

BLIND CRITERION AND ORACLE BOUND FOR INSTANTANEOUS AUDIO SOURCE SEPARATION USING ADAPTIVE TIME-FREQUENCY REPRESENTATIONS

Emmanuel Vincent and Rémi Gribonval

METISS group, IRISA-INRIA

Campus de Beaulieu

35042 Rennes Cedex, France

{emmanuel.vincent, remi.gribonval}@irisa.fr

ABSTRACT

The separation of multichannel audio mixtures is often addressed by the masking approach, which consists of representing the mixture signal in the time-frequency domain and associating each time-frequency bin with a small number of active sources. Adaptive time-frequency representations can increase the disjointness of the sources compared to fixed representations. However their use has not been conclusive so far. In this paper, we propose a new criterion for the blind estimation of an adapted representation of an instantaneous mixture and explain how to compute the oracle representation leading to the best possible performance given reference source signals. Experimental results suggest that a small separation performance improvement can indeed be achieved using adaptive representations, but that complementary approaches must be investigated to obtain larger improvements.

1. INTRODUCTION

Blind audio source separation is the task of recovering the J source signals $\mathbf{s}(t) = [s_j(t)]_{1 \leq j \leq J}$ underlying an I -channel mixture audio signal $\mathbf{x}(t) = [x_i(t)]_{1 \leq i \leq I}$. This task has been tackled by various approaches depending on the type of mixing process. We focus here on underdetermined instantaneous mixtures of the form

$$\mathbf{x}(t) = \mathbf{A} \mathbf{s}(t) \quad (1)$$

where \mathbf{A} is a $I \times J$ mixing matrix modeling the source spatial positions and $I < J$. Separation can be addressed via the masking approach, which consists of representing the mixture signal in a domain where the sources have almost disjoint support, identifying the mixing matrix and associating each bin of the representation with a small number of active sources based on spatial cues [1, 2]. Fixed time-frequency representations, such as a Short-Time Fourier Transform (STFT) or a Modified Discrete Cosine Transform (MDCT) [3], are often employed. The overlap between the source representations can be minimized by choosing an appropriate window [1]. Nevertheless, it often remains rather large in some time-frequency bins, resulting in “musical noise” artifacts [1].

One approach to potentially reduce this overlap is to select the most adapted time-frequency representation among a large “library” of representations according to some estimated overlap criterion. Classical examples of such libraries include the local Cosine Packet (CP) and Wavelet Packet (WP) libraries [4]. This approach was first suggested in [5] and applied recently in [6] where two criteria were defined for the selection of adapted CP bases. However the Signal-to-Distortion Ratio (SDR) resulting from the

best criterion was 0.4 dB worse than with the MDCT on average [6]. It could not be determined from the results whether this performance limitation was due mainly to some intrinsic properties of CP bases or to the chosen criteria for the selection of the best basis and the active sources in each time-frequency bin.

In this paper, we answer this question using the standard statistics concept of oracle estimators, *i.e.* algorithms that compute the values of the separation parameters leading to the best possible performance given reference source signals. By definition, these algorithms do not address the blind source separation task, but provide instead intrinsic performance bounds for each separation approach. Some oracle estimators were recently proposed for source separation via time-invariant filtering or time-frequency masking on a fixed representation [7]. In the following, we design an oracle estimator of the best basis within a given library and combine it with existing oracle or blind estimators of the active sources in each time-frequency bin. We also propose a new blind basis selection criterion and evaluate the results both for CP and WP bases.

The rest of the paper is structured as follows. We briefly recall the principle of time-frequency masking in Section 2 and define a blind criterion and an oracle criterion for the estimation of the active sources in each time-frequency bin. We extend these criteria to the selection of the best basis in Section 3 and evaluate their performance on audio signals in Section 4. We conclude in Section 5 and point out further research directions.

2. MASKING ON A FIXED TIME-FREQUENCY BASIS

The masking approach to source separation can be conducted on any invertible representation of the data. For simplicity, we focus on orthogonal time-frequency bases, such as MDCT, CP and WP bases [3, 4], which allow the computation of exact oracle estimators [7] and a fair comparison of the results [6] since they involve the same number of coefficients. Given a fixed orthogonal basis $\mathcal{B} = \{\phi_m^{\mathcal{B}}\}_{1 \leq m \leq T}$, the coefficients of any signal $y(t)$ in this basis are obtained by $\langle y, \phi_m^{\mathcal{B}} \rangle = \sum_{t=0}^{T-1} y(t) \phi_m^{\mathcal{B}}(t)$. Also, the signal can be recovered from its coefficients as $y(t) = \sum_{m=1}^T \langle y, \phi_m^{\mathcal{B}} \rangle \phi_m^{\mathcal{B}}(t)$.

2.1. Principle

From now on, we assume that the mixing matrix \mathbf{A} is known, *e.g.* it has been precisely estimated using some clustering technique, so that the coefficients of the mixture channels $x_i(t)$ and the sources $s_j(t)$ satisfy $[\langle x_i, \phi_m^{\mathcal{B}} \rangle]_{1 \leq i \leq I} = \mathbf{A} [\langle s_j, \phi_m^{\mathcal{B}} \rangle]_{1 \leq j \leq J}$. The estimation of the source signals $\hat{s}_j(t)$ is now equivalent to that of their basis coefficients $\langle \hat{s}_j, \phi_m^{\mathcal{B}} \rangle$. Denoting by \mathcal{J}_m the set containing the

indexes of the J^{act} sources contributing most actively to the mixture at the time-frequency bin m with $J^{\text{act}} \leq I$, these coefficients can be expressed as [5, 2]

$$\begin{cases} \langle \hat{s}_j, \phi_m^{\mathcal{B}} \rangle = 0 & \text{if } j \notin \mathcal{J}_m, \\ \left[\langle \hat{s}_j, \phi_m^{\mathcal{B}} \rangle \right]_{j \in \mathcal{J}_m} = \mathbf{A}_{\mathcal{J}_m}^\dagger \left[\langle x_i, \phi_m^{\mathcal{B}} \rangle \right]_{1 \leq i \leq I} \end{cases} \quad (2)$$

where $\mathbf{A}_{\mathcal{J}_m}$ denotes the $I \times J^{\text{act}}$ matrix composed of the columns of \mathbf{A} indexed by $j \in \mathcal{J}_m$, and $\mathbf{A}_{\mathcal{J}_m}^\dagger$ is its $J^{\text{act}} \times I$ pseudo-inverse. The set \mathcal{J}_m is called an activity pattern. When $J^{\text{act}} = 1$, this expression reduces to the binary masking formula in [1, app. A].

2.2. Blind activity patterns

The difficulty of time-frequency masking lies in the blind estimation of the activity patterns $\mathcal{J} = \{\mathcal{J}_m\}_{1 \leq m \leq T}$. When $J^{\text{act}} < I$, the mixture channels $x_i(t)$ can be modeled as the sum of the estimated source signals $\hat{s}_j(t)$ scaled by the coefficients of the mixing matrix $\mathbf{A} = [a_{ij}]_{1 \leq i \leq I, 1 \leq j \leq J}$, plus a residual signal. Blind activity patterns \mathcal{J}^{bl} are then usually determined given $\mathbf{x}(t)$, \mathbf{A} and \mathcal{B} by minimizing the energy of the residual [1, 2]¹

$$e(\mathbf{x}, \mathbf{A}, \mathcal{J}, \mathcal{B}) = \sum_{i=1}^I \sum_{t=0}^{T-1} \left(x_i(t) - \sum_{j=1}^J a_{ij} \hat{s}_j(t) \right)^2. \quad (3)$$

This quantity depends implicitly on \mathcal{B} and \mathcal{J} given expression (2). Due to the orthogonality of basis \mathcal{B} , this criterion can be decomposed as $e(\mathbf{x}, \mathbf{A}, \mathcal{J}, \mathcal{B}) = \sum_{m=1}^T e(\mathbf{x}, \mathbf{A}, \mathcal{J}_m, \phi_m^{\mathcal{B}})$ with

$$e(\mathbf{x}, \mathbf{A}, \mathcal{J}_m, \phi_m^{\mathcal{B}}) = \sum_{i=1}^I \left(\langle x_i, \phi_m^{\mathcal{B}} \rangle - \sum_{j=1}^J a_{ij} \langle \hat{s}_j, \phi_m^{\mathcal{B}} \rangle \right)^2. \quad (4)$$

The blind activity patterns $\mathcal{J}_m^{\text{bl}}$ can thus be computed in each time-frequency bin m independently by selecting the minimum of (4) over all possible patterns \mathcal{J}_m . This criterion can be interpreted as a blind distortion measure on the estimated source signals. Indeed, assuming that there exists some patterns \mathcal{J}_m such that $\langle s_j, \phi_m^{\mathcal{B}} \rangle = 0$ for all $j \notin \mathcal{J}_m$ with $J^{\text{act}} < I$, then this criterion achieves its global minimum $e(\mathbf{x}, \mathbf{A}, \mathcal{J}_m, \phi_m^{\mathcal{B}}) = 0$ for these patterns only and $\langle \hat{s}_j, \phi_m^{\mathcal{B}} \rangle = \langle s_j, \phi_m^{\mathcal{B}} \rangle$ for all j .

2.3. Oracle activity patterns

When reference source signals $s_j(t)$ are available, the separation performance can be assessed by the SDR in decibels (dB), defined as $\text{SDR} = 10 \log_{10} (\sum_{j,t} s_j(t)^2 / \sum_{j,t} (\hat{s}_j(t) - s_j(t))^2)$. The use of this particular measure is justified in [7]. It is then possible to determine the best possible performance given $\mathbf{s}(t)$, $\mathbf{x}(t)$, \mathbf{A} and \mathcal{B} by computing the oracle activity patterns \mathcal{J}^{or} maximizing the SDR, or equivalently minimizing the oracle distortion measure [7]

$$d(\mathbf{s}, \mathbf{x}, \mathbf{A}, \mathcal{J}, \mathcal{B}) = \sum_{j=1}^J \sum_{t=0}^{T-1} (\hat{s}_j(t) - s_j(t))^2. \quad (5)$$

Similarly to above, this measure can be written as $d(\mathbf{s}, \mathbf{x}, \mathbf{A}, \mathcal{J}, \mathcal{B}) = \sum_{m=1}^T d(\mathbf{s}, \mathbf{x}, \mathbf{A}, \mathcal{J}_m, \phi_m^{\mathcal{B}})$ with

$$d(\mathbf{s}, \mathbf{x}, \mathbf{A}, \mathcal{J}_m, \phi_m^{\mathcal{B}}) = \sum_{j=1}^J \left(\langle \hat{s}_j, \phi_m^{\mathcal{B}} \rangle - \langle s_j, \phi_m^{\mathcal{B}} \rangle \right)^2. \quad (6)$$

¹This criterion is actually derived from a Maximum Likelihood (ML) perspective in [1, 2], under the assumption that the residual is Gaussian.

This shows that the oracle activity patterns $\mathcal{J}_m^{\text{or}}$ can be computed in each time-frequency bin m independently by selecting the minimum of (6) over all possible patterns \mathcal{J}_m .

3. SELECTION OF THE BEST BASIS

Let us now assume that we have a large library $\mathcal{L} = \{\mathcal{B}\}$ of orthonormal bases, such as a CP library, a WP library, a set of MDCT bases with various window lengths, or any union or subset of these. The source activity patterns \mathcal{J} can be computed for each basis \mathcal{B} of \mathcal{L} via any criterion (typically $\mathcal{J} = \mathcal{J}^{\text{bl}}$ or $\mathcal{J} = \mathcal{J}^{\text{or}}$). The choice of an adapted basis for each mixture signal can then potentially improve the separation performance compared to prior selection of a fixed basis for all signals.

3.1. Blind basis

In a blind context, the best basis $\mathcal{B}^{\text{bl}}(\mathcal{J})$ can be estimated by minimizing the residual energy criterion (3). Note that this basis depends on the criterion used to compute \mathcal{J} . Under the constraint of binary masking ($J^{\text{act}} = 1$), this minimization problem is equivalent to the maximization of the sum of the energies of the estimated sources scaled by the mixing coefficients. It is thus similar to the heuristic maximization of the source energies proposed in [6], except that a common basis is estimated for all sources instead of a specific basis per source.

The choice of criterion (3) is justified by the fact that the blind estimation of the activity patterns and that of the best basis ultimately share the same goal, that is the minimization of the distortion on the estimated sources using some relevant blind distortion measure. Assuming that there exists some bases \mathcal{B} and associated activity patterns \mathcal{J} such that $\langle s_j, \phi_m^{\mathcal{B}} \rangle = 0$ for all $j \notin \mathcal{J}_m$ and all m , this criterion achieves its global minimum $e(\mathbf{x}, \mathbf{A}, \mathcal{J}, \mathcal{B}) = 0$ for these bases and patterns only and $\hat{s}_j(t) = s_j(t)$ for all j . Similar two-way optimization problems can be found in other contexts. For instance, in the close context of sparse coding, the same sparsity criterion is often used for the computation of the atom weights within an overcomplete basis and for the adaptation of the basis to the analyzed signal [8].

3.2. Oracle basis

When reference source signals $s_j(t)$ are available, it is also possible to select the oracle basis $\mathcal{B}^{\text{or}}(\mathcal{J})$ resulting in the best possible SDR. As stated above, this is equivalent to minimizing the oracle distortion measure (5).

3.3. The cosine packet and wavelet packet libraries

In the general case, the selection of the best basis is computationally intensive, since all possible bases \mathcal{B} of \mathcal{L} must be tested and the activity patterns \mathcal{J}_m and the corresponding criterion values (4) or (6) must be computed for each element $\phi_m^{\mathcal{B}}$ of each basis. However, efficient global minimization algorithms exist when the library has a tree structure such that different bases share common elements [9]. CP and WP libraries satisfy this property.

CP bases window the signal into overlapping time frames of variable length, while WP bases filter it into overlapping frequency subbands of variable bandwidth [4]. Both types of bases are defined by a maximum packet depth D , in addition to a bell type for CP and a wavelet filter type for WP. Each basis element is indexed

by $m = (n, k)$ where n is a node of the library tree, denoting one of $2^{D+1} - 1$ possible frames or subbands, and $1 \leq k \leq K_n$. The depth D determines the minimum length of the time frames or the minimum bandwidth of the frequency subbands, which are equal respectively to $T \times 2^{-D+1}$ samples or $F_s \times 2^{-D+1}$, where T is the length of the signals and F_s the sampling frequency.

4. EXPERIMENTS

We evaluated the performance of the above estimators for the separation of the 20 three-source stereo ($J = 3, I = 2$) instantaneous speech and music mixtures considered in [7]. Music sources were taken from synchronized multitrack recordings², while speech sources were unrelated. The source signals were sampled at 22.05 kHz and had a duration of 2^{18} samples (11.9 s). The mixing matrix was fixed. The source and mixture signals are available online as part of the BSS Oracle toolbox³.

Separation was performed by binary masking ($J^{\text{act}} = 1$) using either MDCT bases, blind CP/WP bases $\mathcal{B}^{\text{bl}}(\mathcal{J})$ or oracle CP/WP bases $\mathcal{B}^{\text{or}}(\mathcal{J})$, and for each type of basis either blind activity patterns \mathcal{J}^{bl} given the true mixing matrix or oracle activity patterns \mathcal{J}^{or} . The Wavelab toolbox⁴ was used to compute CP and WP coefficients and search for the best basis tree given criterion values for each node. CP bases were built from a sine bell, and WP bases from a ‘symmlet-8’ filter [4].

The average SDR over all mixtures is shown in Figure 1 as a function of the MDCT window length L or the maximum packet depth D . Using oracle bases with oracle activity patterns, the best SDR was achieved with $L \simeq 1800$ samples (80 ms) for the MDCT, $D = 9$ for CP bases, corresponding to a minimum frame length of 1024 samples (46 ms), and $D = 18$ for WP bases, corresponding to a minimum subband bandwidth of 0.17 Hz. These settings were also optimal for other estimators, except for blind WP bases with blind activity patterns where $D = 13$ was best.

The SDR values corresponding to these settings of L and D are summarized in Table 1. With blind activity patterns, blind CP bases provided an average SDR improvement of 0.4 dB compared to the MDCT, while blind WP bases resulted in a SDR deterioration of 0.8 dB. The best blind CP basis performed better than the MDCT basis for all mixtures but one. This stands in contrast with the source-specific CP bases estimated from [6], which led to a SDR deterioration of 3.0 dB compared to the MDCT. The difference between this figure and that reported in Section 1 is due to the use of different data and to our slightly different definition of the SDR, which gives more weight to badly estimated sources.

Still considering blind activity patterns, the comparison of the results for blind vs. oracle bases shows a SDR difference of 0.7 dB for CP bases and 1.0 dB for WP bases. This suggests that some performance improvement could potentially be achieved in the future by using an improved blind basis estimation criterion, but that it will be at best equal to this small difference for these data.

Comparing the results for blind vs. oracle patterns, it appears that a better improvement up to 2.6 dB for the MDCT, 3.0 dB for CP bases and 3.1 dB for WP bases could be obtained in the future using a better criterion for the blind computation of the activity patterns, e.g. based on the modeling of the time-frequency structure

²These recordings are distributed under Creative Commons licenses. Their authors are Alex Q. Another Dreamer, Brian Smith, Carl Leth, Espi Twelve, Jim’s Big Ego, Mister Mouse and Mokamed.

³http://bass-db.gforge.inria.fr/bss_oracle/

⁴<http://www-stat.stanford.edu/~wavelab/>

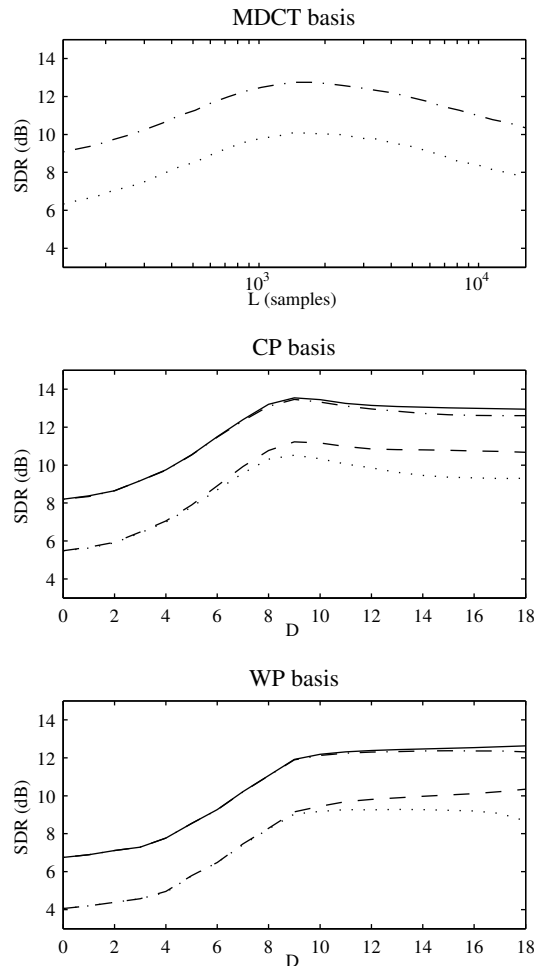


Figure 1: Average separation performance for three-source stereo mixtures by binary masking on a MDCT, CP or WP basis as a function of the window length L or the maximum packet depth D . Plain: oracle basis and activity patterns $\mathcal{B}^{\text{or}}(\mathcal{J}^{\text{or}})$. Dash-dotted: blind basis and oracle activity patterns $\mathcal{B}^{\text{bl}}(\mathcal{J}^{\text{or}})$. Dashed: oracle basis and blind activity patterns $\mathcal{B}^{\text{or}}(\mathcal{J}^{\text{bl}})$. Dotted: blind basis and activity patterns $\mathcal{B}^{\text{bl}}(\mathcal{J}^{\text{bl}})$.

Table 1: Maximum performance for each curve of Figure 1.

SDR (dB)		Blind patterns	Oracle patterns
Blind basis	MDCT	10.1	12.7
	CP	10.5	13.5
	WP	9.3	12.4
Oracle basis	CP	11.2	13.5
	WP	10.3	12.6

of audio signals. Interestingly, CP bases would then still provide a better performance than the MDCT and the blind basis estimation criterion would be near-optimal. Larger improvements are impossible for these data within the masking framework of Section 2.1.

A detailed insight of the separation performance can be obtained by computing the residual energy e_n and the oracle distort-

tion d_n for each node n of the basis tree via summation of (4) and (6) over the corresponding basis elements $m = (n, k)$ and by drawing a scatter plot of these quantities. The resulting plots with oracle activity patterns \mathcal{J}^{or} and blind activity patterns \mathcal{J}^{bl} are shown in Figure 2, either for all possible time frames of the CP library or for all time frames of the best blind CP basis $\mathcal{B}^{\text{bl}}(\mathcal{J}^{\text{bl}})$. It can be seen that both quantities are more correlated for oracle activity patterns than for blind activity patterns, which explains the fact that the residual energy criterion is near-optimal for the selection of the best basis with oracle activity patterns, but not with blind activity patterns. Moreover, with the best blind basis, the oracle distortion appears much larger for two time frames than for other time frames. This supports the observation in [6] that performance is generally better than indicated by the SDR on most time frames, hence perceptually more acceptable.

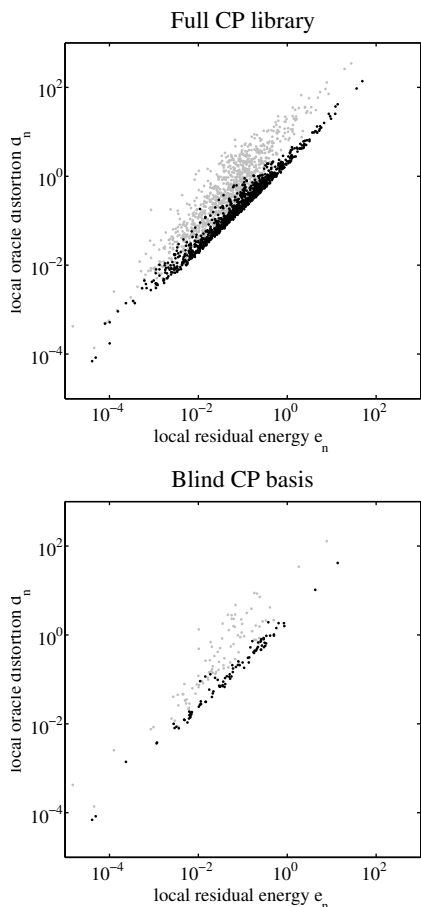


Figure 2: Comparison of the local residual energy e_n and the local oracle distortion d_n for one test signal. Top: all time frames of the CP library with maximum packet depth $D = 9$. Bottom: all time frames of the best blind CP basis $\mathcal{B}^{\text{bl}}(\mathcal{J}^{\text{bl}})$. Black: oracle activity patterns \mathcal{J}^{or} . Gray: blind activity patterns \mathcal{J}^{bl} .

5. CONCLUSION

We studied the problem of instantaneous audio source separation via time-frequency masking on orthogonal time-frequency bases.

We extended the use of the residual energy criterion for the blind estimation of the source activity patterns to that of an adapted basis and proposed an oracle basis estimator leading to the best possible performance given reference source signals. We emphasize that this oracle estimator does not address the blind source separation task, but provides an upper performance bound. Blind CP bases resulted in an average SDR improvement of 0.4 dB compared to the MDCT for the separation of three-sources stereo mixtures by binary masking, while the best possible improvement was limited to 1.1 dB. This shows that adaptive representations are only a step towards perfect separation and that alternative approaches to masking must be used in parallel. We plan in particular to integrate the separation method proposed in [10] within the framework of adaptive representation, using the same criterion for the blind estimation of the source coefficients and that of an adapted basis. We will also consider the extension of source separation methods based on the STFT [1, 2] to overcomplete adaptive representations, which would allow the separation of convolutive mixtures. Finally, we will study the effect of imprecise estimation of the mixing system.

6. ACKNOWLEDGMENT

The authors would like to thank Mark Plumbley for careful reading of early versions of this paper.

7. REFERENCES

- [1] Ö. Yilmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
- [2] J. P. Rosca, C. Borss, and R. V. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. III–877–880.
- [3] J. P. Princen and A. B. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 5, pp. 1153–1161, 1986.
- [4] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press, 1998.
- [5] R. Gribonval, "Piecewise linear source separation," in *Proc. SPIE*, vol. 5207 Wavelets: Applications in Signal and Image Processing X, 2003, pp. 297–310.
- [6] 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, 2007.
- [7] 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.
- [8] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?" *Vision Research*, vol. 37, pp. 3311–3325, 1997.
- [9] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 713–718, 1992.
- [10] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.