NON-SQUARE BLIND SOURCE SEPARATION UNDER COHERENT NOISE BY BEAMFORMING AND TIME-FREQUENCY MASKING

Radu Balan, Justinian Rosca, Scott Rickard

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

ABSTRACT

To be applicable in realistic scenarios, blind source separation approaches should deal evenly with non-square cases and the presence of noise. We consider an additive noise mixing model with an arbitrary number of sensors and possibly more sources than sensors (the non-square case) when sources are disjointly orthogonal. We formulate the maximum likelihood estimation of the coherent noise model, suitable when sensors are nearby and the noise field is close to isotropic, and also under the direct-path far-field assumptions. The implementation of the derived criterion involves iterating two steps: a partitioning of the time-frequency plane for separation followed by an optimization of the mixing parameter estimates. The structure of the solution is surprising at first but logical: it consists of a beamforming linear filter, which reduces noise, and a filter across time-frequency domain to separate sources. The solution is applicable to an arbitrary number of microphones and sources. Experimentally, we show the capability of the technique to separate four voices from two, four, six, and eight channel recordings in the presence of isotropic noise. [1]

1. INTRODUCTION

Source separation promises to further a variety of applications of speech enhancement and separation beyond what is possible today with classical microphone array techniques [1]. In particular for audio signals (the domain of interest in this work), a variety of BSS techniques have been introduced in recent years. Few work on real audio data (e.g. [2, 3, 4]), even fewer with noisy data [5], and most deal with the "square" case of source separation (equal number of sources and sensors). Claims of generalization to the non-square case exist, however most often it is not clear how techniques would scale, neither from an algorithmic perspective nor in terms of computational properties.

[6] introduced a BSS technique for the separation of an arbitrary number of sources from just *two* mixtures provided the time-frequency representations of sources do not overlap. The key observation in the technique is that each time-frequency (TF) point depends on at most one source and its associated mixing parameters. This deterministic hypothesis was called *W-disjoint orthogonality* and is reviewed in section 2.2. 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. Such an approach was used in [7] as well.

The deterministic signal model was extended to a stochastic signal model in [8], 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.

In contrast to the case of [9], in this paper we analyze the estimators when noise comes from an isotropic diffuse noise field, as studied in differential microphone microphone array literature [1]. Such a model is consistent with the assumption about the microphone array geometry, whereby microphone spacing is as small as one centimeter.

The ICA literature scarcely discusses the noise case [10]. BSS and deconvolution results of a theoretical nature in dealing with noise were presented in [5]. For the two-channel system in [4], the ML estimator of the mixing parameters was derived in the presence of Gaussian sensor noise. However the noise element represented a technicality in that it was considered in the limit zero in order to be able to derive parameter update equations. Nonetheless the approach proved effective on real non-noisy data.

In this paper we deal with the multi-channel case from an algorithmic perspective. We present a novel approach to BSS exploiting TF properties of the input data, and of the noise, which are readily applied to speech separation on two, four, and six channels. For this, we extend the ML

[1] In Proceedings of the 4th ICA-BSS Conference, April 2003, Japan

estimators derived before (under the W-disjoint orthogonality assumption). The ML approach considers both mixing parameters and sources, unlike in [4] where the optimization was over mixing parameters only. The estimation algorithm iterates two optimization steps. First, likelihood is optimized over the set of mixing parameters for each source separately. Second the partition of TF points is optimized. Unlike [9] here we consider and compare the isotropic noise field (characterized by a sinc coherence matrix) to the uncorrelated noise field, and show the gains offered by taking into account the right noise model. For the purposes of this paper we consider the anechoic mixing model only. However the method presented can be extended to arbitrary complex mixing models.

The organization of the paper is as follows. Section 2 presents the signal mixing model and a statistical motivation of the W-disjoint orthogonality signal model. Section 3 shows the derivation of the ML estimator of mixing parameters and source signals, and its implementation by an iterative procedure. Section 4 experimentally highlights the capability of the system to deal with noisy echoic data, and its scaling properties. Experiments with two, four, six, and eight inputs show increased separation capability and decreased artifacts with an increase in the number of inputs on data ranging from anechoic to echoic.

2. MIXING MODEL AND SIGNAL ASSUMPTION

2.1. The Mixing Model

Consider the measurements of L source signals by a equispaced linear array of D sensors under 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:

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

where $n_1, \ldots, n_D$ are the sensor noises, and $(a_{d,l}; \tau_{d,l})$ are the attenuation and delay parameters of source l to sensor d. For the far-field model and equispaced sensor array, the attenuations $a_{d,l}$ and delays $\tau_{d,l}$ are linearly distributed across the sensors (i.e. with respect to index d). Thus we can define the average attenuation a_l, and delay τ_l, so that

$$
a_{d,l} = (d-1)a_l, \quad \tau_{d,l} = (d-1)\tau_l, \quad 1 \le d \le D, 1 \le l \le L \tag{2}
$$

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 $|\tau_l| \le \Delta, \forall l$.

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 (1) turns into

$$
X_d(k, \omega) = \sum_{l=1}^{L} (1 - (d-1)a_l) e^{-i\omega(d-1)\tau_l} S_l(k, \omega) + N_d(k, \omega) \tag{3}
$$

or, more compactly,

$$
X(k, \omega) = \sum_{l=1}^{L} Z_l(\omega) S_l(k, \omega) + N(k, \omega) \tag{4}
$$

with

$$
Z_l(\omega) = \begin{bmatrix} 1 & (1 - a_l)e^{-i\omega\tau_l} & \cdots & (1 - (D-1)a_l)e^{-i\omega(D-1)\tau_l} \end{bmatrix}^T \tag{5}
$$

and X, N the D-vectors of measurements, respectively noises. When no danger of confusion arises, we drop the arguments k, ω.

We assume the noise is Gaussian distributed with a covariance matrix of the form

$$
R_n = \sigma^2 \Gamma_n \tag{6}
$$

where σ^2 is the average noise field spectral power, and Γ_n the coherence matrix. The uncorrelated noise field is characterized by the identity matrix,

$$
\Gamma_n = I_D \tag{7}
$$

whereas the isotropic, diffuse noise field has the coherence matrix given by (see [1])

$$
\begin{bmatrix}
1 & \mathrm{sinc}(\omega\tau_{max}) & \cdots & \mathrm{sinc}(\omega\tau_{max}(D-1)) \\
\mathrm{sinc}(\omega\tau_{max}) & 1 & \cdots & \mathrm{sinc}(\omega\tau_{max}(D-2)) \\
\vdots & & \ddots & \vdots \\
\mathrm{sinc}(\omega\tau_{max}(D-1)) & \cdots & \mathrm{sinc}(\omega\tau_{max}) & 1
\end{bmatrix} \tag{8}
$$

Our problem is: given measurements $(x_1(t), \ldots, x_D(t))_{1 \le t \le T}$ of the system (1) we want to determine the ML estimates of the mixing parameters $(a_l, \tau_l)_{1 \le l \le L}$ and the source signals $(s_1(t), \ldots, s_L(t))_{1 \le t \le T}$ in the presence of isotropic diffuse noise. When the number of sources is greater than the number of mixtures the problem is degenerate. In order to solve this we rely on the W-disjoint orthogonality assumption.

2.2. The W-Disjoint Orthogonal Signal Model

Two signals s_1 and s_2 are called *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:

$$
S_1(k, \omega) S_2(k, \omega) = 0 \;\; , \;\; \forall k, \omega \tag{9}
$$

For L sources $S_1, \ldots, S_L$ the definition generalizes to:

$$
S_i(k, \omega) S_j(k, \omega) = 0 \;\; , \;\; \forall\, 1 \le i \ne j \le L, \; \forall k, \omega \tag{10}
$$

Disjoint orthogonality has been extensively studied as the basis for time-frequency processing in [11]. Relation (9) does indeed hold in an approximate sense for real speech signals and a large class of real signals. In [9] we additionally proved that (9) can be seen as the limit of a stochastic model introduced in [8]. In this paper we assume that (10) is satisfied for all practical purposes. In addition, we assume that noise is Gaussian distributed with zero mean and coherence given by (8).

3. THE MAXIMUM LIKELIHOOD ESTIMATOR OF SIGNAL AND MIXING PARAMETERS

In this section we derive the joint maximum likelihood estimator of parameters and source signals under assumption 10. The source signals naturally partition the time-frequency plane into L disjoint subsets $\Omega_1, \ldots, \Omega_L$, where each source signal is non-zero (i.e. active). Thus the signals are given by the collection Ω_1,. . .,Ω_L and one complex variable S that defines the active signal:

$$S_l(k,\omega) = S(k,\omega)1_{\Omega_l}(k,\omega) \tag{11}$$

Let the model parameters θ consist of the mixing parameters (a_l, τ_l), $1 \leq l \leq L$, the partition $(\Omega_l)_{1\leq l\leq L}$ and S. Based on equations 4 and 6, its likelihood and maximum log-likelihood estimator are given by:

$$\mathcal{L}(\theta) = \prod_{l=1}^{L} \prod_{(k,\omega)\in\Omega_l} \frac{1}{\pi^D \sigma^{2D}} \exp\{-\frac{1}{\sigma^2} Y_l^*(k,\omega)\Gamma_n^{-1}(\omega)Y_l(k,\omega)\}$$
$$\hat{\theta}_{ML} = \text{argmin}_\theta \sum_{l=1}^{L} \sum_{(k,\omega)\in\Omega_l} Y_l^*(k,\omega)\Gamma_n^{-1}(\omega)Y_l(k,\omega)\} \tag{12}$$

where $Y_l(k,\omega) = X(k,\omega) - Z_l(\omega)S_l(k,\omega)$. For any partition $(\Omega_1, \ldots,\Omega_L)$ we define the selection map Σ : TF-plane $\rightarrow$ $\{1, \ldots, L\}$, $\Sigma(k,\omega) = l$ iff $(k,\omega) \in \Omega_l$. Clearly Σ defines a unique partition. Optimizing over S in (12) we obtain

$$\hat{S} = \frac{Z_l^* \Gamma_n^{-1} X}{Z_l^* \Gamma_n^{-1} Z_l} \tag{13}$$

where $l = \Sigma(k,\omega)$. Let us denote by $A = (a_l, \tau_l)_{1\leq l\leq L}$ the mixing parameters. Inserting (13) into (12), the optimization problem reduces to:

$$(\hat{A}, \hat{\Sigma}) = \text{argmax}_{A,\Sigma} J(A, \Sigma) \tag{14}$$

where:

$$J(A,\Sigma) = \sum_{(k,\omega)} \frac{|Z_{\Sigma(k,\omega)}^* \Gamma_n^{-1} X(k,\omega)|^2}{Z_{\Sigma(k,\omega)}^* \Gamma_n^{-1} Z_{\Sigma(k,\omega)}} \tag{15}$$

Note the criterion to maximize depends on a set of continuous parameters A, and a selection map Σ. A typical optimization algorithm for such a criterion works as follows. The optimization is done in two steps: first the optimization over the continuous parameters, and then the optimization over the selection map (or, equivalently, the partition). Such a procedure is iterated until the criterion reaches a saturation floor. Because the criterion is bounded above, we are guaranteed it will converge. Next we describe solutions for the two optimization problems.

3.1. Optimal Partition

Given a set of mixing parameters, $A = (a_l, \tau_l)_{1\leq l\leq L}$, the optimal selection map is simply given by

$$\hat{\Sigma}(k,\omega) = \text{argmax}_l \frac{|Z_{\Sigma(k,\omega)}^* \Gamma_n^{-1} X(k,\omega)|^2}{Z_{\Sigma(k,\omega)}^* \Gamma_n^{-1} Z_{\Sigma(k,\omega)}} \tag{16}$$

The partition is then immediate: $\Omega_l = \{(k,\omega)|\Sigma(k,\omega) = l\}$.

3.2. Optimal Mixing Parameters

Now given a partition $(\Omega_l)_{1\leq l\leq L}$, the optimal mixing parameters are obtained independently for each l by:

$$(\hat{a_l}, \hat{\tau_l}) = \text{argmax}_{a_l,\tau_l} \sum_{(k,\omega)\in\Omega_l} \frac{|Z_{\Sigma(k,\omega)}^* \Gamma_n^{-1} X(k,\omega)|^2}{Z_{\Sigma(k,\omega)}^* \Gamma_n^{-1} Z_{\Sigma(k,\omega)}} \tag{17}$$

Note that both the denominator and numerator depend on ω, unlike the independent noise case when the numerator was independent of k and ω (see [9]). Therefore a 2-dimensional optimization procedure is required in order to solve (17). In the numerical simulations presented next we used the gradient descent method to search for the optimum.

Summing these findings, the optimization algorithm becomes:

3.3. ML Algorithm

- Step 0. Initialize $(a_l^0, \tau_l^0)_{1\leq l\leq L}$ with random values so that $|a_l^0| < 1$ and $|\tau_l^0| < \Delta$; Set $s = 0$, $J^s = 0$, and choose a stopping threshold ϵ;
- Step 1. Find the optimal partition $(\Omega_l^{s+1})_{1\leq l\leq L}$, and selection map, Σ^{s+1} by solving (16) with $a_l = a_l^s$, $\tau_l = \tau_l^s$;
- Step 2. Apply gradient descent to (17) until it converges to a local optimum $(a_l^{s+1}, \tau_l^{s+1})$ for each $1 \leq l \leq L$, and subset of time-frequency points Ω_l^{s+1};
- Step 3. Set $s = s + 1$, and compute $J^s = J(A^s, \Sigma^s)$. If $(J^s - J^{s-1})/J^s > \epsilon$ then go to Step 1; otherwise:
- Step 4. Estimated parameters after s iterations are $a_l = a_l^s$, $\tau_l = \tau_l^s$, and $\Omega_l = \Omega_l^s$. The source signal are then computed by converting the estimated time-frequency representations back into the time domain.

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 the non-square case.

The solution (11,13) can be understood in the following way. Once the mixing parameters have been estimated, we apply two independent linear filters. One linear filter is across the spatial channels (13) and performs a beamforming in order to reduce the output noise. The other (11) is across time-frequency domain and solves the source separation problem by selecting those time-frequency points where, by our W-disjoint orthogonality assumption, only one source is active.

4. EXPERIMENTAL RESULTS

We implemented the algorithm and applied it to realistic synthetic mixtures generated with a ray tracing model. 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 four setups corresponding to anechoic mixing, low echoic (reverberation time $18\,\mathrm{ms}$), echoic (reverberation time $130\,\mathrm{ms}$), and strong echoic (reverberation time $260\,\mathrm{ms}$). 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 256 samples and 50% overlap. Coherent noise was added on each channel. The average input signal-to-interference-ratio (SIR) was about $-5\,\mathrm{dB}$. The average individual signal-to-noise-ratio (SNR) was $10\,\mathrm{dB}$ (i.e. SNR of one source with respect to noise only). Each test was performed three times with independent noise realizations that were filtered to the isotropic diffuse noise coherence.

The optimization problem (17) was solved by performing 30 gradient descent steps at each iteration (Step 2 of the algorithm). Experimentally, the optimization algorithm converged very fast. In at most five iterations it reached 0.1% of the local maximum. Also experimentally, we noticed the algorithm converges more often to the true direct-path parameters when we add small noise to the diagonal of (8). In fact we chose Γ_n as the sum between (8) and 0.01 times the identity matrix.

In the following we present the results obtained as described above. To compare results, we used three criteria: output average signal to interference ratio gain (includes other voices and noise), segmental SNR, and signal distortion, defined as follows:

$$\mathrm{SIRgain} = \frac{1}{N_f}\sum_{k=1}^{N_f} 10\log_{10}\left(\frac{\| S_o \|^2}{\| \hat{S} - S_o \|^2}\frac{\| X - S_i \|^2}{\| S_i \|^2}\right) \tag{18}$$

$$\mathrm{segSNR} = \frac{1}{N_f}\sum_{k=1}^{N_f} 10\log_{10}\frac{\| S_i \|^2}{\| \hat{S} - S_i \|^2} \tag{19}$$

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

where: $\hat{S}$ 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; N_f is the number of frames where the summand is above $-10\,\mathrm{dB}$ for SIR gain and segmental SNR, and $-30\,\mathrm{dB}$ for distortion. The summands for SIR gain and segmental SNR computation were saturated at $+30\,\mathrm{dB}$, and were saturated at $+10\,\mathrm{dB}$ for distortion. Ideally, $\mathrm{SIRgain}$ should be a large positives, whereas $\mathrm{distortion}$ should be a large negative.

SIR gains are presented in Figure 1, segmental SNR in Figure 2, and the distortion values are given in Table 1. Results show separation of all voices particularly for $D \geq 4$. A sample of input and outputs for $D = 4$ is given in Figure 3. Also SIR gains tend to improve with an increase in the number of sensors. This indicates that separation power of the system increases. Also, one can notice a decrease in performance as we move from anechoic to echoic data. Interestingly, the 8 microphone setup seems to increase by little (if any) compared to the 6-mic case. This seems to be due to the simplified mixing model (enforcing an anechoic model, when in fact it is echoic).

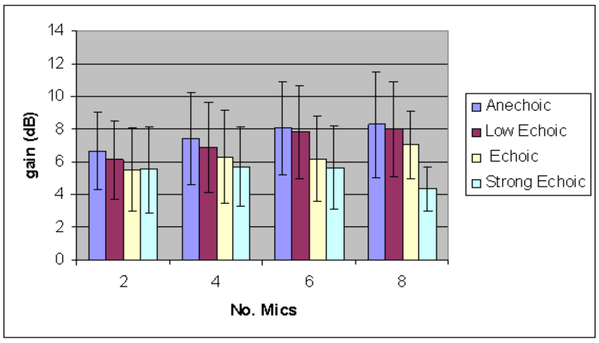

Fig. 1. SIR gains for 2-8 microphones on four data types (anechoic, low echoic, echoic, and strongly echoic). Each bar includes one standard deviation bounds.

5. CONCLUSIONS

Real source separation scenarios are rarely square. On the contrary, situations constantly vary between the so called

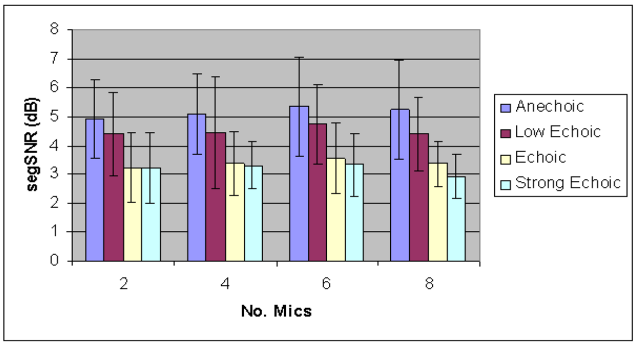

Fig. 2. Segmental SNRs for 2-8 microphones on four data types (anechoic, low echoic, echoic, and strongly echoic). Each bar includes one standard deviation bounds.

D	$Anechoic$	$LowEch$	$Echoic$	$StrongEch$
2	-3.98 (1.35)	-3.49 (1.17)	-2.58 (0.92)	-2.61 (1.01)
4	-4.36 (1.41)	-3.69 (1.53)	-2.79 (0.92)	-2.70 (0.78)
6	-4.43 (1.68)	-3.74 (1.10)	-2.88 (0.93)	-2.61 (0.85)
8	-4.36 (1.71)	-3.57 (1.18)	-2.61 (0.73)	-2.01 (0.50)

Table 1. Distortions for -5dB input SIR and 10dB individual input SNR: mean (standard deviation) for $D = 2, 4, 6, 8$.

degenerate case and the over specified case. Particularly for small microphone arrays, noise is coherent. By being able to deal evenly with such cases and in the presence of coherent noise, the present approach opens the door to audio source separation in realistic scenarios.

This was possible by exploiting the time frequency sparseness of signals within the more general noisy signal model. Our source separation algorithm implements the maximum likelihood estimator for both mixing parameters and source signals under a direct-path mixing model and for a linear array of sensors. We presented an iterative procedure to optimize the likelihood, similar in spirit to hybrid optimization algorithms. Interestingly enough, the optimal solution consists of a beamforming filter, which reduces output noise, followed by time-frequency processing for source separation.

The resulting algorithm exhibits nice scaling properties both algorithmically and experimentally. The former refers to scalability in the number of inputs (here we used two, four, six, and eight microphone linear arrays). The latter views the increased separation power on echoic data (as showed by SIR gain and segmental SNR) at decreasing or

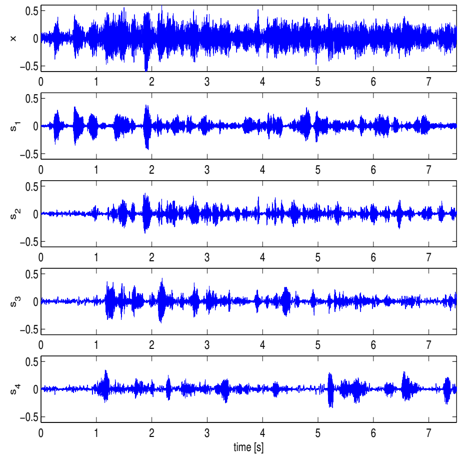

Fig. 3. Example of 6-channel algorithm behavior on mixture of coherent noise and four voices. The separated outputs are s_1-s_4.

relatively constant artifacts with an increase in the number of inputs.

Future work could address the question whether anything is to be gained by considering an echoic model. This extension is naturally feasible in this approach.

6. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[2] L.Parra, "Convolutive blind source separation based on multiple decorrelation," in *IEEE-ICNN*, 1997.

[3] Jorn Anemuller and Birger Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proceedings of the second international workshop on independent component analysis and blind signal separation*, Petteri Pajunen and Juha Karhunen, Eds., Helsinki, Finland, June 19–22 2000, pp. 215–220.

[4] S. Rickard, R. Balan, and J. Rosca, "Real-time blind source separation using DUET," in *3rd International Conference on Independent Component Analysis and Blind Source Separation (ICA2001), San Diego, CA*, December 2001.

[5] E. Moulines, J.F. Cardoso, and E. Gassiat, "Maximum likelihood for blind source separation and deconvolution of noisy signals using mixture models," in *Proceedings ICASSP*. 1997, pp. 3617–3720, IEEE Press.

[6] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2000, IEEE Press, June 5-9, 2000, Istanbul, Turkey.

[7] M. Aoki, M. Okamoto, S. Aoki, and H. Matsui, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. & Tech.*, vol. 22, no. 2, pp. 149–157, 2001.

[8] R. Balan and J. Rosca, "Statistical properties of STFT ratios for two channel systems and applications to blind source separation," in *Proceedings ICA 2000, Helsinki*, Petteri Pajunen and Juha Karhunen, Eds. 2000, pp. 429–434, Otamedia, Helsinki, Finland, June 2000.

[9] R. Balan, J. Rosca, and S. Rickard, "Scalable non-square blind source separation in the presence of noise," in *sent to IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2003), Hong-Kong, China*, April 2003.

[10] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley and Sons, 2001.

[11] S. Rickard and O. Yilmaz, "On the W-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002), Orlando, Florida, USA*, May 2002.