



Audio Engineering Society Convention Paper

Presented at the 114th Convention
2003 March 22–25 Amsterdam, The Netherlands

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

An Independent Component Analysis Approach to Automatic Music Transcription

Samer A. Abdallah¹, Mark D. Plumbley¹

¹ *Queen Mary, University of London, Mile End Road, London E1 4NS, United Kingdom.*

Correspondence should be addressed to Samer A. Abdallah (samer.abdallah@elec.qmul.ac.uk)

ABSTRACT

We used Independent Component Analysis (ICA) with sparse coding to analyze music spectral sequences. We modelled an audio spectrum as an approximate mixture of the spectra of individual notes, using our ICA approach to “unmix” this to find the individual notes and note spectra. Notes are assumed to be approximately independent, and sparse (mostly off). Results on synthesized harpsichord music are encouraging, producing an approximate piano-roll transcription, and a passable rendition of the original music when resynthesized. We are currently working to extend and improve this through the use of temporal information of note activities and to handle more complex timbral behaviour.

INTRODUCTION

In the last few years, Independent Component Analysis (ICA) has emerged as a promising method for

Blind Signal Separation (BSS), where n sensors (e.g. microphones) receive different mixtures from m independent signals (sources) [1]. ICA has been ap-

plied to EEG analysis and audio source separation, where normally $m = n$. Some sources are known to have a high probability of being zero, a so-called *sparse source*[2]. In this case, the technique of *sparse coding* [3] can be used, and can allow us to separate more sources than there are sensors. This results in only a few sources being “on” at any particular time.

In this paper, we describe an approach to automatic transcription of polyphonic music based on this sparse ICA approach.

We will describe a sparse coding ICA system which operates on a single musical audio spectral sequence rather than multiple audio waveforms [4]. The intention is that observed spectra can be considered to approximate a mixture of the spectra of individual musical notes, and the notes will be “more independent” and “more sparse” than the original spectra. We should therefore be able to extract notes and harmonic profiles from only the audio spectrogram, without any pre-training on individual notes.

GENERATIVE MODEL

We use a causal latent variable model in which the observed vectors $\mathbf{x} \in \mathbb{R}^n$ are noisy linear mixtures of m basis vectors, with sources s_i assumed to be independent random variables and arranged into a vector $\mathbf{s} \in \mathbb{R}^m$.

Thus, we have $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{e}$, where \mathbf{A} is an $n \times m$ matrix encoding the basis vectors and \mathbf{e} is a Gaussian random vector representing additive noise. An equivalent graphical model is illustrated in Fig. 1.

The conditional density of the generated vector \mathbf{x} is given by

$$P(\mathbf{x}|\mathbf{A}, \mathbf{s}) = \left[\frac{\det \mathbf{\Lambda}_e}{(2\pi)^n} \right]^{1/2} \exp -\frac{1}{2} \mathbf{e}^T \mathbf{\Lambda}_e \mathbf{e}. \quad (1)$$

As mentioned above, the elements of the source vector \mathbf{s} are assumed to be independent, and are assumed to be drawn from a continuous density $p(\cdot)$, giving

$$P(\mathbf{s}) = \prod_{i=1}^m p(s_i). \quad (2)$$

The resulting density model for the observed data is

$$P(\mathbf{x}|\mathbf{A}) = \int_{\mathbb{R}^m} P(\mathbf{x}|\mathbf{A}, \mathbf{s})P(\mathbf{s}) \, ds. \quad (3)$$

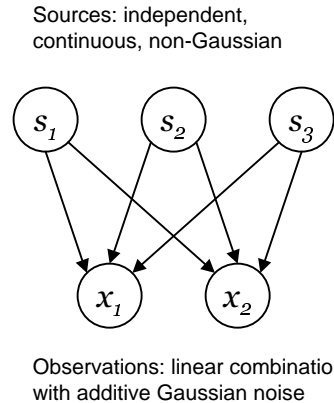


Fig. 1: Graphical model of sparse coder with a three-dimensional latent space.

If \mathbf{A} is known and \mathbf{x} is observed, then \mathbf{s} can be estimated from the posterior density,

$$P(\mathbf{s}|\mathbf{A}, \mathbf{x}) = \frac{P(\mathbf{x}|\mathbf{A}, \mathbf{s})P(\mathbf{s})}{P(\mathbf{x}|\mathbf{A})}. \quad (4)$$

A maximum *a posteriori* (MAP) estimate can then be found at the posterior mode:

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s}} P(\mathbf{s}|\mathbf{A}, \mathbf{x}). \quad (5)$$

If the posterior is sufficiently smooth, this can be found by gradient ascent on the logarithm of this function.

Estimation of the mixing matrix

To estimate the basis matrix \mathbf{A} , we maximise the following objective function:

$$\mathcal{L} = E_{\mathbf{x}} \log P(\mathbf{x}|\mathbf{A}), \quad (6)$$

where $E_{\mathbf{x}}$ denotes the expectation over the observed distribution of \mathbf{x} . This leads to the update rule

$$\begin{aligned} \Delta \mathbf{A} &= \eta \mathbf{\Lambda}_e \int (\mathbf{x} - \mathbf{A}\mathbf{s})\mathbf{s}^T P(\mathbf{s}|\mathbf{A}, \mathbf{x}) \, ds \\ &= \eta \mathbf{\Lambda}_e E_{\mathbf{s}|\mathbf{x}, \mathbf{A}} \mathbf{e}\mathbf{s}^T, \end{aligned} \quad (7)$$

where η is a learning rate parameter.

After further approximations and a covariant gradient modification [5], we get:

$$\Delta \mathbf{A}_{LS} = \eta \mathbf{A} [\gamma(\hat{\mathbf{s}})\hat{\mathbf{s}}^T - \mathbf{I}]. \quad (8)$$

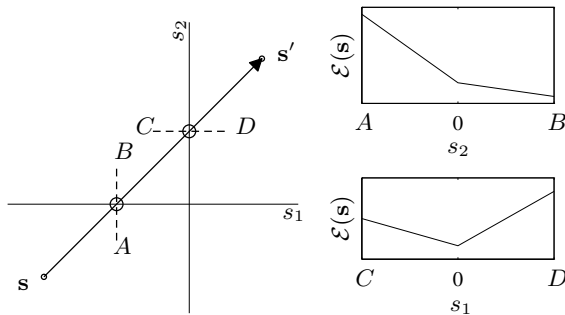


Fig. 3: A Two dimensional illustration of the operation of the modified active set optimiser. Given the proposed step from \mathbf{s} to \mathbf{s}' , and the local behaviour of the gradient along the segments AB and CD , the modified optimiser would truncate the step at the *second* zero crossing, and inactivate the *first* coordinate, s_1 .

where $\gamma(\mathbf{s}) \stackrel{\text{def}}{=} -\nabla_{\mathbf{s}} \log P(\mathbf{s})$. Since $P(\mathbf{s})$ is factorial, its gradient will be an element-wise function defined in terms of the prior density $p(s)$:

$$[\gamma(\mathbf{s})]_i = \gamma(s_i), \quad \gamma(s) \stackrel{\text{def}}{=} -\frac{d}{ds} \log p(s). \quad (9)$$

Modification of the Prior

The parameters μ and α control the width and relative mass of the central peak in the prior (Fig. 2):

$$f(s) = \begin{cases} Ce^{-|s|} & : |s| \geq \mu, \\ CKe^{-\alpha|s|} & : |s| < \mu, \end{cases} \quad (10)$$

where C is a normalisation constant, and $K = e^{\mu(\alpha-1)}$ to ensure continuity. The associated function $\gamma(s)$, illustrated in Fig. 2(b), is given by

$$\gamma(s) = \begin{cases} \text{sgn } s & : |s| \geq \mu, \\ \alpha \text{sgn } s & : |s| < \mu. \end{cases} \quad (11)$$

This prior results in a thresholding behaviour when computing $\hat{\mathbf{s}}$.

Optimization Algorithm

We modified the approach of Lewicki and Sejnowski [6] to estimate the unknown parameters. To estimate the value of \mathbf{A} , we marginalize out \mathbf{s} and maximize the expected value of $\log P(\mathbf{x}|\mathbf{A})$: this is the learning stage. If we have an estimate of \mathbf{A} , we can calculate an estimate of \mathbf{s} by maximising $P(\mathbf{s}|\mathbf{A}, \mathbf{x})$.

We used a ‘‘sparsified’’ Laplacian prior for \mathbf{s} , so that $\log p(s)$ is piecewise linear.

A modified ‘active set’ quasi-Newton optimisation algorithm was developed to deal with the discontinuities in the gradient of the objective function used in maximum *a posteriori* estimation in the sparse coding model (Fig. 3).

RESULTS

We applied our method to short-term Fourier magnitude spectra derived from some synthesised harpsichord music. The basis vectors, the ‘independent components,’ were found to be the *note* spectra, and so the system as a whole, after training, was essentially doing an optimal linear decomposition of the input spectra into a set of atomic note spectra, but where the atomic spectra were themselves learned from the music in an unsupervised fashion.

On an extract of Bach: Partita in A minor, 55 non-zero spectra (columns of \mathbf{A}) were extracted, of which 49 corresponded to notes (Fig. 4). A harmonic relationship between spectral peaks for each note emerged from the analysis, without anything in the model that required harmonics. The estimated sparse sources (\mathbf{s}) gave a ‘‘piano-roll’’ transcription of the music (Fig. 5, Fig. 6), and resynthesized audio rendered a passable (if hesitant) reproduction of the piece.

Finally, we also tested the approach with input \mathbf{x} consisting of 4 consecutive spectrogram frames instead of a single spectrum. As well as basis vectors representing steady-state notes, many of the basis vectors learned by the algorithm were tuned to note onsets (Fig. 7).

DISCUSSION

We believe that this approach to Automatic Music Transcription is unique, in that it is based simply on the assumption that notes are sparse and approximately independent. There is no need to build in any heuristic knowledge from psychophysical observations, as used in e.g. blackboard models [7, 8], or to assume any harmonic structure [9]. Since it is not necessary to build in any spectral profiles or harmonic assumptions, this approach may be particularly applicable to recordings of new or unknown instruments. The sparse coding ICA approach al-

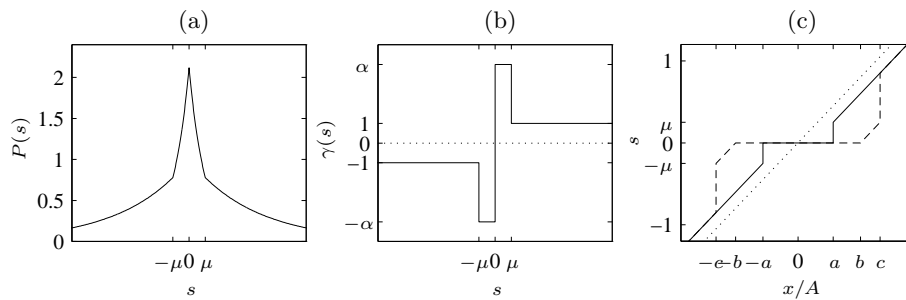


Fig. 2: Approximation to ‘sparsified’ Laplacian, constructed piecewise from exponential segments. (a) Prior, $p(s)$, (b) $\gamma(s) = -(\text{d}/\text{d}s) \log p(s)$, and (c) shrinkage operator. Where the shrinkage operator is multi-valued, the actual value reached depends on the details of the optimisation procedure used to find \hat{s} . The two options are shown as the solid and dashed lines; the dotted line shows the line $\hat{s} = x/A$.

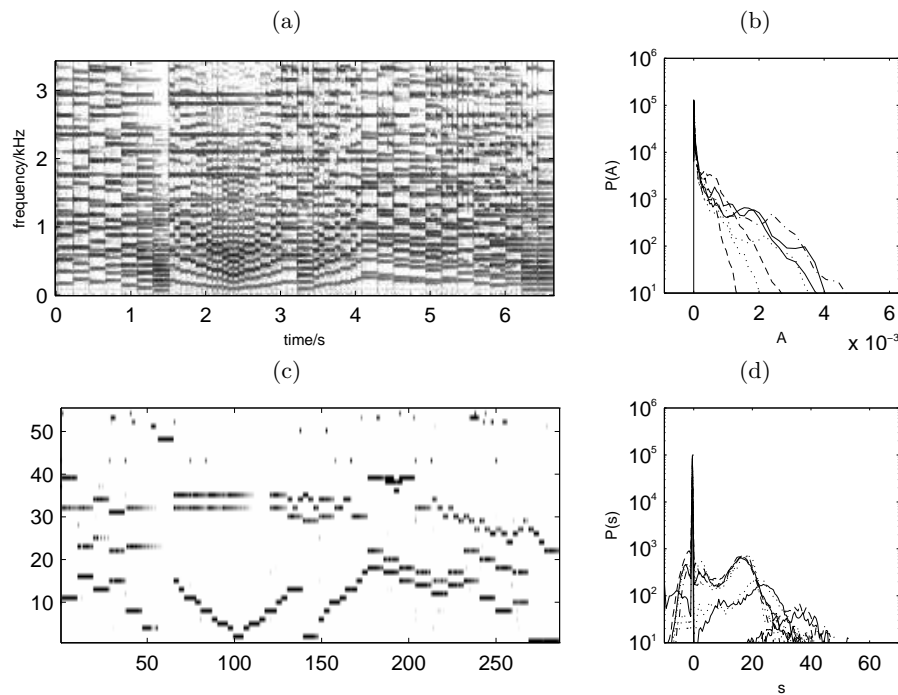


Fig. 5: Sparse coding of an extract of Bach: Partita showing at the top (a) the input spectrogram and (b) histograms of input activity at a selection of frequencies, and at the bottom (c) the output activity and (d) output histograms. The output is visibly sparser than the input. The output histograms also show some interesting structure: their bimodality reflects the on/off nature of the notes present in the music.

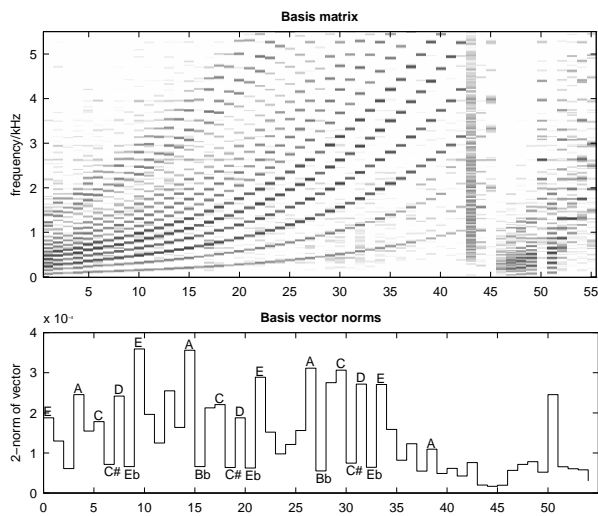


Fig. 4: The 55 non-zero basis vectors obtained from sparse coding synthetic harpsichord music. The lower plot show the lengths (2-norms) of the basis vectors, showing that the basis vectors corresponding to the most structurally important pitches in the key (which is A minor) had the largest norms.

lows us to solve the “chicken and egg” problem of note extraction and spectral analysis together, using a rigorous probabilistic model.

This method could form the basis of an automatic transcription system, and was found to be capable of doing audio to midi conversion even with an extremely simple final stage. However, this was with a very restricted data set, and it remains to be seen how it performs with other types of music played on real instruments. The notes produced by the synthetic harpsichord were very consistent, and real instruments produce much more variable sounds.

We do not expect to build a complete polyphonic transcription system based on a simple linear generative model like this. However, what this approach illustrates is that there is enough structure in music for musically relevant aspects to emerge through unsupervised learning; in this case, it was the independent existences of notes, each with a certain spectral structure.

Given the limitations of the linear generative model, the system performed quite well, but at a high computational cost. It was also somewhat tricky to ob-

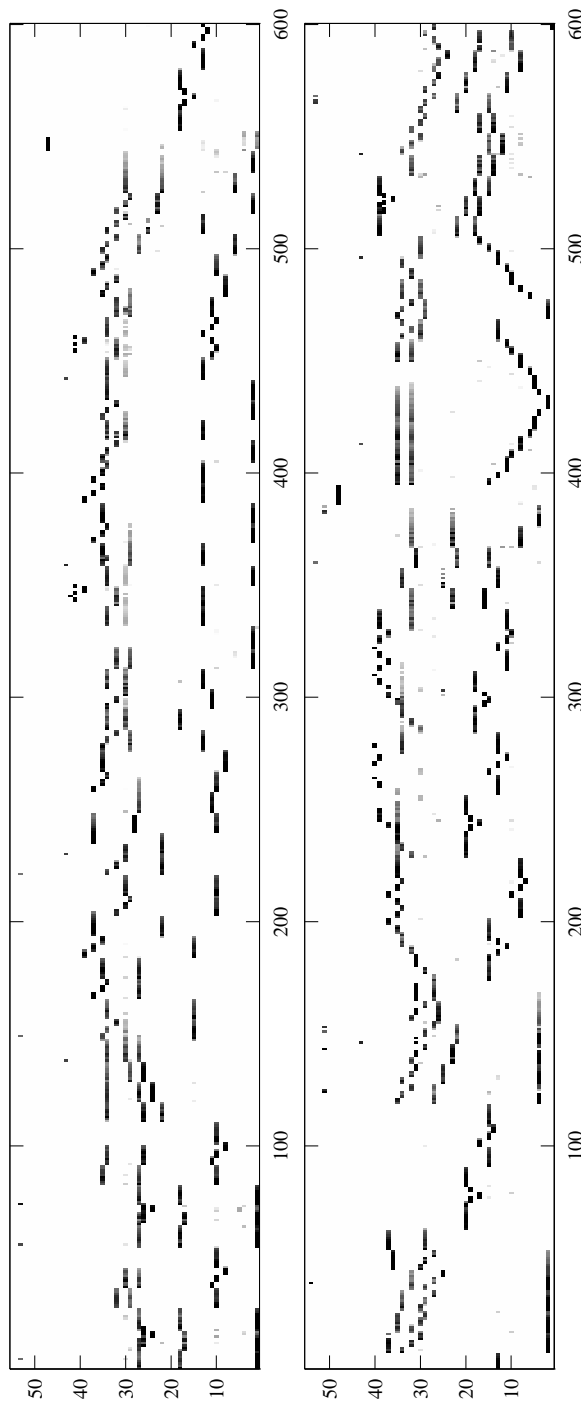


Fig. 6: A longer extract of ‘piano-roll’ output showing how the system performs with varying degrees of polyphony.

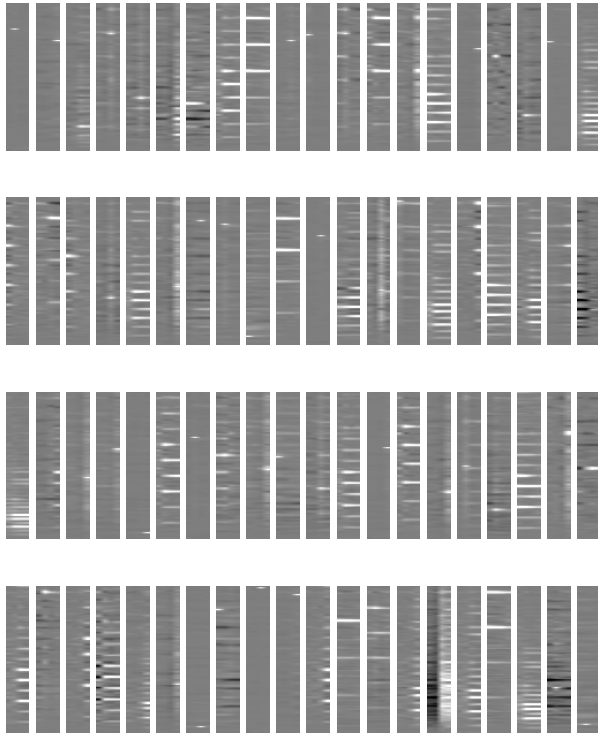


Fig. 7: Basis trained on 2-dimensional strips of spectrogram. Each strip is four pixels wide. Many of the basis vectors appear to be attuned to note onsets, and one (bottom row, sixth from right) seems to represent wide-band onset activity.

tain the appropriate parameters for the prior distribution $p(s)$. In general, it seems that probabilistic inference in this sort of noisy, multiple-cause graphical model is a powerful technique, but cannot be applied to large amounts of data without some restrictions on the graph topology to allow more efficient inference, or preprocessing to reduce the number of variables involved.

We saw that the marginal distributions of the basis vectors were bimodal, and that this was probably related to the binary nature of note activation. It may be possible to formalise this with an explicit mixture model. Olshausen and Millman [10] developed a sparse coder using a mixture-of-Gaussians prior, but again, this is at a high computational cost due to the need to integrate over a multimodal posterior.

The current approach assumes white noise in the spectral domain. Even if a white noise model was appropriate in the time domain, observation of the Fourier magnitude spectra suggests that the ‘noise-like’ activity is not distributed uniformly across the spectrum, but tends to be proportional to signal activity—this is a manifestation of *spectral leakage* and fundamentally, is due to phase information ‘leaking’ into the supposedly phase invariant magnitude spectrum because of the windowing process. This could be modelled as a multiplicative noise process, but since there is generally more activity at the lower end of the spectrum, it would be much simpler, as an initial step, to assume additive, uncorrelated, but non-uniform Gaussian noise, with a smaller variance at high frequencies. This would result in the higher harmonics of musical notes receiving more weight in the sparse coder.

The main difficulty with the use of sparse coding is the computational cost. One way of overcoming this would be to reduce the connectivity of the system considered as a network. It may be possible to do this by using a more efficient method such as ICA, to do an initial phase of redundancy reduction, followed by a topographic organisation to localise the residual dependencies. Then, a sparse coder could be implemented with less than full connectivity. Hoyer and Hyvärinen [11] have made steps in this direction but their algorithm does not exploit the expected locality of the residual dependencies.

The initial experiments with two-dimensional spec-

rogram patches suggested that note onsets would be an important feature; the use of wider patches should result in more reliable note detection by learning the typical profile of a note in time as well as frequency. The main obstacle to this is the large amount of data involved in representing a patch. Unlike visual images, these spectrogram ‘auditory images’ are not translation invariant in the frequency direction and hence the patches used as input must span the whole frequency range.

CONCLUSIONS

We considered the problem of automatic polyphonic music transcription to be one of identifying ‘independent’ notes that mix and give rise to observed spectra. With this motivation, we used the techniques of Independent Component Analysis (ICA) and sparse coding (where extracted parameters are mostly ‘off’) to analyse spectra of polyphonic music.

We found that we could indeed separate out notes from simple polyphonic music played on synthesized harpsichord, and produce a piano-roll description, which could be resynthesized into a passable (if hesitant) rendition of the piece. It remains to be seen how it performs with other types of music played on real instruments.

We are currently working on developments to overcome limitations of this approach. For example, the synthesized harpsichord fits much better to the generative model used here than real-world sounds do. Other instruments with vibrato or significant spectral change over time, requiring perhaps a 2D spectral subspace space to contain their “timbre track”, will require at least a subspace rather than vector in the generative model.

ACKNOWLEDGEMENTS

This work is supported by grant GR/R54620 from the UK Engineering and Physical Sciences Research Council.

REFERENCES

- [1] A Hyvärinen, J Karhunen, and E Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.
- [2] D. J Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.
- [3] 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.
- [4] M. D Plumbley, S. A Abdallah, J. P Bello, M. E Davies, G Monti, and M. B Sandler. Automatic music transcription and audio source separation. *Cybernetics and Systems*, 33(3):603–627, September 2002.
- [5] D. J. C MacKay. Maximum likelihood and covariant algorithms for independent component analysis, 1996. Unpublished manuscript available at <http://wol.ra.phy.cam.ac.uk/mackay>.
- [6] M. S Lewicki and T. J Sejnowski. Learning over-complete representations. *Neural Computation*, 12:337–365, 2000.
- [7] K Kashino, K Nakadai, T Kinoshita, and H Tanaka. Application of Bayesian probability network to music scene analysis. In *Working Notes of the IJCAI-95 Computational Auditory Scene Analysis Workshop*, pages 52–59, Aug 1995.
- [8] D Godsmark and G. J Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27:351–366, 1999.
- [9] A Klapuri. Automatic transcription of music. Msc thesis, Department of Information Technology, Tampere University of Technology, April 1998.
- [10] B. A Olshausen and K. J Millman. Learning sparse codes with a mixture-of-gaussians prior. In S. S A., T. K Leen, and K.-R Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 841–847. MIT Press, 2000.
- [11] P. O Hoyer and A Hyvärinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 42(12):1593–1605, 2002.