ORACLE ESTIMATION OF ADAPTIVE COSINE PACKET TRANSFORMS FOR UNDERDETERMINED AUDIO SOURCE SEPARATION

Andrew Nesbit and Mark D. Plumbley

Department of Electronic Engineering
Queen Mary, University of London
Mile End Road, London, E1 4NS, UK
{andrew.nesbit,mark.plumbley}@elec.qmul.ac.uk

ABSTRACT

We address the problem of instantaneous, underdetermined audio source separation by time-frequency masking. Using oracle estimators, we determine experimental upper performance bounds, by assuming that we have reference sources available, and that we know, or have estimated, the mixing structure. Oracle estimation of four musical sources from two-channel mixtures demonstrates a potential for SDR improvements of up to 12.7 dB, compared to semi-blind methods. We also show that using adaptive cosine packet transforms, rather than fixed-basis STFTs, can improve performance by up to 2.2 dB. Finally, by allowing more than one non-zero source coefficient per time-frequency index, improvements of up to 7.7 dB could be possible.

Index Terms— Audio systems, Cosine transforms, Time-frequency analysis

1. INTRODUCTION

The aim of *blind source separation* (BSS) is to recover a set of individual *sources* from an observed *mixture* of those sources. Typically we have little or no information on the sources themselves or the mixing process. In this paper we consider the *instantaneous* case for audio signals, which means that neither delays nor reverberations occur in the mixing process. We construct the following model:

$$x = As \tag{1}$$

where $x = x(n) = [x_i(n)]_{1\le i\le I}$ represents an I-channel mixture signal, $s = s(n) = [s_i(n)]_{1\le j\le J}$ is a vector of J source signals, and $0 \le n < N$ is the discrete-time index. The *mixing matrix* is $A = [a_{i,j}]_{1\le i\le I, 1\le j\le J}$, and has real, constant entries.

We are particularly interested in modelling two-channel audio mixtures with more than two sources, because this can approximate the *panpotted mono*[1] mixing technique. Generally, mixtures of this type, that is, when $I < J$, are called *underdetermined*. This case is particularly challenging because even if we know or have estimated A, standard matrix inversion techniques do not give a unique solution.

We take the *time-frequency masking* approach to try to solve the problem [1]. Consider a linear, invertible transform, T, which is specified by its basis, $\mathcal{B} = \{\phi_\gamma^{\mathcal{B}}\}_{\gamma\in\Gamma}$, where Γ is an index set. Then the transform coefficients of any signal $y = y(n)$ are given by $T\{y\}(\gamma) = \langle y, \phi_\gamma^{\mathcal{B}}\rangle = \sum_{n=0}^{N-1} y(n)\phi_\gamma^{\mathcal{B}}(n)$. After transformation by T, Equation (1) assumes the following form:

$$[\langle x_j, \phi_\gamma^{\mathcal{B}}\rangle]_{1\le i\le I} = A[\langle s_j, \phi_\gamma^{\mathcal{B}}\rangle]_{1\le j\le J}. \tag{2}$$

Usually, T represents the short-time Fourier transform (STFT) [1]. More recently, the application of the modified discrete cosine transform (MDCT) [2] and the adaptive cosine packet (CP) transform [3, 4] has been explored.

Such transformations represent the sources in such a way that the number of source coefficients, $\langle s_j, \phi_\gamma^{\mathcal{B}}\rangle$, which are non-zero at each time-frequency index, γ, is relatively small, giving a *sparse* representation [5]. The sources can then be estimated against certain criteria which assign energy from the source coefficients in Equation (2) to the estimated sources. Inversion by T^{-1} follows. In this paper, we assume that A is known or has been estimated. That is, we study the *semi-blind*, rather than the *blind*, case of source separation. In Section 2 we describe how *oracle estimation* techniques can find experimental upper bounds for audio source separation performance [6]. These methods determine the best possible source estimates that time-frequency masking can yield for a particular mixture, according to some performance criterion.

It has previously been shown that using adaptive cosine packet (CP) transforms, which use variably sized windows to try to capture the time-varying nature of the source signals better, has the potential to improve estimation performance compared to using fixed-basis transforms such as the STFT or MDCT [4, 7]. Our aim is to develop more performant oracle estimation methods for mixtures comprised of harmonically overlapping sources, such as music, under the relaxed assumption that *one or more* non-zero sources are allowed per time-frequency index (see Section 2). This extends previous work on oracle estimation of adaptive CP transforms, which allow only one active source per γ [7]. In Section 4, we compare results between semi-blind and oracle estimation, and show that the potential improvements are significant.

2. TIME-FREQUENCY MASKING

Denote by J'_γ the assumed number of active (non-zero) source coefficients at γ. Then $\mathcal{J}_\gamma = \{j : \langle s_j, \phi_\gamma^{\mathcal{B}}\rangle \neq 0\}$ is the set of all J'_γ sources contributing to $[\langle x_i, \phi_\gamma^{\mathcal{B}}\rangle]_{1\le i\le I}$, and is called the *local activity pattern* at γ. Our mixing model then takes the following form

$$[\langle x_i, \phi_\gamma^{\mathcal{B}}\rangle]_{1\le i\le I} = A_{\mathcal{J}_\gamma}[\langle s_j, \phi_\gamma^{\mathcal{B}}\rangle]_{j\in\mathcal{J}_\gamma} \tag{3}$$

where $A_{\mathcal{J}_\gamma}$ is the $I \times J'_\gamma$ submatrix of A formed by taking columns A_j, and $[\langle s_j, \phi_\gamma^{\mathcal{B}}\rangle]_{j\in\mathcal{J}_\gamma}$ is formed by taking rows of $[\langle s_j, \phi_\gamma^{\mathcal{B}}\rangle]_{1\le j\le J}$,

Andrew Nesbit has been supported by EPSRC Grant EP/E045245/1, and by the European Commission through the SIMAC project (number of contract: 507142; key action IST-2.3.1.7 Semantic-based knowledge systems).

[1] In this context, the word *mono* describes the sources, not the mixtures.

whenever $j \in \mathcal{J}_\gamma$. The *global activity pattern* is given by $\mathcal{J} = \{\mathcal{J}_\gamma\}_{\gamma\in\Gamma}$.

Let us assume that A is known, and that the number of active sources at any time-frequency index is less than or equal to the number of mixtures ($J'_\gamma \leq I$). Then Equation (3) reduces to an (over-)determined mixture, and estimation of the sources is then possible according to the following equation [3]:

$$\begin{cases} \langle \hat{s}_j, \phi^{\mathcal{B}}_\gamma \rangle = 0 & \text{if } j \notin \widehat{\mathcal{J}}_\gamma \\ [\langle \hat{s}_j, \phi^{\mathcal{B}}_\gamma \rangle]_{j\in\widehat{\mathcal{J}}_\gamma} = A^{+}_{\widehat{\mathcal{J}}_\gamma} [\langle x_i, \phi^{\mathcal{B}}_\gamma \rangle]_{1\leq i\leq I} & \text{otherwise} \end{cases} \quad (4)$$

where $\widehat{\mathcal{J}}_\gamma$ is an estimate of $\mathcal{J}_\gamma$ and $A^{+}_{\widehat{\mathcal{J}}_\gamma}$ denotes the (Moore-Penrose) pseudoinverse of $A_{\widehat{\mathcal{J}}_\gamma}$. Time frequency masking can then be interpreted as the problem of estimating local activity patterns. (Efficient estimation of $\mathcal{J}_\gamma$ for the $J'_\gamma \leq J$ case is currently an open problem.)

2.1. Semi-blind methods

In the case that $J'_\gamma < I$, we can express the mixture channels as $\langle x_i, \phi^{\mathcal{B}}_\gamma \rangle = \sum_{j\in\mathcal{J}'_\gamma} \langle s_j, \phi^{\mathcal{B}}_\gamma \rangle + \nu^{\mathcal{B}}_\gamma$, where $\nu^{\mathcal{B}}_\gamma$ is the sum of the interfering sources with $j \notin \mathcal{J}_\gamma$, and modelled as Gaussian white noise. This motivates us to estimate $\mathcal{J}_\gamma$ by minimising sums of squared residuals [1, 8]:

$$e(x, A, \mathcal{J}, \mathcal{B}) = \sum_{\gamma\in\Gamma} \sum_{i=1}^{I} \left(\langle x_i, \phi^{\mathcal{B}}_\gamma \rangle - \sum_{j=1}^{J} a_{i,j} \langle \hat{s}_j, \phi^{\mathcal{B}}_\gamma \rangle \right)^2, \quad (5)$$

which depends implicitly on Equation (4), and yields maximum likelihood (ML) estimates of the active sources. An important special case is when $J'_\gamma = 1$. Then we can derive the ML estimates [9, pp. 657–661] and Equation (5) in an equivalent way by modelling the singleton activity set with a uniform prior probability, $\Pr(\mathcal{J}_\gamma = \{j\}) = \frac{1}{J}, 1 \leq j \leq J$ [3].

In the $J'_\gamma \leq I$ case, Equation (5) is no longer motivated in the same way as before, and so we define a new semi-blind criterion. Assume that the coefficients $\langle s_j, \phi^{\mathcal{B}}_\gamma \rangle$ follow a Laplacian prior distribution, independently and identically for all j and γ. Then we find the maximum a posteriori (MAP) estimate of the sources by minimising

$$e'(x, A, \mathcal{J}, \mathcal{B}) = \sum_{j=1}^{J} |\langle \hat{s}_j, \phi^{\mathcal{B}}_\gamma \rangle|. \quad (6)$$

which depends implicitly on Equation (4). This is one of the usual approaches to sparse source separation [5]. However, our experiments in Section 4 consider only cases in which $I = 2$, and so we can determine $\hat{s}$ according to Equation (6) using an efficient, geometrically motivated algorithm [10].

2.2. Oracle methods

Oracle estimators determine those J'_γ and $\widehat{\mathcal{J}}_\gamma$ which give the best possible separation performance for each mixture, by optimising against some criterion [6]. Here we follow the approach and notation of [7]. These techniques require us to know the original sources, s, and the mixing system, A. Oracle estimates allow us to judge the difficulty of separating the sources from a given mixture, and to gain insight into the upper performance bounds of our class of separation algorithms, subject to $J'_\gamma \leq I$. Because their computation depends on knowing the reference source signals, oracle estimators are very useful for evaluating algorithms, rather than for practical (semi-)blind source separation.

The oracle estimate of s is that $\hat{s}$ which minimises a distortion measure such as

$$\mathbf{d}(s, x, A, \mathcal{J}, \mathcal{B}) = \sum_{n=0}^{N-1} \sum_{j=1}^{J} (\hat{s}_j(n) - s_j(n))^2. \quad (7)$$

The advantages of defining $\mathbf{d}$ in this way are that (a) minimising it is equivalent to maximising the signal to distortion ratio (SDR)

$$\text{SDR} = 10 \log_{10} \frac{\sum_{n=0}^{N-1} \sum_{j=1}^{J} (s_j(n))^2}{\mathbf{d}(s, x, A, \mathcal{J}, \mathcal{B})}, \quad (8)$$

with which we shall evaluate the all methods; and that (b) it satisfies additivity constraints required for computing oracle bases (see Section 3).

2.3. Oracle masks for orthonormal transforms

For signals represented by an orthonormal transform, Equation (7) is equal to the following [7]:

$$\mathbf{d}(s, x, A, \mathcal{J}, \mathcal{B}) = \sum_{\gamma\in\Gamma} \sum_{j=1}^{J} \left(\langle \hat{s}_j, \phi^{\mathcal{B}}_\gamma \rangle - \langle s_j, \phi^{\mathcal{B}}_\gamma \rangle \right)^2. \quad (9)$$

It is clear that minimising $\mathbf{d}(s, x, A, \mathcal{J}, \mathcal{B})$ is equivalent to minimising at each γ independently, by computing oracle local activity patterns:

$$\widehat{\mathcal{J}}^{\text{ora}}_\gamma = \underset{\mathcal{J}_\gamma\in\mathcal{P}_\gamma}{\arg\min} \sum_{j=1}^{J} \left(\langle \hat{s}_j, \phi^{\mathcal{B}}_\gamma \rangle - \langle s_j, \phi^{\mathcal{B}}_\gamma \rangle \right)^2 \quad (10)$$

where $\langle \hat{s}_j, \phi^{\mathcal{B}}_\gamma \rangle$ on the right hand side is given by Equation (4), and $\mathcal{P}_\gamma$ the set of all possible activity patterns, subject to J'_γ. If J'_γ is small then an exhaustive search over all $\widehat{\mathcal{J}}_\gamma \in \mathcal{P}_\gamma$ is computationally feasible.

2.4. Oracle masks for non-orthogonal transforms

The STFT is often used in time-frequency masking, but because it is non-orthogonal, the equivalence of Equations (7) and (9) no longer holds. The optimal oracle activity patterns must be determined jointly by a full combinatorial search over all time-frequency indices, γ. This is computationally infeasible for any realistic signal, so we can only compute *near-optimal* oracle activity patterns by minimising $\mathbf{d}(s, x, A, \mathcal{J}, \mathcal{B})$ at each γ separately, according to Equations (9) and (10).

3. ADAPTIVE LOCAL COSINE BASES

Adaptive cosine packet (CP) transforms partition the signal with overlapping sine windows of variable length [11]. Such an orthonormal decomposition is often sparser than fixed-basis representations (such as STFT or MDCT), because it gives 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. It has been shown that, in some cases, this can improve separation performance [4, 7].

Consider a *dictionary* of possible CP bases, $\mathcal{D} = \{\mathcal{B}\}$. We represent $\mathcal{D}$ as a complete binary tree with depth D. Each of the $2^{D+1} - 1$ nodes represents a block in the windowed signal, and the shortest possible block size is given by $K_{\text{min}} = 2^{-D}N$, corresponding to a node at depth D. The aim is to choose the basis which is best adapted to the time-varying nature of the signal, by minimising some

specified criterion or cost function (Equations (5), (6), and (9)). Such a basis, $\mathcal{B}$, is called the *best orthogonal basis*. Fortunately there exist efficient algorithms which take advantage of the tree representation of $\mathcal{D}$ and determine the best basis in $O(N \log_2 N)$ operations, and an exhaustive search is not necessary [12].

There are conceptually two steps in using adaptive CP transforms for time-frequency masking. Firstly, we determine the best basis $\mathcal{B}$ in which to decompose the signal, and set $\mathcal{B}^{\text{sb}} = \mathcal{B}$ or $\mathcal{B}^{\text{ora}} = \mathcal{B}$. This involves computing all possible $\mathcal{J}$. Secondly, we estimate the global activity pattern, $\mathcal{J}$, in the best basis $\mathcal{B}$, and set $\widehat{\mathcal{J}}^{\text{sb}} = \widehat{\mathcal{J}}$ or $\widehat{\mathcal{J}}^{\text{ora}} = \widehat{\mathcal{J}}$ accordingly. This choice of $\widehat{\mathcal{J}}$ minimises the cost of representing s in the best basis so found.

The best basis in a semi-blind context is estimated according to the same criterion as the semi-blind activity patterns. As we stated above, this is because we want to choose the basis which determines the semi-blind activity pattern with minimum cost. We select the basis which minimises the cost given by Equation (5) for the $J'_\gamma < I$ case, and Equation (6) for the $J'_\gamma = I$ case. These cost functions satisfy additivity constraints required by the best basis algorithm [12]. The resulting basis, $\mathcal{B}^{\text{sb}}$, is optimised for the estimated the semi-blind activity pattern $\widehat{\mathcal{J}}^{\text{sb}}$.

Estimation of the oracle best basis is very similar to that of the semi-blind best basis. The only differences are that the cost to minimise is given by Equation (9) and that s must be known. This basis, $\mathcal{B}^{\text{ora}}$, is optimised for estimation of the oracle activity pattern $\mathcal{J}^{\text{ora}}$.

In previous work, oracle estimators were developed for adaptive CP transforms for *binary masking* ($J'_\gamma = 1$) [7]. It can give good results, especially for speech signals [1], but for realistic mixtures with harmonically related sources, such as music, there will be more than one active active source at (almost) all γ. The formulation here thus allows $J'_\gamma \leq 2$, we extend the $J'_\gamma = 1$ framework. In Section 5, we show that this significantly increases performance.

4. EXPERIMENTS

We test our algorithms on eight different pieces of music by various artists, each one comprised of $J = 4$ sources. As we had access to the original multitracked data, we were able to synthesise instantaneous mixtures, with $I = 2$, to simulate a panpotted mono mixing process. The pitched sources were harmonically related so that overlapping harmonics between different sources were expected. Each source had a sample rate of 44.1 kHz, with a resolution of 16 bits per sample. Extracts of length 2^{17} samples (≈ 3.0 s) were taken from each source.

We generated ten random mixing matrices according to

$$A^{(r)} = \begin{pmatrix} \cos\theta_1^{(r)} & \cos\theta_2^{(r)} & \cos\theta_3^{(r)} & \cos\theta_4^{(r)} \\ \sin\theta_1^{(r)} & \sin\theta_2^{(r)} & \sin\theta_3^{(r)} & \sin\theta_4^{(r)} \end{pmatrix}, \qquad (11)$$

where $1 \leq r \leq 10$. Each $\theta_j^{(r)}$ was selected randomly, independently and uniformly distributed over $[0, \pi/2]$. This makes for $10 \times 8 = 80$ mixtures.

For experiments on semi-blind criteria, we tested the $J'_\gamma = 1$ and $J'_\gamma = 2$ cases. For oracle criteria, we used $J'_\gamma = 1$, $J'_\gamma = 2$ and $J'_\gamma \leq 2$. For fixed-basis transforms, we tested the STFT and MDCT, with block lengths ranging in powers of two from $K = 2^5$ (≈ 0.7 ms) to $K = 2^{17}$ (≈ 3.0 s). For the CP transforms, we tested dictionaries with shortest block size ranging from $K_{\text{min}} = 5$ ($D = 12$) to $K_{\text{min}} = 17$ ($D = 0$), to select the best basis. For each combination of transform (STFT, MDCT or CP), criterion (Equations (5), (6), and (9)), block length (K or K_{min}), and assumed J'_γ, we estimated $80 \times 4 = 320$ sources and computed the mean of their SDR measures. Results are presented in Figure 1 and Table 1.

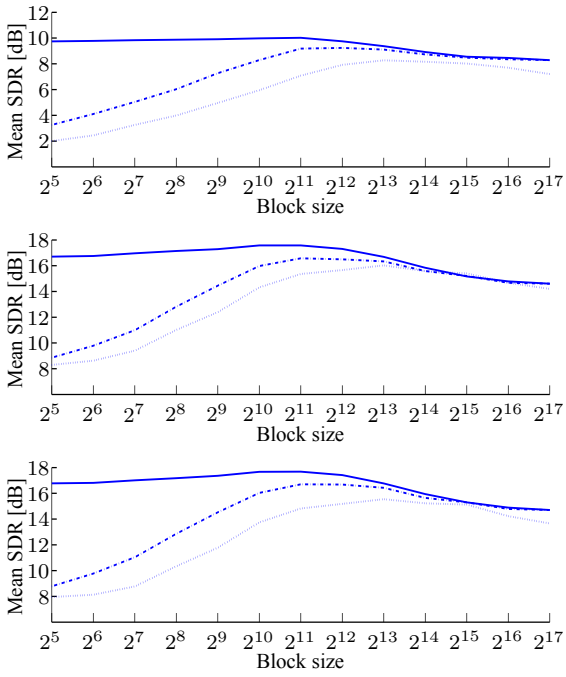

Fig. 1. Mean oracle performance for $J'_\gamma = 1$ (top plot), $J'_\gamma = 2$ (middle plot) and $J'_\gamma \leq 2$ (bottom plot), with STFT (dotted line), MDCT (dashed-dotted line) and CP (solid line). The horizontal axes indicate the block size, K, for STFT and MDCT, and the shortest block size, K_{min}, for CP transforms.

5. DISCUSSION

The oracle methods perform significantly better than the semi-blind methods, because they use reference sources to find (near-)optimal source estimates with respect to the SDR. In the $J'_\gamma = 1$ case, this improvement ranges from 2.5 dB (STFT) to 4.7 dB (CP). The improvement is even larger when $J'_\gamma = 2$, and ranges from 10.6 dB (STFT) to 12.7 dB (CP). The obvious conclusion we can draw from this is that by developing more performant (semi-)blind estimation criteria (to improve upon Equations 5 and 6), we should be able to estimate more accurately the $\mathcal{J}_\gamma$ and obtain significant performance increases.

For the oracle estimators, for each of the $J'_\gamma = 1$, $J'_\gamma = 2$ and $J'_\gamma \leq 2$ cases, CP gives the highest mean SDR, followed in decreasing order by MDCT and STFT. Firstly, let us note that the improvement due to using the adaptive CP transforms, compared to the STFT, ranges from 1.6 dB ($J'_\gamma = 2$) to 2.2 dB ($J'_\gamma \leq 2$). Secondly, we note that the performance increase due to relaxing the constraints on J'_γ is even larger. Comparing the $J'_\gamma = 1$ and $J'_\gamma \leq 2$ cases, we can see that these improvements are 7.2 dB (STFT), 7.5 dB (MDCT) and 7.7 dB (CP). These results both reinforce and extend previous results, in which only the $J'_\gamma = 1$ case was considered for oracle estimation of adaptive CP transforms [7].

Although, for the oracle methods, the performance increase by using adaptive CP transforms is significant, much greater improvements can be made by going from $J'_\gamma = 1$ to $J'_\gamma \leq 2$ regardless

Crit.	J'_γ	Trans.	K	K_{min}	Mean SDR [dB]
sb	$J'_\gamma = 1$	STFT	2^{14}	-	5.8
		MDCT	2^{11}	-	5.0
		CP	-	2^{11}	5.3
	$J'_\gamma = 2$	STFT	2^{13}	-	5.4
		MDCT	2^{13}	-	4.9
		CP	-	2^{12}	4.9
ora	$J'_\gamma = 1$	STFT	2^{13}	-	8.3
		MDCT	2^{12}	-	9.2
		CP	-	2^{11}	10.0
	$J'_\gamma = 2$	STFT	2^{13}	-	16.0
		MDCT	2^{11}	-	16.6
		CP	-	2^{10}	17.6
	$J'_\gamma \leq 2$	STFT	2^{13}	-	15.5
		MDCT	2^{11}	-	16.7
		CP	-	2^{11}	17.7

Table 1. Best mean SDR [dB] for semi-blind and oracle criteria.

of the transform. This echoes the performance improvements seen by comparing semi-blind and oracle results above. Note that for the MDCT and CP transforms, setting $J'_\gamma \leq 2$ gives the best mean SDR, followed in decreasing order by setting $J'_\gamma = 2$ and finally $J'_\gamma = 1$.[2] However, the performance increase seen by comparing the $J'_\gamma = 2$ and $J'_\gamma \leq 2$ cases for the MDCT and CP transforms is relatively small (0.1 dB). This indicates that there are relatively few γ for which the oracle estimators determined a local activity pattern with $J'_\gamma = 1$, so that even though our transforms give sparse representations, there is still overlap of harmonically related sources. It would therefore be interesting to examine other adaptive transforms which attempt to make the $J'_\gamma = 1$ case more realistic, so that (semi-)blind estimation criteria based on binary masking could be alternatively used.

From Figure 1, we see that for the fixed-basis transforms, the mean SDR decreases significantly as K decreases, and that for small values K_{min}, the CP-based methods perform much better than the fixed basis methods. The reason is that the adaptive CP transforms allow for longer block sizes over those segments of the signal for which a lower distortion is attained, whereas the fixed-basis transforms are restricted to uniformly small block sizes.

Interestingly, for the semi-blind methods, using the STFT gives the best performance: 5.8 dB ($J'_\gamma = 1$) and 5.4 dB ($J'_\gamma = 2$), and the mean SDR for all semi-blind methods is lower when we set $J'_\gamma = 2$, than when $J'_\gamma = 1$. This is slightly counterintuitive, and indicates that it is important to verify the applicability of the MAP-based estimation criterion (Equation (6)) to sources with significant amounts of harmonic overlap.

6. CONCLUSION

In this paper, we formulated semi-blind and oracle methods for source separation by time-frequency masking, based on STFT, MDCT and adaptive CP transforms. Oracle results showed that by allowing more than one active source per time-frequency index, by making good estimations of the activity patterns, and by adapting the CP bases to the signal structures, adaptive CP methods have the potential to perform significantly better than fixed-basis methods. A significant problem for the future is to search for (semi-)blind methods which can approach the potential performance gain indicated by the oracle methods, particularly for the methods based on adaptive transforms.

[2] This trend does not hold strictly for STFT-based methods because estimators for non-orthogonal transforms are near-optimal, rather than optimal. A difference of several fractions of a decibel could then be expected.

7. ACKNOWLEDGMENTS

The authors thank Emmanuel Vincent for many helpful discussions during the course of this work.

8. REFERENCES

[1] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[2] M. Davies and N. Mitianoudis, "Simple mixture model for sparse overcomplete ICA," *IEE Proc. Vision, Image and Signal Processing*, vol. 151, no. 1, pp. 35–43, Feb. 2004.

[3] R. Gribonval, "Piecewise linear source separation," in *Proc. SPIE (Wavelets X)*, vol. 5207, pp. 297–310. San Diego, CA, USA, 3–7 Aug. 2003.

[4] A. Nesbit, M. D. Plumbley, and M. E. Davies, "Audio source separation with a signal-adaptive local cosine transform," *Signal Processing*, vol. 87, no. 8, pp. 1848–1858, Aug. 2007.

[5] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13, no. 4, pp. 863–882, 2001.

[6] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

[7] E. Vincent and R. Gribonval, "Blind criterion and oracle bound for instantaneous audio source separation using adaptive time-frequency representations," in *Proc. WASPAA2007*, New Paltz, NY, USA, 21–24 Oct. 2007.

[8] J. Rosca, C. Borss, and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *Proc. ICASSP '04*, Montreal, Canada, 17–21 May 2004, vol. 3, pp. 885–888.

[9] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, second edition, 1992.

[10] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, Nov. 2001.

[11] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, second edition, 1999.

[12] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Information Theory*, vol. 38, no. 2, pp. 713–718, Mar. 1992.