UNDERDETERMINED SPARSE BLIND SOURCE SEPARATION WITH DELAYS

Rayan Saab, Özgür Yılmaz, Martin J. McKeown, Rafeef Abugharbieh

ABSTRACT

In this paper, we address the problem of under-determined blind source separation (BSS), mainly for speech signals, in an anechoic environment. Our approach is based on exploiting the sparsity of Gabor expansions of speech signals. For parameter estimation, we adopt the clustering approach of DUET [19]. However, unlike in the case of DUET where only two mixtures are used, we use all available mixtures to get more precise estimates. For source extraction, we propose two methods, both of which are based on constrained optimization. Our first method uses a constrained ℓ^q $(0 < q \leq 1)$ approach, and our second method uses a constrained "modified" ℓ^1 minimization approach. In both cases, our algorithms use all available mixtures, and are suited to the anechoic mixing scenario. Experiments indicate that the performances of the proposed algorithms are superior compared to DUET in many different settings.

1. INTRODUCTION

BSS algorithms can be categorized according to the assumptions they make about the mixing model. Thus, one can classify them as either instantaneous, anechoic or echoic and as either under-determined, even-determined, or over-determined. Moreover, to make the problem tractable, one usually needs to make certain assumptions about the nature of the sources. Several methods exist that attempt to solve the BSS problem under various assumptions and conditions, see, e.g., [15].

One such method, independent component analysis (ICA) (e.g., [9], [2]) is based on the assumption that the sources are statistically independent. In this case, ICA can extract n sources from n recorded *instantaneous* mixtures of these sources. In fact, [11] and [10] show that ICA can be expanded into the case of instantaneous under-determined mixtures (where there are more sources than mixtures).

An alternative approach to the BSS problem for under-determined *instantaneous* mixtures is to assume that the sources have a sparse expansion with respect to some basis (or redundant dictionary). In this case, one can formulate the source extraction problem as a constrained ℓ^1 minimization problem, which typically yields a convex program [4], [12], [13].

There are also algorithms that demix under-determined *anechoic* mixtures. One such algorithm, presented in [1] employs a complex independent component analysis technique to solve the BSS problem for electroencephalographic (EEG) data. This approach involves solving a permutation problem as well as identifying whether sources extracted from various bands correspond to one another or not.

Another algorithm to demix under-determined *anechoic* mixtures, called DUET, was proposed by Yilmaz and Rickard [19]. This algorithm uses sparsity of speech signals in the short time Fourier transform (STFT) domain to construct binary time-frequency masks, which are then used to extract several sources from only two mixtures. The advantage of DUET is that it is very fast; the disadvantage is that it, by construction, uses only two mixtures even if there are more mixtures available. Another under-determined *anechoic* demixing algorithm was proposed by Bofill in [3]. This algorithm estimates the attenuation coefficients by using a scatter plot technique and the delays by maximizing a kernel function. To extract the sources, [3] solves a complex constrained ℓ^1 minimization problem via second order cone programming. This algorithm, like DUET, uses only two of the available mixtures.

In this paper we propose two algorithms to solve the problem of *anechoic* mixing when we have *more than two mixtures* available. We adopt the two stage approach as formalized by Theis et al. [18]. The first (parameter estimation) stage of both proposed algorithms uses clustering of feature vectors constructed from the Gabor (or STFT) coefficients of the mixtures to estimate the mixing parameters (see Section 3). In the second (source extraction) stage, one algorithm uses an ℓ^q minimization based approach, with $q < 1$ while the other algorithm uses a modified ℓ^1 minimization approach, to estimate the sources in the STFT domain (see Sections 4.1 and 4.2, respectively).

Simulation results highlight the benefits of the proposed algorithms.

2. MIXING MODEL AND STATEMENT OF THE PROBLEM

Suppose we have n time domain sources $s_1(t), \ldots, s_n(t)$ and m mixtures $x_1(t), \ldots, x_m(t)$ such that

$$x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t - \delta_{ij}),\ i = 1, 2, \ldots, m \tag{1}$$

where $m < n$ and $a_{kj} \in \mathbb{R}^+$ and $\delta_{kj} \in \mathbb{R}$ are attenuation coefficients and time delays associated with the path from the j^{th} source to the i^{th} receiver, respectively. Equation (1) defines an *anechoic* mixing model. Without loss of generality, we set $\delta_{1j} = 0$ for $j = 1, \ldots, n$ and we scale the source functions s_j such that

$$\sum_{i=1}^{m} a_{ij}^2 = 1 \tag{2}$$

for $j = 1, \ldots, n$.

Using an appropriate window and taking the STFT of $x_1, \ldots, x_m$, the mixing model (1) can now be expressed as

$$\hat{\mathbf{x}}(\tau, \omega) = A(\omega)\hat{\mathbf{s}}(\tau, \omega), \tag{3}$$

where

$$\hat{\mathbf{x}} = [\hat{x}_1 \ldots \hat{x}_m]^T,\ \hat{\mathbf{s}} = [\hat{s}_1 \ldots \hat{s}_n]^T, \tag{4}$$

$\hat{x}_i$ and $\hat{s}_j$ denote the STFT of x_i and s_j, respectively, and

$$A(\omega) = \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ a_{21}e^{-i\omega\delta_{21}} & \ldots & a_{2n}e^{-i\omega\delta_{2n}} \\ \vdots & \vdots & \vdots \\ a_{m1}e^{-i\omega\delta_{m1}} & \ldots & a_{mn}e^{-i\omega\delta_{mn}} \end{bmatrix}. \tag{5}$$

Rayan Saab and Rafeef Abugharbieh are with the Department of Electrical and Computer Engineering, The University of British Columbia.

Özgür Yılmaz is with the Department of Mathematics, The University of British Columbia.

Martin J. McKeown is with Department of Medicine (Neurology), Pacific Parkinson's Research Centre, The University of British Columbia

Note that (2) ensures that the column vectors of A have unit norm.

Henceforth, we will replace the continuous STFT with the equivalent discrete counterpart, i.e., the Gabor coefficients – the samples of the STFT of s on a sufficiently dense lattice in the TF plane given by

$$\hat{s}_j[k,l] = \hat{s}_j(k\tau_0, l\omega_0) \tag{6}$$

where τ_0 and ω_0 are the time-frequency lattice parameters. Similar notation will be used for the mixing matrix A and the Gabor coefficients of the mixtures x_i.

Our approach throughout the rest of the paper is based on the following assumption about the Gabor expansions of speech signals.

Sparsity assumption: *Gabor expansions of speech signals are sparse in that few coefficients will capture most of the signal power (see e.g., [19]).*

3. MIXING MODEL RECOVERY

In this section, we will describe our method for the recovery of the mixing model parameters, i.e., the delay and attenuation coefficients. Successful extraction of these parameters relies on the sparsity of speech in the STFT domain, which we briefly discuss next.

3.1. Speech Sparsity in the STFT domain

In [19], it was shown that Gabor expansions of speech signals are sparse. For further illustration of the sparsity exhibited by speech in the STFT domain, in Figure 1 we show the average cumulative powers of the *sorted Gabor (STFT) coefficients*, the average cumulative power of the time domain sources and of their Fourier (DFT) coefficients. We observe that the STFT with a window of 64ms duration demonstrates superior performance in terms of sparsity, capturing 98% of the total signal power with only approximately 9% of the coefficients.

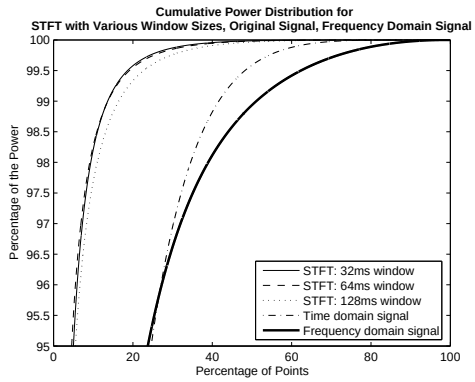


Fig. 1. Average cumulative power of 50 3-second speech signals in time domain, frequency (Fourier) domain, and TF domain for window sizes of 32ms, 64ms and 128ms. The STFT with 32ms and 64ms window length yield significantly sparser representations of the data (more power in fewer coefficients).

3.2. Parameter Estimation via Clustering

Consider the feature vectors at each TF point [k,l] given by

$$\mathbf{F}[k,l] := \begin{bmatrix} \left|\frac{\hat{x}_1[k,l]}{\|\hat{\mathbf{x}}[k,l]\|}\right| & \cdots & \left|\frac{\hat{x}_m[k,l]}{\|\hat{\mathbf{x}}[k,l]\|}\right| \\ \hat{\Delta}_{21}[k,l] & \cdots & \hat{\Delta}_{m1}[k,l] \end{bmatrix}. \tag{7}$$

Here $\|\cdot\|$ denotes the Euclidean norm and as in [19],

$$\hat{\Delta}_{j1}[k,l] := -\frac{1}{l\omega_0}\angle\frac{\hat{x}_j[k,l]}{\hat{x}_1[k,l]}. \tag{8}$$

Remarks:

1. The first m-dimensions of the resulting feature vectors correspond to points on the unit sphere of $\mathbb{R}^m$.
2. If only one source s_J is nonzero at a TF point, the feature vector at that TF point will reduce to

$$\begin{aligned}\mathbf{F}[k,l] &= [\ a_{1J} \quad \cdots \quad a_{mJ} \quad \cdots \quad \delta_{2J} \quad \cdots \quad \delta_{mJ}\ ] \\ &:= \mathbf{F}_J\end{aligned} \tag{9}$$

   Thus, the feature vectors calculated at any TF point $[k,l]$ at which source J is the only active source will be identical, and equal to $\mathbf{F}_J$.

The sparsity assumption for the sources in the TF domain means that we expect an abundance of points where only one source is active. Thus, we use a clustering approach, such as k-means, in the feature space to estimate the delay and attenuation parameters of the mixing model. In summary, the proposed **Parameter Estimation Algorithm** is as follows.

1. Compute the mixture vector $\hat{\mathbf{x}}[k,l]$ at every TF point $[k,l]$.
2. At every TF point $[k,l]$, compute the corresponding feature vector $F[k,l]$, as in (7),
3. Perform some clustering algorithm (e.g., k-means) to find the n cluster centers in the feature space. The cluster centers will yield preliminary estimates $\bar{a}_{ij}$ and $\bar{\delta}_{ij}$ of the mixing parameters a_{ij} and δ_{ij}, respectively, via (9).
4. Normalize the attenuation coefficients to obtain the *final attenuation parameter estimates* $\tilde{a}_{ij}$, i.e.,

$$\tilde{a}_{ij} := \bar{a}_{ij}/(\sum_{i=1}^{m}\bar{a}_{ij}^2)^{1/2}.$$

   The *final delay parameter estimates* are given by $\tilde{\delta}_{ij} := \bar{\delta}_{ij}$.

4. SOURCE SEPARATION

The cluster centers, obtained as described above, can be used to construct the estimated mixing matrix

$$\tilde{A}[l] = \begin{bmatrix} \tilde{a}_{11}e^{-il\omega_0\tilde{\delta}_{11}} & \cdots & \tilde{a}_{1n}e^{-il\omega_0\tilde{\delta}_{1n}} \\ \tilde{a}_{21}e^{-il\omega_0\tilde{\delta}_{21}} & \cdots & \tilde{a}_{2n}e^{-il\omega_0\tilde{\delta}_{2n}} \\ \vdots & \vdots & \vdots \\ \tilde{a}_{m1}e^{-il\omega_0\tilde{\delta}_{m1}} & \cdots & \tilde{a}_{mn}e^{-il\omega_0\tilde{\delta}_{mn}} \end{bmatrix}. \tag{10}$$

Note that each column vector of $\tilde{A}[l]$ is a unit vector in $\mathbb{C}^m$.

Our goal in this section is to compute "good" estimates $s_1^e, s_2^e, ..., s_n^e$ of the original sources $s_1, s_2, ..., s_n$, which must satisfy

$$\tilde{A}[l]\hat{\mathbf{s}}^{\mathbf{e}}[k,l] = \hat{\mathbf{x}}[k,l], \tag{11}$$

where $\hat{\mathbf{s}}^e = [\hat{s}_1^e, ...\hat{s}_n^e]^T$ is the vector of source estimates in the TF domain. At each TF point $[k,l]$, (11) provides m equations (corresponding to the m available mixtures) with $n > m$ unknowns $(\hat{s}_1^e, ...\hat{s}_n^e)$. Assuming that this system of equations is consistent, it has infinitely many solutions. To choose a reasonable estimate among these infinitely many solutions, we shall exploit the sparsity of the sources in the TF domain.

4.1. Algorithm 1: Constrained ℓ^q minimization

We wish to find, at each time frequency point, the "sparsest" $\hat{\mathbf{s}}^{\mathbf{e}}$ that solves (11). This problem can be formally stated as

$$\min_{\hat{\mathbf{s}}^e} \|\hat{\mathbf{s}}^e\|_{\text{sparse}} \quad \text{subject to} \quad \tilde{A}\hat{\mathbf{s}}^{\mathbf{e}} = \hat{\mathbf{x}}, \tag{12}$$

where $\|\mathbf{x}\|_{\text{sparse}}$ denotes some measure of sparsity of the vector $\mathbf{x}$.

Given a vector $\mathbf{x} = (x_1, \ldots, x_n) \in C^n$, one measure of its sparsity is given by the number of the non-zero components of $\mathbf{x}$, commonly denoted by $\|\mathbf{x}\|_0$. Replacing $\|\mathbf{x}\|_{\text{sparse}}$ in (12) with $\|\mathbf{x}\|_0$, we get the so-called P_0 problem, e.g., [7]. Solving P_0 is, in general, combinatorial and the solution is very sensitive to noise. More importantly, the sparsity assumption we have for the Gabor coefficients of speech signals essentially suggests that most of the coefficients are very small, however not identically zero. In this case, P_0 fails tremendously, since the sparsity is measured as the number of nonzero components which does not take into account the size of the components at all. Alternatively, one can consider

$$\|\mathbf{x}\|_q := (\sum_i |x_i|^q)^{1/q},$$

where $0 < q \leq 1$, as a measure of sparsity. Here, the smaller we choose q, the more importance we attach to the sparsity of $\mathbf{x}$, e.g., [6]. Moreover, this sparsity measure takes into account the size of the components.

Motivated by this, we propose to compute the vector of source estimates $\hat{\mathbf{s}}^e$ by solving at each TF point $[k, l]$ the P_q problem which is defined by replacing $\|\mathbf{x}\|_{\text{sparse}}$ in (12) with $\|\mathbf{x}\|_q$.

4.1.1. Solving P_q

The optimization problem P_q is not convex, thus computationally challenging. Under certain conditions on the sparsity of $\mathbf{x}$, it can be shown that a near minimizer can be obtained by solving the convex P_1 problem, see [5, 7, 14]. However, we do not want to impose any a priori conditions on the sparsity of the Gabor coefficients of the source vectors. Without such conditions, only local optimization algorithms are known in the literature [14]. Below, we show that the P_q problem with $0 < q < 1$ can be solved in combinatorial time.

Theorem 1 *Let $A = [\mathbf{a}_1|\mathbf{a}_2|\ldots|\mathbf{a}_n]$ be an $n \times m$ matrix with $n > m$, $A_{ij} \in \mathbb{C}$, and the column vectors $\mathbf{a}_i$ have unit norm. Suppose A is full rank. For $0 < q < 1$, the P_q problem*

$$\min_{\mathbf{s}} \|\mathbf{s}\|_q \quad \textit{subject to} \quad A\mathbf{s} = \mathbf{x},$$

where $\mathbf{x} \in \mathbb{C}^n$, has a solution $\mathbf{s}^* = (s_1^*, \ldots s_n^*)$ which has $k \leq m$ non-zero components. Moreover, if the non-zero components of $\mathbf{s}^*$ are $s_{i(j)}^*$, $j = 1, \ldots, k$, then the corresponding column vectors $\{\mathbf{a}_{i(j)} : j = 1, \ldots, k\}$ of A are linearly independent.

The proof of this theorem is elementary and can be found in [17].

Theorem 1 makes the P_q problem computationally tractable, shows that there are only finitely many solutions of P_q, and suggests a combinatorial algorithm to solve P_q. More precisely, let $\mathcal{A}$ be the set of all $m \times m$ invertible sub-matrices of A (we know that $\mathcal{A}$ is non-empty as A is full rank). Then the solution of P_q will be given by the solution of

$$\min \|B^{-1}\mathbf{x}_B\|_q \quad \text{where} \quad B \in \mathcal{A}. \tag{13}$$

Here for $B = [\mathbf{a}_{i(1)}|\cdots|\mathbf{a}_{i(m)}]$, $\mathbf{x}_B := [x_{i(1)} \cdots x_{i(m)}]$. Noting that $\#\mathcal{A} \leq \binom{n}{m}$, (13) is a combinatorial problem.

Based on the discussion in the previous sections, we now present our **separation algorithm**. At each TF point $[k, l]$:

1. Construct the estimated mixing matrix $\tilde{A}[l]$ as in (10),
2. Find the estimated source vector $\hat{\mathbf{s}}^e[k, l]$ by solving:

$$P_q\text{: } \min_{\hat{\mathbf{s}}^e[k,l]} \|\hat{\mathbf{s}}^e[k, l]\|_q \quad \text{subject to} \quad \tilde{A}[l]\hat{\mathbf{s}}^{\mathbf{e}}[k, l] = \hat{\mathbf{x}}[k, l]. \tag{14}$$

   for some $0 < q < 1$. (In the experiments section, we shall observe that a choice of $0.1 \leq q \leq 0.5$ is desirable.)
3. After repeating steps 1 and 2 for all TF points, reconstruct $\mathbf{s}^{\mathbf{e}}(t)$, the time domain estimate of the sources, from the estimated Gabor coefficients.

4.2. Algorithm 2: Constrained "Modified" ℓ^1 Minimization

Motivated by the sparsity of the sources, we consider a "modified" version of (14) with $q = 1$. We are thus trying to find the solution of

$$P_{1,mod}\text{: } \min_{\hat{\mathbf{s}}^e[k,l]} \|\hat{\mathbf{s}}^e[k, l]\|_1 \text{ subject to } \begin{cases} \tilde{A}[l]\hat{\mathbf{s}}^e[k, l] = \hat{\mathbf{x}}[k, l] \\ \hat{\mathbf{s}}^e[k, l]_{\text{cl}} \neq 0 \end{cases} \tag{15}$$

where $\hat{\mathbf{s}}^e[k, l]_{\text{cl}}$ is the component of the source estimate that corresponds to the column of $\tilde{A}[l]$ that has the largest inner product with the mixture feature-vector $\mathbf{x}[k, l]$. In other words, we are imposing an additional constraint by looking for a solution that incorporates the source that is "closest" to the observation. Thus, we are assuming that the observation vector is comprised mainly of a contribution from one source; the other sources could also be active, but to a lesser degree. This leads to the following algorithm. At each TF point $[k, l]$:

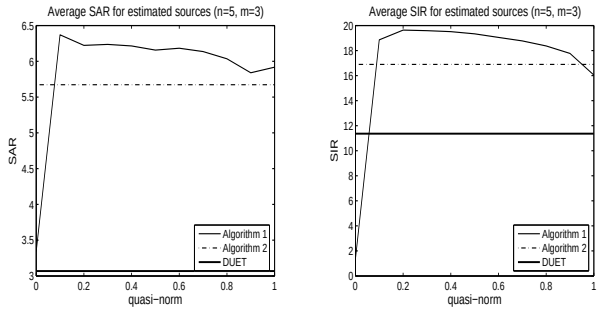
1. Construct the estimated mixing matrix $\tilde{A}[l]$ as in (10),
2. Find the estimated source vector $\hat{\mathbf{s}}^e[k, l]$ by solving (15).
3. After repeating steps 1 and 2 for all TF points, reconstruct $\mathbf{s}^{\mathbf{e}}(t)$, the time domain estimate of the sources, from the estimated Gabor coefficients.

4.3. Variation to the Algorithms

We introduce a variation to the algorithm whereby ρ is a user set parameter utilized to improve the algorithms' estimates of the sources and also to increase sparsity. Thus, at each TF point we remove the estimates of the smallest sources whose combined power contribution is less than $100(1 - \rho)\%$ of the total power at that TF point. Consequently, for $\rho = 0$ only the highest estimate is kept, and for $\rho = 1$ all estimates are kept.

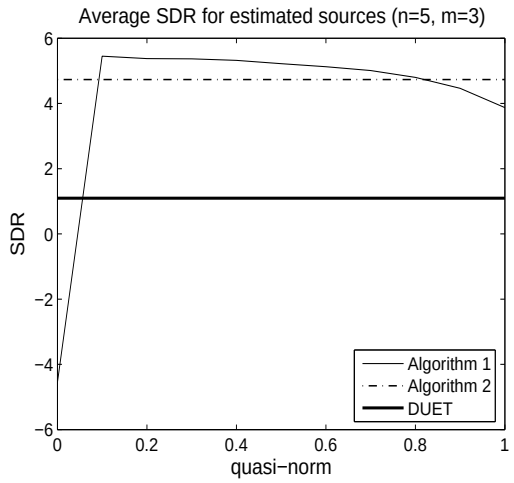
5. EXPERIMENTS AND RESULTS

To assess the performance of the algorithm, we used an anechoic room mixing model [16]. The simulated scenario involved 3 microphones and 5 sources placed in the room. We performed an experiment extracting the sources from the mixtures using both proposed algorithms, and repeated the experiment 60 times by varying the speech sources and their locations in the room. We also compare the results to those of DUET. We report the results in Figure 2 in terms of Signal-to-Artifact (SAR), Signal-to-Interference (SIR) and Signal-to- Distortion (SDR) Ratios defined as in [8]. We can see that for algorithm 1, the best SDR performance occurs at $q = 0.1$, while the best SIR performance occurs at $q = 0.2$. In other simulations that we have performed involving different numbers of sources and sensors, and which we do not report here for lack of space, we note that the best performance is usually obtained for $0.1 < q < 0.5$.



(a) SAR

(b) SIR



(c) SDR

Fig. 2. SAR, SIR and SDR obtained from demixing 5 mixtures of 10 sources using the two proposed algorithms with various choices for the quasi-norm q (for Algorithm 1) and with $\rho = 0.4$ for both algorithms.

6. CONCLUSION

We have presented two BSS algorithms for under-determined, anechoic mixtures that adopt a two step approach combining the strengths of constrained minimization and DUET. In the first stage of both algorithms we compose appropriate feature vectors and use them to extract the parameters of the mixing model via clustering. Following this *blind mixing model recovery* stage we perform a *blind source extraction stage*, for which we proposed two algorithms. The first algorithm is based on ℓ^q minimization, and the second algorithm is based on a modified version of ℓ^1 minimization. Both algorithms perform the demixing at every significant TF point separately, because our mixing matrix is frequency dependent. Our experimental results show that both algorithms outperform DUET. Moreover, it is noteworthy that algorithm 1 outperforms algorithm 2 for $q < 0.8$. The importance of these algorithms is that both allow the use of all available mixtures to perform separation in an under-determined anechoic mixing environment.

7. REFERENCES

[1] J. Anemuller, T. Sejnowski, and S. Makeig, "Complex independent component analysis of frequency domain electroencephalographic data," in *Neural Networks*, 2003, pp. 16:1311–1323.

[2] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[3] P. Bofill, "Underdetermined blind separation of delayed sound sources in the frequency domain," *Neurocomputing*, vol. 55, pp. 627–641, 2003.

[4] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," in *International Workshop on Independent Component Analysis and Blind Signal Separation (ICA)*, Helsinki, Finland, June 19–22 2000, pp. 87–92.

[5] D. Donoho, "Compressed Sensing," *Preprint*. [Online]. Available: http://www-stat.stanford.edu/~donoho/Reports/2004/CompressedSensing091604.pdf

[6] ——, "Sparse components of images and optimal atomic decompositions," *Constructive Approximation*, vol. 17, pp. 352–382, 2001.

[7] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization," in *Proc. Natl. Acad. Sci. USA 100 (2003), 2197-2202*.

[8] R. Gribonval, L. Benaroya, E. Vincent, and C. Fevotte, "Proposal for performance measurement in source separation," in *Proceedings of 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, 2003, pp. 763–768.

[9] C. Jutten, J. Herault, P. Comon, and E.Sorouchiary, "Blind separation of sources, parts i,ii and iii," *Signal Processing*, vol. 24, pp. 1–29, 1991.

[10] T.-W. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Proc. Letters*, vol. 6, no. 4, pp. 87–90, April 1999.

[11] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," in *Neural Computation*, 2000, pp. 12:337–365.

[12] Y. Li, A. Cichocki, and S. Amari, "Sparse component analysis for blind source separation with less sensors than sources," in *Proceedings of 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Riken. Kyoto, Japan: ICA, Apr. 2003, pp. 89–94.

[13] Y. Li, A. Cichocki, S. Amari, S. Shishkin, J. Cao, and F. Gu, "Sparse representation and its applications in blind source separation," in *Seventeenth Annual Conference on Neural Information Processing Systems (NIPS-2003)*, Vancouver, Dec. 2003.

[14] D. Malioutov, "A sparse signal reconstruction perspective for source localization with sensor arrays," Master's thesis, MIT, 2003.

[15] P. O'Grady, B. Pearlmutter, and S. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, 2005.

[16] S. Rickard, "Personal communication," 2005.

[17] R. Saab, O. Yılmaz, M. McKeown, and R. Abugharbieh, "A sparsity based approach for blind demixing of underdetermined multi-channel anechoic mixtures," *Preprint*.

[18] F. Theis and E. Lang, "Formalization of the two-step approach to overcomplete bss," in *Proc. 4th Intern. Conf. on Signal and Image Processing (SIP'02) (Hawaii)*, N. Younan, Ed., 2002.

[19] O. Yilmaz and S. Rickard, "Blind source separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.