument samples web page. URL <http://theremin.music.uiowa.edu/~web/sound/>, January 1999.
- [20] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7:333-339, 1996.
- [21] E. Saund. A multiple cause mixture model for unsupervised learning. *Neural Computation*, 7:51-71, 1995.
- [22] P. Smaragdis. Information theoretic approaches to source separation. Master's thesis, MAS Department, Massachusetts Institute of Technology, 1997.
- [23] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer. Structured audio: Creation, transmission, and rendering of parametric sound representations. *Proceedings of the IEEE*, 86:922-940, 1998.