

SOURCE EXTRACTION FROM TWO-CHANNEL MIXTURES BY JOINT COSINE PACKET ANALYSIS

Andrew Nesbit, Mike Davies, Mark Plumbley and Mark Sandler

Centre for Digital Music
Department of Electronic Engineering
Queen Mary, University of London
Mile End Road, London, E1 4NS, United Kingdom

email: {Andrew.Nesbit, Mike.Davies, Mark.Plumbley, Mark.Sandler}@elec.qmul.ac.uk

ABSTRACT

This paper describes novel, computationally efficient approaches to source separation of underdetermined instantaneous two-channel mixtures. A best basis algorithm is applied to trees of local cosine bases to determine a sparse transform. We assume that the mixing parameters are known and focus on demixing sources by binary time-frequency masking.

We describe a method for deriving a best local cosine basis from the mixtures by minimising an l^1 norm cost function. This basis is adapted to the input of the masking process.

Then, we investigate how to increase sparsity by adapting local cosine bases to the expected output of a single source instead of to the input mixtures. The heuristically derived cost function maximises the energy of the transform coefficients associated with a particular direction. Experiments on a mixture of four musical instruments are performed, and results are compared. It is shown that local cosine bases can give better results than fixed-basis representations.

1. INTRODUCTION

Blind source separation is a broad term which describes a set of techniques which aim to estimate individual *sources* from a number of observed *mixtures* of those source signals. Cases in which the number of mixtures is greater than, or equal to, the number of sources are called (*over-*)*determined*. These cases have been well studied, commonly through the application of independent component analysis (ICA) [5].

In contrast to the overdetermined case, *underdetermined* blind source separation considers cases in which there are more sources than mixtures. In this work, we deal with underdetermined, *instantaneous, two-channel* mixtures of $n > 2$ time-domain audio sources:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \end{pmatrix} \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix} \quad (1)$$

where s_j is the j th source, x_i is the i th mixture, a_{ij} is the positive real amplitude (mixing parameter) of the j th source in the i th mixture (observation), and $1 \leq j \leq n$ and $i = 1, 2$.

A mixture model given by Equation 1 may represent, for example, a music signal with a conventional “pan-potted

stereo” mixing method. Indeed, our experiments will concentrate on mixtures in which each source is a musical instrument.

The blind source separation problem may be split conceptually into two successive subproblems [10]. *Identification* is the first, and involves determining the mixing parameters a_{ij} . Once the mixing parameters are known, the second subproblem, *filtering*, involves separating each source s_j from the mixtures to yield an estimated source \hat{s}_j . The degenerate unmixing estimation technique (DUET) [11] provides an example of this partitioning into subproblems: The mixing parameters are identified by constructing a histogram from which the values may be read. Once this has been done, the sources are estimated by time-frequency masking (Section 2). DUET is one method which may be applied to mixtures in the form of Equation 1. (DUET was originally developed for blind source separation of *anechoic* mixtures, in which the mixture may include relative delays as well as relative amplitude gains. Instantaneous mixtures are a special case of anechoic mixtures—simply set all relative delays to zero—and so DUET may be used as stated.)

The current paper concentrates on the filtering phase. We assume that the mixing parameters are known or have been estimated. As such, these methods are equally applicable to other *non-blind* scenarios, in which the mixing parameters are known.

Section 2 describes filtering by time-frequency masking. Subsequently, Section 3 describes computationally efficient methods for adapting time-frequency representations to try to match the time-varying signal characteristics better. These methods apply the best basis algorithm [3] to a tree of local cosine bases. In Section 4, we compare and contrast different techniques and representations: The short-time Fourier transform (STFT, which lies at the heart of the filtering stage in DUET), the modified discrete cosine transform (MDCT), a best local cosine basis derived from a mixture of the sources, and a best local cosine basis which sparsifies the representation at the output of the filtering process.

2. TIME-FREQUENCY MASKING

Consider a real- or complex-valued linear transform T applied to the mixtures x_1 and x_2 in Equation 1. This gives transformed mixtures $\tilde{x}_1 = Tx_1$ and $\tilde{x}_2 = Tx_2$ with the same mixing structure as Equation 1.

A *sparse* transform has most coefficients very close to zero and only a few large coefficients. This will represent the mixtures in the desired way, such that the sources have (approximately) disjoint support in the transform domain.

Andrew Nesbit is supported by a research grant from the Semantic Interaction with Music Audio Contents (SIMAC) project (EU-FP6-IST-507142), and by the Department of Electronic Engineering, Queen Mary, University of London.

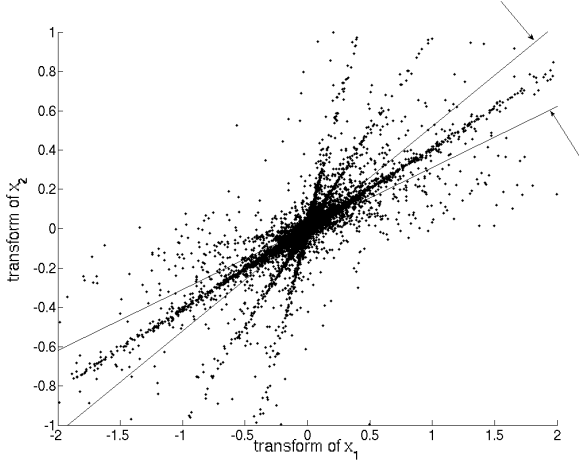


Figure 1: Scatterplot of sparse transforms \tilde{x}_1 and \tilde{x}_2 (MDCT). The arrows point to the upper and lower bounds of the symmetric threshold. It can be seen that between these bounds are points clustered around a straight line, whose gradient indicates the ratio of the mixing parameters for one source.

Individual sources can be estimated from \tilde{x}_1 and \tilde{x}_2 by constructing *binary time-frequency masks*. This assumes that at each point in the transform domain, energy from at most one source dominates, that is, it assumes a sparse transform. Then, the mask can be used to filter (extract) the coefficients belonging to a particular source.

The masks are constructed as follows. The ratio of mixing coefficients for the j th source can be interpreted as an angle

$$\theta_j = \arctan\left(\frac{a_{2j}}{a_{1j}}\right) \quad (2)$$

where the inverse tangent is computed in the first quadrant of the plane. If the transform $Ts_j = \tilde{s}_j$ of the j th source is sparse and real-valued, then its coefficients tend to cluster along the line defined by θ_j . A scatterplot of \tilde{x}_1 and \tilde{x}_2 shows that the ratio of the mixing parameters for each source may be found by visual inspection (Figure 1). For simplicity we will consider manually defined symmetric thresholds u , although such restrictions are not generally necessary. A binary time-frequency mask $M_{\theta_j, u}$ then captures the coefficients which fall “close” to the line corresponding to θ_j and discards all others:

$$M_{\theta_j, u} = \begin{cases} 1 & \text{if } \theta_j - \frac{u}{2} < \arctan\left(\frac{\tilde{x}_2}{\tilde{x}_1}\right) < \theta_j + \frac{u}{2} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

This mask $M_{\theta_j, u}$ estimates the mixture coefficients carrying most of the energy for source j , for the given symmetric threshold u .

Once the masks have been constructed, they can be used to determine $\hat{\tilde{s}}_j$, an estimation of the transformed j th source:

$$\hat{\tilde{s}}_j = M_{\theta_j, u} \cdot (\tilde{x}_1 \cos \theta_j + \tilde{x}_2 \sin \theta_j) \quad (4)$$

We apply the mask to this linear combination of \tilde{x}_1 and \tilde{x}_2 because it allows extraction of sources which occur entirely in

one mixture. This is in contrast to techniques such as DUET which apply the mask to \tilde{x}_1 only, which can be problematic if a source is mostly or entirely represented in \tilde{x}_2 .

The use of time-frequency masking, in various forms, has been successfully applied to instantaneous mixtures in the MDCT domain [4], instantaneous mixtures in the STFT domain [1] and anechoic mixtures in the STFT domain [11]. An initial study of the effects of different sparse transforms appears in [8].

3. ADAPTING THE TIME-FREQUENCY REPRESENTATION

Transforms such as the STFT and MDCT have constant-length analysis windows and fixed bases for the entire duration of the signal. This gives fixed time-frequency resolution.

In order to better match the time-varying characteristics of the mixtures and sources, in the current work, we adapt the bases and window lengths to the input signals by constructing a transform whose basis functions are adaptively selected local cosine bases. This allows longer windows over intervals requiring fine frequency resolution (at the expense of coarser time resolution), and shorter windows over intervals with broadband frequency content (giving finer time resolution). If a signal is decomposed in such a basis, then we anticipate that its transform may be sparser than transforms which decompose the signal in a fixed basis.

3.1 Trees of Local Cosine Bases

In our method, we use a *cosine packet tree* composed of local cosine bases, which are briefly introduced here. For more details on these structures, see [6].

The basis functions of the linear transform are defined over dyadic-length (powers of 2) intervals $[c_{pd}, c_{p+1, d}]$. The endpoints are given by

$$c_{pd} = 2^{-d}Np - \frac{1}{2} \quad (5)$$

where N is the length of the (time-domain) signal, and defines a binary tree structure where the depth of a node is given by d up to a maximum depth D ($0 \leq d \leq D$), and the position of a node at level d is given by p ($0 \leq p < 2^d$). A pair of indices (p, d) corresponds to a node in the tree, and identifies a signal space \mathbf{W}_d^p spanned by an orthogonal *local cosine basis*:

$$\left\{ w_{pd}[n] \sqrt{\frac{2}{2^{-d}N}} \cos \left[\pi \left(k + \frac{1}{2} \right) \frac{n - c_{pd}}{2^{-d}N} \right] \right\} \quad (6)$$

where $0 \leq k < 2^{-d}N$ indexes the functions in the basis. The smooth window w_{pd} localises the basis functions over a dyadic interval $[c_{pd}, c_{p+1, d}]$ and partly overlaps with its immediately adjacent windows $w_{p-1, d}$ and $w_{p+1, d}$. Furthermore, the window must satisfy special properties [6]. Figure 2 is an example tree of local cosine bases.

Each signal space \mathbf{W}_d^p is orthogonal to \mathbf{W}_d^q whenever $p \neq q$, and $\mathbf{W}_j^p = \mathbf{W}_{j+1}^{2p} \oplus \mathbf{W}_{j+1}^{2p+1}$. This means that the union of the bases corresponding to the children of any node comprise an orthogonal basis of the space corresponding to that node. (The length- N signal being analysed is in the signal space \mathbf{W}_0^0 .) Bases occurring deeper in the tree correspond to

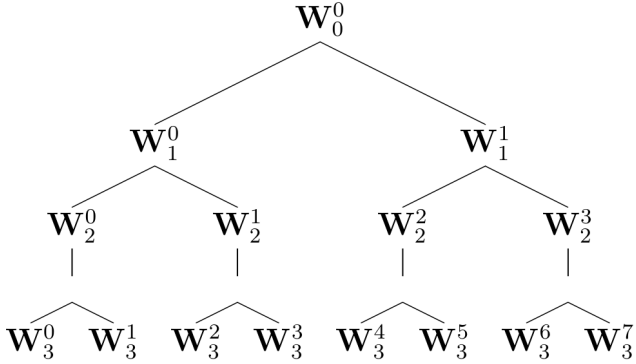


Figure 2: Tree of signal spaces spanned by local cosine bases. Deeper nodes correspond to more localised basis functions (shorter analysis windows).

shorter time intervals and so are better for representing sections of the signal with highly time-varying characteristics; bases occurring higher in the tree are better for representing sections which need better frequency resolution at the cost of coarser time resolution. This tree structure offers a computationally efficient method for computing a good basis.

3.2 Selecting the Best Basis

A tree of local cosine bases describes many possible orthogonal bases for representing a signal. A complete binary tree provides a *dictionary* of orthogonal bases from which the optimal basis can be adaptively chosen to represent the signal. The l^1 cost of representing a length- N signal x in the basis $B = \{b_m\}$ is given by

$$C(x, B) = \sum_{m=1}^N \frac{|\langle x, b_m \rangle|}{\|x\|} \quad (7)$$

and provides a convenient measure of sparsity [2]. The *best basis* is the one which minimises this cost. The computationally efficient Coifman-Wickerhauser algorithm takes advantage of the binary structure and determines the best basis in $O(N \log_2 N)$ time [3].

Figure 3 depicts a tree of local cosine bases adapted to an audio recording of a glockenspiel showing the original time-domain signal partitioned into dyadic intervals, each of which correspond to a basis in the tree. The bars of the glockenspiel are struck in the first half of the signal and so relatively short basis functions have been adapted to capture the transients. The notes all ring out and decay in the second half of the signal; here, long basis functions have been chosen because the signal varies relatively slowly over time.

3.3 Adapting to the Input

We consider two natural ways by which the local cosine basis may be adapted. The first method attempts to maximise the sparsity of the average of the two time-domain mixtures \tilde{x}_1 and \tilde{x}_2

$$x_a = \frac{1}{2}(x_1 + x_2) \quad (8)$$

by minimising the l^1 cost described in Section 3.2. This method will be referred to as *CP1*. Results for CP1, are given

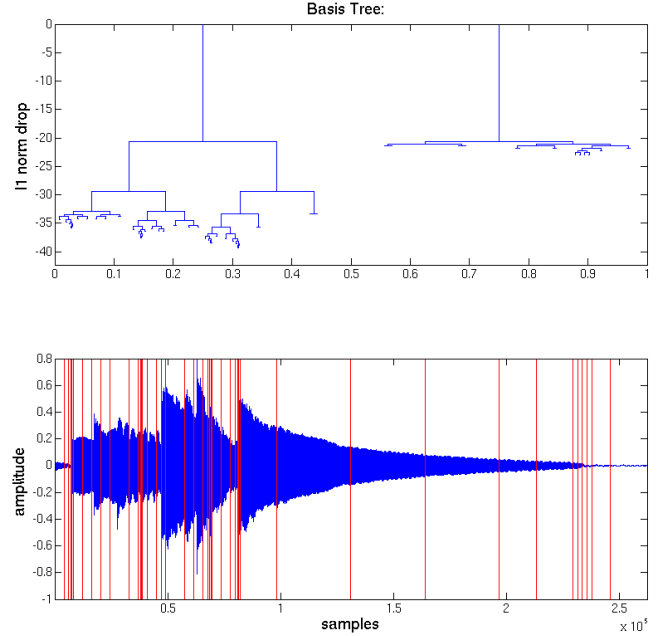


Figure 3: Glockenspiel. Upper plot is a local cosine best basis tree computed by minimising the l^1 norm to a maximum depth $D = 10$. Lower plot is the time-domain signal partitioned into intervals; the width of each interval is determined by the depth of the corresponding basis in the tree.

in Section 4, indicating some improvement over fixed-basis methods.

3.4 Adapting to a Single Source

One issue with the CP1 method is that it models mixtures of the sources rather than the sources themselves. For example, consider a music signal: If a percussive note with broadband frequency content and a tonal note with fine frequency content occur at same time, then the basis selected to cover that time interval may not be particularly well adapted to either. Furthermore, the basis may not adapt to transients well as tonal content tends to have more energy.

To overcome this possible limitation, we propose to adapt one basis to the expected output of the time-frequency mask for each source. This will select a basis for each source with the intention that each such basis will capture the time-frequency structures of that source better than the basis determined by CP1.

A heuristically motivated cost function is now developed, based on this intuitive reasoning. Whereas the CP1 method minimises the l^1 cost of expressing a signal in some basis, here we maximise the energy of the local cosine coefficients associated with a particular source angle θ_j . The mixing parameters for a given source are known; the representation which has greatest sparsity for this source has local cosine coefficients clustered around these mixing parameters. By selecting a basis which maximises the energy of coefficients that cluster around θ_j we would expect that a sparse representation will be generated.

Therefore we use the following cost function:

$$C(x_1, x_2, B, \theta_j, u) = - \sum_{m=1}^N \Lambda_{\theta_j, u} \langle (x_1 \cos \theta_j + x_2 \sin \theta_j), b_m \rangle^2 \quad (9)$$

where

$$\Lambda_{\theta_j, u} = \begin{cases} 1 & \text{if } \theta_j - \frac{u}{2} < \arctan\left(\frac{\langle x_2, b_m \rangle}{\langle x_1, b_m \rangle}\right) < \theta_j + \frac{u}{2} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

and $B = \{b_m\}$ is a basis from the dictionary of bases derived from the complete local cosine tree. The binary mask $\Lambda_{\theta_j, u}$ has a similar form to Equation 3, but instead of masking a transformed mixture \tilde{x}_1 or \tilde{x}_2 , it masks local cosine coefficients in the basis B . Again, the fast tree-searching algorithm of Coifman and Wickerhauser finds the best basis corresponding to this cost function. For the rest of this paper, this method will be referred to as *CP2*.

This method learns an overcomplete dictionary of bases adapted to different sources. In this sense, it may be considered to be equivalent to techniques based on, for example, the matching pursuit algorithm [7]. However, the advantage of this method stems from the representation of local cosine bases as tree structures which allows us to apply the fast tree-searching algorithm to determine the best basis.

4. RESULTS

A stereo mixture of four musical instrument sources¹ were used to test the source extraction methods. The sources are excerpts from “real-world” multitracked music, are harmonically related and so overlapping partial frequencies were expected.

Each of the four sources s_1, \dots, s_4 corresponds to the sounds produced by one instrument, where s_1 is percussion, s_2 is acoustic guitar (“guitar 1”), s_3 is male vocal and s_4 is another acoustic guitar (“guitar 2”). All sources were converted to a sample resolution of 16 bits and a sample rate of 22050 Hz for 2^{18} samples (11.9 s).

The mixtures x_1 and x_2 were generated by instantaneously mixing as follows:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.90 & 0.71 & 0.50 & 0.28 \\ 0.09 & 0.29 & 0.50 & 0.72 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix} \quad (11)$$

The resulting mixture is a realistic simulation of a pan-potted stereo downmix, and we have found this to be relatively challenging for standard signal extraction methods.

4.1 Measuring Performance

For each estimated source, we wish to make numerical evaluations of the contribution of unwanted sources (interference) and the distortion due solely to the separation process (artifacts). We do this by measuring the *Source to Interference Ratio (SIR)* and the *Source to Artifacts Ratio (SAR)*. Furthermore, in order to simplify direct comparisons, the *Source to*

source	transform	K	D	SIR	SAR	SDR
s_1	STFT	512	-	30.31	3.47	3.46
	MDCT	512	-	24.95	1.23	1.20
	CP1	-	9	25.48	2.17	2.14
	CP2	-	8	26.15	3.38	3.35
s_2	STFT	1024	-	30.13	7.03	7.01
	MDCT	1024	-	35.87	7.07	7.06
	CP1	-	6	31.30	7.60	7.57
	CP2	-	8	28.24	9.28	9.22
s_3	STFT	1024	-	36.48	-0.23	-0.23
	MDCT	1024	-	40.94	0.77	0.77
	CP1	-	7	36.50	3.16	3.16
	CP2	-	8	32.84	4.22	4.21
s_4	STFT	1024	-	29.20	4.65	4.63
	MDCT	1024	-	29.70	4.12	4.10
	CP1	-	7	30.46	6.44	6.41
	CP2	-	9	24.64	6.91	6.82

Table 1: Results of source extraction. The STFT and MDCT block sizes are specified by K . The maximum depths of local cosine trees are given by D .

Distortion Ratio (SDR) is computed; this combines both the SIR and SAR into a single numerical measure of total relative distortion. Methods for computing these measurement criteria are explained in detail in [9]. Whenever these measures are used, they will be stated in units of decibels (dB).

4.2 Experiments

Time-frequency masks were constructed and applied to the mixture channels x_1 and x_2 represented by the following transforms:

- **STFT** at block sizes K , Hamming-windowed, and with $K/2$ overlap on consecutive blocks. This is essentially the filtering component of DUET [11]. The STFT is a complex-valued transform, so the binary masks were determined based on the magnitude of the STFT.
- **MDCT** critically sampled, with various block sizes K
- **CP1** at various maximum tree search depths D
- **CP2** at various maximum tree search depths D

All experiments used the same symmetric masking threshold $u = 0.17$. This value u was determined experimentally to give good results overall and ensured that masks for adjacent sources do not overlap.

Relatively short and medium block sizes K were chosen for STFT and MDCT testing: $K = 256, 512, 1024$. The reason is that informal listening tests reveal this captures transients well for these transforms, and longer frame sizes seriously degrade note attacks.

The depth d of a node in a local cosine tree corresponds to basis functions of length 2^{18-d} (the example mixture has length 2^{18}). The maximum tree depths tested were $D = 10$, so that the smallest basis functions have length 256 (equal to the smallest K).

4.3 Discussion

It is clear from Table 1 that the effects of the artifacts dominate interference, since the SAR values are typically some 20 dB or more lower than SIR. This is not too surprising, since these methods are based on binary masking, and we would expect the masking process to introduce some artifacts.

¹Another Dreamer, *Personalized Perfection*, 2004. Source files available online <http://anotherdreamer.infobeing.net/personalizedperfection.htm> subject to the Creative Commons Attribution-NonCommercial 1.0 license.

For STFT and MDCT, the best estimation of the percussion source s_1 is given by a shorter analysis window than those analysing the other (tonal) source estimates. Similarly, CP1 gives best results on s_1 when the algorithm is allowed to search deeper in the tree than for the other sources (depth 9 rather than 6 or 7). This suggests that shorter basis functions are indeed preferred for representing signals with broadband noisy content and that longer basis functions are preferred for representing tonal structures.

The results presented generally confirm that the sparser local cosine basis representations (CP1 and CP2) typically reduce distortion (improve SDR) when compared to STFT and MDCT. With the exception of the percussion source s_1 , where STFT gives the best result, Table 1 shows that CP1 gives better SDR than both STFT and MDCT, and in turn CP2 gives better SDR than CP1. CP2 appears to reduce artifacts (increase SAR) at the expense of some increase in interference (reducing SIR) in comparison to CP1, resulting in an overall improvement in SDR (since the effect of artifacts dominates).

Even on source s_1 , informal listening tests have revealed that source extraction based on the STFT suffers from objectionable “pipe noise” artifacts, while the other methods do not exhibit such artifacts to the same degree. For future work we propose to confirm this effect using more formal listening tests.

5. FURTHER WORK

Results have shown that adapting a local cosine basis to the output can give good results. However, the energy-based cost function (Equation 9) is derived from heuristic reasoning and intuition. It may be the case that a more subtle cost function is required to represent the estimated source more sparsely. In particular, the current energy-based cost function considers only coefficients of the estimated source without regarding the coefficients of the other sources. Therefore, the next step is to manually examine the basis functions which are adapted to a particular source direction and determine the most suitable cost function for this sort of joint adaptation of local cosine bases.

Similarly, the CP1 technique (Section 3.3) minimises the l^1 cost of \bar{x}_a , the average of the input mixtures. Alternatively, one could minimise the average l^1 cost of both \bar{x}_1 and \bar{x}_2 .

The tree structure described in Section 3 is not necessarily tied to local cosine bases. It should be possible to apply a tree-like framework to other transforms, such as the STFT. This would give access to phase information so the framework could be used to separate anechoic mixtures.

All techniques in this paper assume the mixing parameters are already known (the non-blind case). In practical situations this information may not be available and so the mixing structure would need to be identified. It would be useful to study the sensitivity of the sparse representations to the accuracy of the mixing parameter estimates.

Finally, the performance measures, SIR, SAR and SDR, may not correspond well to a subjective human assessment of separation performance. Informal listening tests show that each representation imparts a noticeably different timbre to the extracted sources. Therefore, we believe that listening tests would give a more meaningful, practical measure of separation performance.

6. CONCLUSIONS

This paper has described novel research on local cosine packet representations for source separation of two-channel instantaneous mixtures. One method is CP1 (Section 3.3) and the other is CP2 (Section 3.4).

The advantage of the CP1 and CP2 techniques is that they adapt a basis to match the time-varying characteristics of the signal. CP2 takes this idea one step further and adapts the basis to the time-frequency mask used for separating each separate source. Searching a tree of local cosine bases is fast and gives promising results. This is in contrast to representations with fixed bases, which are not adapted to the signal under analysis. On a stereo mixture of percussion, voice and two guitars, we have shown that our new methods CP1 and CP2 improve source extraction performance for all but one of the four sources. For future work we propose to investigate further possible improvements through the use of alternative sparsifying cost functions.

REFERENCES

- [1] D. Barry, B. Lawlor, and E. Coyle. Real-time sound source separation: Azimuth discrimination and resynthesis. In *Proceedings of the AES 117th Convention*, San Francisco, CA, USA, 28–31 October 2004.
- [2] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
- [3] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions of Information Theory*, 38(2):713–718, March 1992.
- [4] M. Davies and N. Mitianoudis. Simple mixture model for sparse overcomplete ICA. *IEE Proceedings on Vision, Image and Signal Processing*, 151(1):35–43, February 2004.
- [5] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, 2001.
- [6] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, second edition, 1999.
- [7] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, December 1993.
- [8] V. Y. F. Tan and C. Févotte. A study of the effect of source sparsity for various transforms on blind audio source separation performance. In *Proceedings of the Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS’05)*, Rennes, France, 16–18 November 2005.
- [9] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Speech and Audio Processing*, 2004. Preprint, to appear.
- [10] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies. Blind audio source separation. Technical Report C4DM-TR-05-01, Centre for Digital Music, Queen Mary, University of London, 24 November 2005. Available online http://www.elec.qmul.ac.uk/people/emmanuelv/VincentEtAl05_bass_tutorial.pdf.
- [11] Ö. Yılmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004.