GENERALIZED SPARSE SIGNAL MIXING MODEL AND APPLICATION TO NOISY BLIND SOURCE SEPARATION

Justinian Rosca, Christian Borss, Radu Balan

Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540
{justinian.rosca,christian.borss,radu.balan}@scr.siemens.com

ABSTRACT

Sparse constraints on signal decompositions are justified by typical sensor data used in a variety of signal processing fields such as acoustics, medical imaging, or wireless, but moreover can lead to more effective algorithms. The specific sparseness assumption used in this work is that the maximum number of statistically independent sources active at any time and frequency point in a mixture of signals is small. This is shown to result from an assumption of sparseness of the sources themselves, and allows us to solve the maximum likelihood formulation of the non-instantaneous acoustic mixing source estimation problem. We consider an additive noise mixing model with an arbitrary number of sensors and possibly more sources than sensors, when sources satisfy the sparseness assumption above. The solution obtained is applicable to an arbitrary number of microphones and sources, but works best when the number of sources simultaneously active at any time frequency point is a small fraction of the total number of sources.

1. INTRODUCTION

The idea of a sparse signal representation is to transform signal data into a domain where data can be parsimoniously described (for instance by a superposition of a small number of basis) or, more generally, has a small ℓ^p-norm $(0 \leq p \leq 1)$ [1, 2, 3]. Typical signal transformations are the Fourier, the wavelet transform, or independent component analysis (ICA) transformations adapted to the data [4]. Taking sparseness as a prior assumption about signal models is often justified by the nature of signals (e.g. natural images [5], sounds [6]). More importantly, the assumption can lead to effective algorithms for signal separation. This has been the case in applications ranging from audio source separation to medical and image signal processing [7]. The subject of this paper is twofold. First we discuss the form of our sparsity assumption, then we present its application to blind source separation of noisy real-world audio signals.

Our sparsity assumption is given by a constraint on the maximum number of statistically independent sources present in a mixture of signals at any time and frequency point. This is a generalization of the assumption – that time-frequency representations of any two sources do not overlap – used in [8], which introduced a BSS technique for the separation of an arbitrary number of sources from just *two* mixtures. The key observation in the technique was that each time-frequency (TF) point depended on at most one source and its associated mixing parameters. This deterministic hypothesis was called *W-disjoint orthogonality*. In anechoic non-noisy environments, it is possible to extract the mixing parameters from the ratio of the TF representations of the mixtures. Using the mixing parameters, one can partition the TF representation of the mixtures to produce the original sources.

The deterministic signal model was extended to a stochastic signal model in [9], where each time-frequency coefficient was modeled as a product between a continuous random variable and a 0/1 discrete Bernoulli random variable (indicating the "presence" of the source). This way signals can be modeled as independent random variables, and one can derive the maximum likelihood (ML) estimator of the mixing parameters. The approach has good results on real data even for more sources than sensors and has been further analyzed in the literature. However, the sparse nature of the signal estimates implies that their time-domain reconstruction by time-frequency masking will contain artefacts. The problem is alleviated in [10] by combination of masking and ICA.

In this paper we deal with a multi-channel $(D > 2)$ extension in the presence of noise by exploring a generalization of the sparsity assumptions from [8, 9, 11]. We further extend the ML estimators derived before. The ML approach considers both mixing parameters and sources, unlike in [12] where the optimization was over mixing parameters only.

The rest of the paper is organized as follows. Section 2 presents a statistical motivation of the sparseness assumption, its generalization, and the mixing model. Section 3 shows the derivation of the ML estimator of the mixing parameters and the source signals. Section 4 focuses on the capability of the algorithm under the assumption that either one or only two sources are active at any time frequency point. Experiments use eight sensors and four voice mixtures in the presence of noise and show enhanced intelligibility of speech under the more general sparsity assumption.

2. MIXING MODEL ASSUMPTIONS

2.1. Sparseness and the Generalized W-Disjoint Orthogonality Hypothesis

In [12] we called two signals s_1 and s_2 *W-disjoint orthogonal*, for a given windowing function $W(t)$, if the supports of the windowed Fourier transforms of s_1 and s_2 are disjoint, that is:

$$S_1(k,\omega)S_2(k,\omega) = 0 \ , \ \forall k, \omega \tag{1}$$

This deterministic assumption implies that the signals are in general statistically dependent, which is not the case. Yet, relation

IN PROCEEDINGS OF ICASSP 2004, MONTREAL CANADA, MAY 2004

(1) is satisfied in an approximate sense (e.g. in particular by real speech signals [13]). Furthermore, [11] shows that relation (1) can be seen as the limit of a stochastic model introduced in [9].

Here we generalize this model and show that it follows from a sparsness prior. We call L signals $s_1, s_2, \ldots, s_L$ *generalized W-disjoint orthogonal* (or *N-term W-disjoint orthogonal*) if for every time-frequency point (t, ω), there are $L-N$ indices $\{j_{N+1}, \ldots, j_N\}$ in $\{1, 2, \ldots, L\}$ so that

$$S_{j_k}(t, \omega) = 0 \ , \ \forall N+1 \leq k \leq L. \tag{2}$$

We briefly review the model and signal class from [9]. It states that the time-frequency coefficient $S(k, \omega)$ of a (speech) signal $s(t)$ factors as a product of a continuous random variable, say $G(k, \omega)$, and a 0/1 Bernoulli $V(k, \omega)$:

$$S(k, \omega) = V(k, \omega) G(k, \omega) \tag{3}$$

This formula models sparse signals. Denoting by q the probability of V to be 1, and by $p(\cdot)$ the p.d.f. of G, the p.d.f. of S turns into

$$p_S(S) = qp(S) + (1-q)\delta(S) \tag{4}$$

with δ, the Dirac distribution. For L independent signals $S_1, \ldots, S_L$, the joint p.d.f. is obtained by conditioning with respect to the Bernoulli random variables. To simplify the notation, we assume all $G(k, \omega)$ have the same distribution $p(\cdot)$, and all $V(k, \omega)$ have the same q. We obtain:

$$\begin{aligned} p(S_1, \ldots, S_L) &= (qp(S) + (1-q)\delta(S))^L \qquad (5) \\ = \sum_{k=0}^{L} q^k (1-q)^{L-k} & \sum_{1 \leq a_1 \leq a_2 \leq \cdots \leq a_k \leq L} \prod_{j=1}^{k} p(S_{a_j}) \prod_{j=k+1}^{L} \delta(S_{a_j}) \end{aligned}$$

where $\{a_1, a_2, \ldots, a_L\} = \{1, 2, \ldots, L\}$.

Next assume $q \ll 1$ and approximate the expansion by only the first N terms. Renormalizing the remaining terms, we obtain

$$\begin{aligned} p_{\text{GWDO}} = & \frac{(1-q)^L}{Z} \prod_{l=1}^{L} \delta(S_l) + q \frac{(1-q)^{L-1}}{Z} \sum_{l=1}^{L} p(S_l) \prod_{j \neq l} \delta(S_j) \\ & + \cdots q^N \frac{(1-q)^{L-N}}{Z} \sum_{1 \leq a_1 \leq a_2 \leq \cdots \leq a_N \leq L} \prod_{j=1}^{N} p(S_{a_j}) \prod_{j=N+1}^{L} \delta(S_{a_j}) \end{aligned} \tag{6}$$

with $\{a_1, a_2, \ldots, a_N, a_{N+1}, \ldots, a_L\} = \{1, 2, \ldots, N\}$, and $Z = (1-q)^{L-N} \frac{(1-q)^{N+1} - q^{N+1}}{1-2q}$. Each term of this expression has an interesting interpretation. More specifically, the rank k term, $0 \leq k \leq N$, is associated to a case when exactly k sources are active, and the rest are zero. The the joint p.d.f. in (6) corresponds to the case when *at most N sources are active simultaneously*, which constitutes the *generalized W-disjoint hypothesis*.

This is the stochastic counterpart of the deterministic constraint implied by (2). Equation (6) shows that the constraint on the signals is a reasonable assumption in the stochastic limit, hence the name p_{GWDO}. In this paper we do assume the joint p.d.f. of the source signals in the short-time Fourier domain is given by (6), with the interpretation that this is not an inconsistent assumption but rather the limit of a stochastic model derived from assumptions of sparsity of the sources.

2.2. Mixing Model

Next we introduce a specific additive noise mixing model for non-instantaneous audio signals, where sensor noises are assumed independently distributed and have Gaussian distributions with zero mean and σ^2 variance.

Consider the measurements of L source signals by a equispaced linear array of D sensors under a far-field assumption where only the direct path is present. In this case, without loss of generality, we can absorb the attenuation and delay parameters of the first mixture $x_1(t)$, into the definition of the sources. Furthermore, for the purposes of this paper we neglect the relative attenuation between sensors.

$$\begin{aligned} x_1(t) &= \sum_{l=1}^{L} s_l(t) + n_1(t) \\ x_k(t) &= \sum_{l=1}^{L} s_l(t - \tau_{k,l}) + n_k(t), \ \ 2 \leq k \leq D \end{aligned} \tag{7}$$

where $n_1, \ldots, n_D$ are the sensor noises, and $\tau_{d,l}$ is the delay of source l to sensor d. For a far-field equispaced sensor array, the delays $\tau_{d,l}$ are linearly distributed across the sensors (i.e. with respect to index d). We can define the average delay τ_l, so that

$$\tau_{d,l} = (d-1)\tau_l, \quad 1 \leq d \leq D, 1 \leq l \leq L \tag{8}$$

Clearly other mixing models can be considered at the expense of increasing the model complexity. We use Δ to denote the maximal possible delay between adjacent sensors, and thus $|\tau_l| \leq \Delta, \forall l$.

We denote by $X_d(k, \omega)$, $S_l(k, \omega)$, $N_d(k, \omega)$ the short-time Fourier transform of signals $x_d(t)$, $s_l(t)$, and $n_d(t)$, respectively, with respect to a window $W(t)$, where k is the frame index, and ω the frequency index. Then the mixing model (7) turns into

$$X_d(k, \omega) = \sum_{l=1}^{L} e^{-i\omega(d-1)\tau_l} S_l(k, \omega) + N_d(k, \omega) \tag{9}$$

When no danger of confusion arises, we drop the arguments k, ω in X_d, S_l and N_d.

Our problem is: *given measurements $(x_1(t), \ldots, x_D(t))_{1 \leq t \leq T}$ of the system (7), estimate the mixing parameters $(\tau_l)_{1 \leq l \leq L}$ and the source signals $(s_1(t), \ldots, s_L(t))_{1 \leq t \leq T}$.*

The approach to this problem is the following: (1) estimate the mixing parameters using the stronger W-disjoint orthogonality assumption and the ML estimator as in e.g. [11], and (2) estimate the source signals under the generalized W-disjoint orthogonality assumption. The latter is presented next.

3. TWO ESTIMATORS OF SIGNALS

In this section we derive the maximum likelihood estimator of source signals, as well as an "ad-hoc" estimator of signals, both under the assumption (2). At every TF point (k, ω) there is a subset of N indices, $\Pi = \{j_1, \ldots, j_N\} \subset \{1, 2, \ldots, L\}$, that specifies which signals are allowed to be nonzero. Beside this, there are exactly N complex unknown variables, $\mathbf{R} = (R_1, \ldots, R_N)$ that define the values of the active signals:

$$\begin{aligned} S_{j_m}(k, \omega) &= R_m(k, \omega) \ , \ 1 \leq m \leq N \qquad (10) \\ S_j(k, \omega) &= 0 \ , \ j \notin \Pi \qquad (11) \end{aligned}$$

Hence the unknown source signals are uniquely defined by $(\Pi, \mathbf{R})$.

3.1. The ML Estimator of $(\Pi, \mathbf{R})$

Given the mixing parameters $(\tau_l)_{1\le l\le L}$, the likelihood of the source signal $(\Pi, \mathbf{R})$ is then

$$\mathcal{L}(\Pi,\mathbf{R}) = \prod_{(k,\omega)} \prod_{d=0}^{D-1} \frac{1}{\pi\sigma^2} exp\{-\frac{1}{\sigma^2}|X_{d+1}(k,\omega) - Y_d(k,\omega)|^2\} \tag{12}$$

where

$$Y_d(k,\omega) = \sum_{l=1}^{N} e^{-id\tau_{j_l(k,\omega)}\omega} R_l(k,\omega)$$

Taking the logarithm and rearranging the expression, we get that $(\Pi, \mathbf{R})$ is the minimizer of:

$$min_{\Pi,R}\; I(\Pi,\mathbf{R}) = \sum_{(k,\omega)} \sum_{d=0}^{D-1} |X_{d+1}(k,\omega) - Y_d(k,\omega)|^2 \tag{13}$$

Then $\mathbf{R}$ is easily obtained at every TF point (k,ω) as a least square solution, namely

$$\hat{R} = (\mathbf{M}^*\mathbf{M})^{-1}\mathbf{M}^* X \tag{14}$$

where $\mathbf{M}$ is the $D\times N$ matrix $\mathbf{M}_{d,l} = e^{-id\tau_{j_l}\omega}$, $0 \le d \le D-1, 1 \le l \le N$. Using Vandermonde determinats, one can show the matrix $\mathbf{M}^*\mathbf{M}$ is invertible if and only if $N \le D$ and $\omega\tau_l \neq \omega\tau_f (mod\, 2\pi)$, for all $l \neq f$. Assume from now on we are in such a case (for instance by choosing $N < D$ and $(\tau_l)_l$ to be distinct from one another and smaller than 1). Note the optimal solution depends on Π through the choice of indices $(j_l)_l$. Next we replace $\hat{R}$ into (13) and the minimization of I turns into the maximization of:

$$max_\Pi\; J(\Pi) = X^*\mathbf{M}(\mathbf{M}^*\mathbf{M})^{-1}\mathbf{M}^*X \tag{15}$$

over all L-choose-N objects. The geometric interpretation of $J(\Pi)$ is the following: it represents the size of the projection of X onto the span of columns of $\mathbf{M}$, $J(\Pi) = \|P_{\mathbf{M}}X\|^2$. Hence the optimal choice $\hat{\Pi}$ represents the closest N-dimensional subspace of $\mathbf{C}^D$ to X among all $\begin{pmatrix} L \\ N \end{pmatrix}$ subspaces spanned by different combinations of N columns of the matrix $\mathbf{M}$.

Solving $max\, J(\Pi)$ is in general a computationally expensive problem, since it requires generating all $\begin{pmatrix} L \\ N \end{pmatrix}$ combinations of columns of $\mathbf{M}$ and computing $J(\Pi)$ for each of them. In [11] we presented a solution for the case $N = 1$. Similarly, for $N = D-1$ and $L = D$ we obtain also a simple solution using the following observation. If $j \in \{1,\dots,L\}$ denotes the missing index in Π, then $J(\Pi) = \|X\|^2 - |a_jX|^2/\|a_j\|^2$ where a_j is the j^{th} row of the $D\times D$ matrix $\mathbf{Q}$, $Q_{d,j} = e^{-id\tau_j\omega}$, $1 \le d,j \le D$.

The algorithm can be modified to deal with an echoic mixing model, or different array configurations at the expense of increased computational complexity. It requires knowledge of the number of sources, however this number is not limited to the number of sensors. It works also in non-square case. The algorithm is guaranteed to converge to a local minimum only.

Since we used (6) as the stochastic limit of (5), the signal estimator we derive is *the maximum á posteriori* with respect to the prior joint p.d.f. (6). However, if one adopts the deterministic point of view regarding (2), our estimator is truly the maximum likelihood estimator.

3.2. An ad-hod estimator of $(\Pi, \mathbf{R})$

We have also derived and used for comparison another estimator of source signals. This estimator is obtained by noticing that the estimates of the source signals have to satisfy the N-term W-disjoint orthogonality hypothesis and they have to fit as well as possible in (7). With these constraints in mind, we implemented a second estimator as follows. For each subset $\Pi = \{j_1,\dots,j_N\}$ of $\{1,2,\dots,L\}$ and every subset $\Gamma = \{g_1,\dots,g_N\} \subset \{1,2,\dots,D\}$ both of N elements, we solve the linear system

$$X_{g_l}(k,\omega) = \sum_{f=1}^{N} e^{-i(g_l-1)\tau_{j_f}\omega} R^{\Gamma,\Pi}_{j_f}(k,\omega)\;\;,\;\; 1 \le l \le N \tag{16}$$

Then average the estimates for some source index j over all subsets Γ,

$$\tilde{R}^{\Pi}_j = \frac{1}{\sum_\Gamma w(\Gamma)} \sum_\Gamma w(\Gamma) R^{\Gamma,\Pi}_j(k,\omega) \tag{17}$$

where the weight w is chosen as $w(\Gamma) = 1/\sqrt{\sum_{g\in\Gamma} g^2}$ because we assume the errors are larger for microphones further away from microphone 1. Next we compute the mean square error

$$K(\Pi) = \frac{1}{\sum_\Gamma w^2(\Gamma)\sum_{j\in\Gamma}|\tilde{R}^\Pi_j|^2} \sum_\Gamma w^2(\Gamma) \sum_{j\in\Gamma} |\tilde{R}^\Pi_j - R^{\Gamma,\Pi}_j|^2$$

and the optimal subset Π of N active sources is estimated by minimizing

$$\tilde{\Pi} = \text{argmin}_\Pi K(\Pi) \tag{18}$$

The signal estimator is then defined by $\tilde{S}_j = \tilde{R}^{\tilde{\Pi}}_j$.

4. EXPERIMENTAL RESULTS

We have implemented the two estimators described and applied them on realistic voice mixtures generated with a ray tracing model. Our main goal is to compare the performance of the approach when N (number of sources active simultaneously) increases. Will this degrade performance, or on the contrary, enhance separation at the cost of increased computational complexity?

Mixtures consisted of four source signals in different room environments and Gaussian noise. The room size was $4\times 5\times 3.2$ m. We used setups corresponding to anechoic and echoic mixing with reverberation time 130 ms. The microphones formed a linear array with 2 cm spacing. Source signals were distributed in the room. Input signals were sampled at 16KHz. For time-frequency representation we used a Hamming window of 512 samples and 50% overlap. Noise was added on each channel. The average (individual) signal-to-noise-ratio (SNR) was 10 dB, while the average input signal-to-interference-ratio (SIR) was about -4.7 dB.

To compare results, we used three criteria: output average signal to interference ratio gain (includes other voices and noise), signal distortion, and mean opinion intelligibility score. The first two are defined as follows:

$$\text{SIRgain} = \frac{1}{N_f}\sum_{k=1}^{N_f} 10\log_{10}\left(\frac{\|S_o\|^2}{\|\hat{S}-S_o\|^2}\frac{\|X-S_i\|^2}{\|S_i\|^2}\right) \tag{19}$$

$$\text{distortion} \;=\; \frac{1}{N_f}\sum_{k=1}^{N_f} 10\log_{10}\frac{\|S_o-S_i\|^2}{\|S_i\|^2} \tag{20}$$

where: N_f is the number of frames where the summand is above -10 dB for SIR gain, and -30 dB for distortion; $\hat{S}$ is the estimated signal that contains S_o contribution of the original signal; X is the mixing at sensor 1, and S_i is the input signal of interest at sensor 1. The summands were saturated at $+30$ dB for SIR gain and $+10$ dB for distortion. Ideally, $\mathrm{SIRgain}$ should be a large positives, whereas $\mathrm{distortion}$ should be a large negative.

We performed tests on noisy data for which SIR level for each source is approximately -4.7 dB, while noise determines an SNR level for the average voice on a channel of 10 dB. Figure 1 shows plots of the wav files of interest for a run of the algorithm where the mixing parameters were obtained using the algorithm described in [11] (Step 1 of the present approach, assuming at most one source active at any time frequency point), while Π and source estimation parameters were determined using the implementation of the present estimators. Average SIR gains shows a degradation in performance from $N = 1$ to $N = 2$, and from anechoic to echoic data (See Fig. 2). However, mean intelligibility scores are best when the number of sources simultaneously active at any time frequency point is $N = 2$. By increasing the number of active sources from $N_1 = 1$ to $N_2 = 2$, the total noise power in the outputs increases by a factor of $\frac{N_2}{N_1}$, or $3dB$, explaining some of the drop in SIR gain in Fig. 2. We conjecture that the sweet spot in separation is when N is a small fraction of the total number of sources.

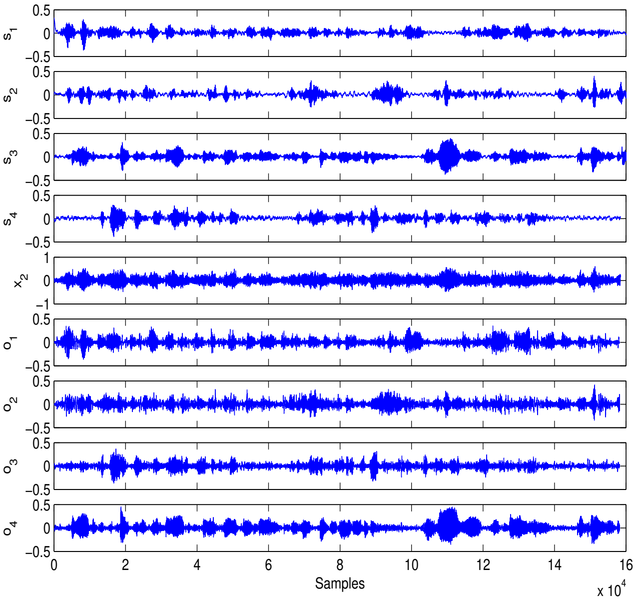

Fig. 1. Example of 8-channel ML algorithm behavior on mixture of noise and four voices each at approximately -4.7 dB input SIR. The first four plots represent the original inputs. The fifth row gives the mixture on channel 2. The separated outputs are presented in the last four rows.

5. CONCLUSIONS

We show that a small number of simultaneously active sources in time-frequency domain is justifiable from a stochastic perspective. This hypothesis, called *generalized W-disjoint orthogonality*, extends the model studied in [11], and is obtained as an asymptotic approximation in the expansion of the joint pdf of sparse sources.

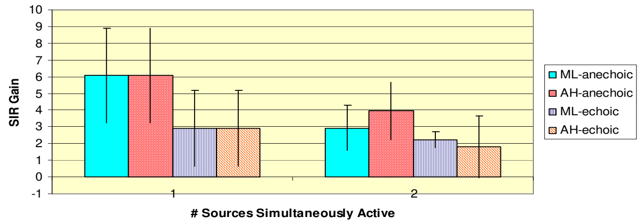

Fig. 2. Average SIR gains and one standard deviation bars for anechoic and echoic experiments with implementations of the ML and Ad-Hoc estimators.

Our source separation algorithm implements both the ML and a heuristic estimator for source signals under a direct-path mixing model and for a linear array of sensors in the presence of noise. Tests with the algorithm on noisy mixtures show that the perceptual quality of separated signals improves at the expense of a smaller reduction in the noise by assuming that two signals are active simultaneously at every time-frequency point rather than one. Future work will integrate these signal estimators with a mixing parameter estimator.

6. REFERENCES

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” Tech. Rep., Stanford University, 1995.

[2] D.L. Donoho and X. Huo, “Uncertainty principles and ideal atomic decomposition,” *IEEE Trans IT*, vol. 47, no. 7, pp. 2845–2862, 2001.

[3] M. Zibulevsky and B. Pearlmutter, “Blind source separation by sparse decomposition in a signal dictionary,” *Neural Computation*, vol. 13(4), pp. 863–882, 2001.

[4] M. Lewicki and T. Sejnowski, “Overcomplete representations,” in *Advances in Neural Information Processing Systems 9*. 1998, MIT Press.

[5] B.A. Olshausen and D.J. Field, “Sparse coding with an overcomplete basis set: A strategy employed by v1,” *Vision Res.*, vol. 37(23), 1997.

[6] A.J. Bell and T.J. Sejnowski, “Learning the higher order structure of a natural sound,” *Network: Computation in Neural Systems*, vol. 7, 1996.

[7] Y. Li, A. Cicochi, and S. Amari, “Sparse component analysis for blind source separation with less sensors than sources,” in *Fourth Int. Symp. on ICA and BSS*, S. Amari, A. Cicochi, S. Makino, and N. Murata, Eds., 2003, pp. 89–94.

[8] A. Jourjine, S. Rickard, and O. Yilmaz, “Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures,” in *ICASSP*, 2000.

[9] R. Balan and J. Rosca, “Statistical properties of STFT ratios for two channel systems and applications to blind source separation,” in *Proc. ICA-BSS*, 2000.

[10] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, “Blind separation of more speech than sensors with less distortion by combining sparseness and ica,” in *IWAENC2003*, 2003, pp. 271–274.

[11] R. Balan, J. Rosca, and S. Rickard, “Scalable non-square blind source separation in the presence of noise,” in *ICASSP2003, Hong-Kong, China*, April 2003.

[12] S. Rickard, R. Balan, and J. Rosca, “Real-time time-frequency based blind source separation,” in *Proc. ICA*, 2001, pp. 651–656.

[13] S. Rickard and O. Yilmaz, “On the W-disjoint orthogonality of speech,” in *Proc. ICASSP*, 2002, vol. 1, pp. 529–532.