A TWO-STAGE FREQUENCY-DOMAIN BLIND SOURCE SEPARATION METHOD FOR UNDERDETERMINED CONVOLUTIVE MIXTURES

Hiroshi Sawada, Shoko Araki and Shoji Makino

NTT Communication Science Laboratories
NTT Corporation
2-4 Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{sawada, shoko, maki}@cslab.kecl.ntt.co.jp

ABSTRACT

This paper proposes a two-stage method for the blind separation of convolutively mixed sources. We employ time-frequency masking, which can be applied even to an underdetermined case where the number of sensors is insufficient for the number of sources. In the first stage of the method, frequency bin-wise mixtures are classified based on Gaussian mixture model fitting. In the second stage, the permutation ambiguities of the bin-wise classified signals are aligned by clustering the posterior probability sequences calculated in the first stage. Experimental results for separating four speeches with three microphones under reverberant conditions show the superiority of the proposed method over existing methods based on time-difference-of-arrival estimations or signal envelope clustering.

1. INTRODUCTION

The technique for estimating individual source components from their mixtures at multiple sensors is known as blind source separation (BSS) [1]. With acoustic applications of BSS, such as solving a cocktail party problem, signals are mixed in a convolutive manner. Let $\mathrm{s}_1, \ldots, \mathrm{s}_N$ be source signals and $\mathrm{x}_1, \ldots, \mathrm{x}_M$ be sensor (microphone) observations. The convolutive mixture model is formulated as

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_l \mathrm{h}_{jk}(l)\, \mathrm{s}_k(t-l), \quad j=1,\ldots,M, \tag{1}$$

where t represents time and $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to microphone j.

If we have a sufficient number of microphones (determined case), i.e., $M \geq N$, linear filters that are estimated for example by independent component analysis (ICA) [2] effectively separate the mixtures. However, if the number of microphones is insufficient (underdetermined case), i.e., $M < N$, such linear filters do not work well. Instead, time-frequency (T-F) masking [3, 4, 5] or l_1-norm minimization [6, 7] is widely used to separate such underdetermined mixtures. Most of the existing techniques for underdetermined convolutive BSS rely on time-difference-of-arrival (TDOA) estimations for each source at the microphones. They work effectively under low reverberant conditions. However, under severe reverberant conditions, TDOA estimations become unreliable and such techniques do not work well. Also, any microphone directivity complicates the TDOA estimation.

This paper proposes a new method for underdetermined convolutive BSS. The method follows the frequency-domain approach, and consists of two stages. In the first stage, frequency bin-wise

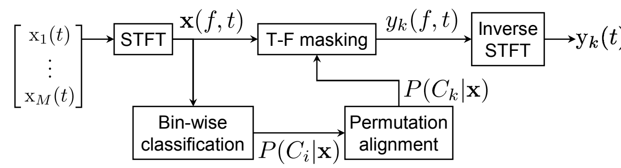

Figure 1: *System flow*

mixtures are classified by the procedure explained in Sec. 3. Since the classification is not based on TDOA estimations, but on a frequency bin-wise instantaneous model, it is robust as regards room reverberations and microphone characteristics as long as the frame size in the frequency analysis is long enough to cover the main part of the impulse responses. In the second stage, the bin-wise classified signals are grouped together for the same original source. This task is almost identical to that of the permutation problem [8, 9] which occurs in frequency-domain ICA-based BSS for a determined case ($M \geq N$). We adapt our recent idea [10], which we developed for a determined case, to an underdetermined case. Section 4 describes the procedure in detail.

2. SYSTEM OVERVIEW

Let us begin with a system overview of the proposed method. Figure 1 shows the system flow, and Fig. 2 shows some examples of spectrograms and classification results for a 3-source 2-microphone case.

2.1. Short-time Fourier transform (STFT)

First, the sensor observations (1) are converted into frequency-domain time-series signals $x_j(f,t)$ by an STFT. The time index t is then down-sampled with a distance equal to the STFT frame shift. If the STFT frame size is long enough to cover the main part of the impulse responses $\mathrm{h}_{jk}(l)$, the time-domain convolutions in (1) can be well approximated with frequency-domain multiplications:

$$x_j(f,t) = \sum_{k=1}^{N} h_{jk}(f) s_k(f,t)\,, \quad j=1,\ldots,M. \tag{2}$$

Let us rewrite (2) in a vector notation:

$$\mathbf{x}(f,t) = \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,t) \tag{3}$$

where $\mathbf{h}_k = [h_{1k}, \ldots, h_{Mk}]^T$ is the vector of frequency responses from source s_k to all sensors, and $\mathbf{x} = [x_1, \ldots, x_M]^T$ is called an observation vector.

978-1-4244-1619-6/07/$25.00 ©2007 IEEE

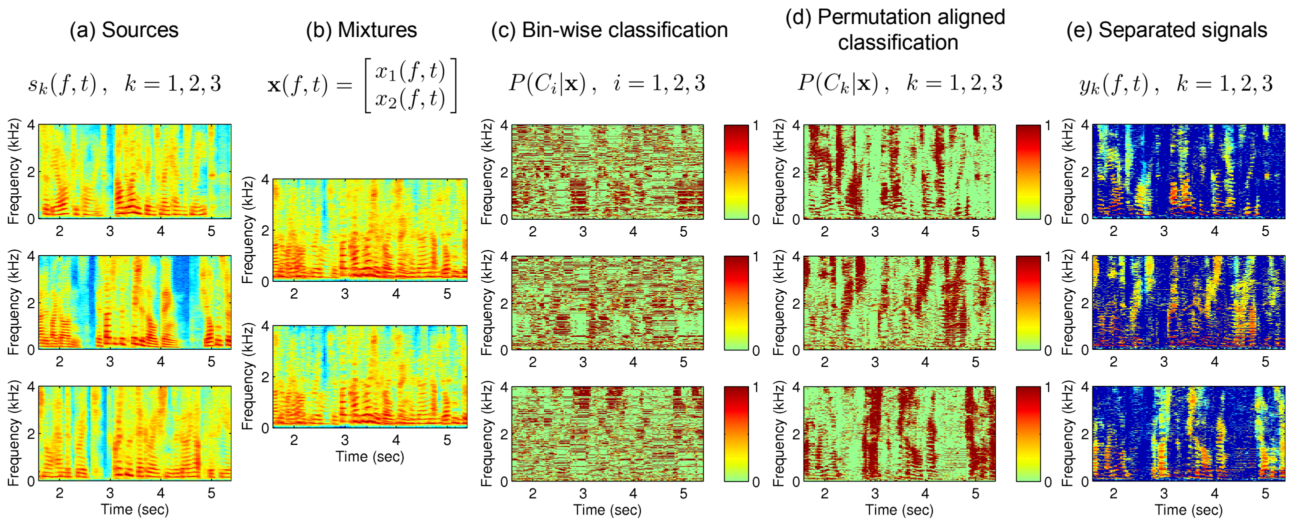

Figure 2: *Example spectrograms for sources (a) mixtures (b) and separated signals (e), and classification results (c) and (d).*

2.2. Time-frequency (T-F) masking and inverse STFT

To separate the mixtures, we employ T-F masking, which is effective for separating sparse sources. A sparse source can be characterized by the fact that the source amplitude is close to zero most of the time. A frequency-domain speech source is a good example. The mixing model (3) can be approximated to a simpler form

$$\mathbf{x}(f,t) \approx \mathbf{h}_k(f) s_k(f,t) \tag{4}$$

for sparse sources s_k. The summation in (3) is eliminated because the sources are assumed to be close to zero most of the time. The subscript k is the index of the most dominant source for each time-frequency slot (f,t). The index k should be identified or estimated for each (f,t) to separate the sources by T-F masking.

Let us classify the observation vectors $\mathbf{x}(f,t)$ into N classes $c_1,\ldots,c_N$, each of which corresponds to a source signal s_k. A vector $\mathbf{x}(f,t)$ should belong to class C_k if the source s_k is the most dominant in the observation $\mathbf{x}(f,t)$. Through the processes of "Bin-wise classification" and "Permutation alignment", which are depicted in Fig. 1 and explained in the next two sections, we calculate the posterior probabilities $P(C_k|\mathbf{x})$ for all source indices k and for all observation vectors $\mathbf{x}(f,t)$. Then, T-F masking for source separation is performed by

$$y_k(f,t) = \begin{cases} x_J(f,t) & \text{if } P(C_k|\mathbf{x}) \geq P(C_{k'}|\mathbf{x})\,, \ \forall k' \neq k \\ 0 & \text{if } P(C_k|\mathbf{x}) < P(C_{k'}|\mathbf{x})\,, \ \exists k' \neq k \end{cases} \tag{5}$$

for $k=1,\ldots,N$ where J is the index of an arbitrarily selected reference microphone that is used to construct the separated signals.

At the end of the flow, time-domain output signals $y_k(t)$ are calculated with an inverse STFT to the frequency-domain separated signals $y_k(f,t)$.

3. BIN-WISE CLASSIFICATION

This section and the next explain how to calculate the posterior probability $P(C_k|\mathbf{x})$ that the k-th source is the most dominant source in an observation vector $\mathbf{x}(f,t)$. The procedure consists of two stages, which are also described in this section and the next.

The first stage classifies observation vectors $\mathbf{x}(f,t)$ in a frequency bin-wise manner. Thus, here we omit the frequency dependence in (3) and (4) for simplicity:

$$\mathbf{x}(t) = \textstyle\sum_{i=1}^{N} \mathbf{h}_i s_i(t) \approx \mathbf{h}_i s_i(t)\,. \tag{6}$$

We changed the source subscript from k to i. This is because there is permutation ambiguity in the frequency bin-wise classifications. The effect of permutations is demonstrated in Fig. 2 (c). Such permutation ambiguities will be aligned in the second stage, which is detailed in the next section.

We see in (6) that classification can be performed according to the information on the vectors $\mathbf{h}_1,\ldots,\mathbf{h}_N$. To remove the effect of source amplitude from $\mathbf{x}$, let us normalize them to have unit norm

$$\mathbf{x}(t) \leftarrow \frac{\mathbf{x}(t)}{||\mathbf{x}(t)||} = \frac{\mathbf{h}_i}{||\mathbf{h}_i||} \cdot \frac{s_i(t)}{|s_i(t)|}\,.$$

Unknown phase $s_i(t)/|s_i(t)|$ ambiguity still remains in $\mathbf{x}(t)$. To model such vectors for each source, we follow the idea described in [6] and employ the following complex Gaussian density function

$$p(\mathbf{x}|\mathbf{a}_i,\sigma_i) = \frac{1}{(\pi\sigma_i^2)^M} \exp\left(-\frac{||\mathbf{x}-(\mathbf{a}_i^H\mathbf{x})\cdot\mathbf{a}_i||^2}{\sigma_i^2}\right) \tag{7}$$

where $\mathbf{a}_i$ is the mean vector with unit-norm $||\mathbf{a}_i||=1$ and σ_i^2 is the variance. Since $(\mathbf{a}_i^H\mathbf{x})\cdot\mathbf{a}_i$ is the orthogonal projection of $\mathbf{x}$ onto the subspace spanned by $\mathbf{a}_i$, $||\mathbf{x}-(\mathbf{a}_i^H\mathbf{x})\cdot\mathbf{a}_i||$ represents the minimum distance between the point $\mathbf{x}$ and the subspace, which implies the probability that $\mathbf{x}$ belongs to the i-th class.

Now the density function for data $\mathbf{x}$ can be given by

$$p(\mathbf{x}|\theta) = \textstyle\sum_{i=1}^{N} \alpha_i\, p(\mathbf{x}|\mathbf{a}_i,\sigma_i) \tag{8}$$

with the parameter set $\theta=\{\mathbf{a}_1,\sigma_1,\alpha_1,\ldots,\mathbf{a}_N,\sigma_N,\alpha_N\}$. The mixture ratios α_i should satisfy $0<\alpha_i<1$ and $\alpha_1+\cdots+\alpha_N=1$. Since there is some computational difficulty involved in maximizing the log-likelihood $\sum_t^T \log p(\mathbf{x}(t)|\theta)$ of T samples directly, we typically employ an EM algorithm to estimate the parameters θ. By introducing the posterior probability

$$P(C_i|\mathbf{x}(t), \theta) = \frac{\alpha_i \, p(\mathbf{x}(t)|\mathbf{a}_i, \sigma_i)}{p(\mathbf{x}(t)|\theta)}, \tag{9}$$

we construct the so-called Q function

$$Q(\theta) = \textstyle\sum_t^T \sum_{i=1}^N P(C_i|\mathbf{x}(t), \theta) \, \log \alpha_i \, p(\mathbf{x}(t)|\mathbf{a}_i, \sigma_i) \tag{10}$$

to be maximized in the EM algorithm.

In the E-step, $P(C_i|\mathbf{x}(t), \theta)$ is calculated by using (9) and (8) based on the current parameters θ. In the M-step, we maximize the Q function by adjusting the parameters in θ. Considering the fact that both $\mathbf{x}$ and $\mathbf{a}_i$ have a unit-norm, the optimal $\mathbf{a}_i$ is given by the eigenvector corresponding to the maximum eigenvalue of

$$\mathbf{R} = \textstyle\sum_t^T P(C_i|\mathbf{x}(t), \theta) \cdot \mathbf{x}(t)\mathbf{x}^H(t) \, .$$

Other parameters are optimized by

$$\sigma_i^2 = \frac{\sum_t^T P(C_i|\mathbf{x}(t), \theta) \cdot ||\mathbf{x}(t) - (\mathbf{a}_i^H \mathbf{x}(t)) \cdot \mathbf{a}_i||^2}{M \cdot \sum_t^T P(C_i|\mathbf{x}(t), \theta)}$$

and $\alpha_i = (1/T) \sum_t^T P(C_i|\mathbf{x}(t), \theta)$.

4. PERMUTATION ALIGNMENT

By performing the operations described in the previous section independently for all frequencies f, we have the posterior probability $P(C_i|\mathbf{x}(f,t))$ according to (9) for $i = 1, \ldots, N$ and all time-frequency slots (f,t). However, since the class order $c_1, \ldots, c_N$ may be different from one frequency to another (Fig. 2 (c)), we need to reorder the indices such that the same index corresponds to the same source over all frequencies (Fig. 2 (d)). In other words, we determine a permutation $\Pi_f : \{1, \ldots, N\} \to \{1, \ldots, N\}$ for all frequencies f, and then update the posterior probabilities by

$$P(C_k|\mathbf{x}) \leftarrow P(C_i|\mathbf{x}) \,\big|_{i=\Pi_f(k)}, \quad k = 1, \ldots, N \, .$$

Such a permutation problem has been extensively studied for frequency-domain ICA-based BSS applied to a determined case ($M \geq N$). The proposed method follows a major approach based on the correlation coefficients $\rho(v_i^f, v_j^g)$ of two sequences $v_i^f(t)$, $v_j^g(t)$ that express the separated signal activities [8, 9, 10]. The correlation coefficient of any two sequences is bounded by $-1 \leq \rho(v_i^f, v_j^g) \leq 1$, and becomes 1 if the two sequences are identical.

Although signal envelopes $v_i^f(t) \leftarrow |y_i(f,t)|$ are widely used [8, 9] to represent the separated signal activities, here we use the posterior probabilities $v_i^f(t) \leftarrow P(C_i|\mathbf{x}(f,t))$ instead. Figures 3 and 4 show examples of separated signal envelopes and posterior probability sequences, respectively. The correlation coefficients calculated with the signal envelopes are

$$\begin{bmatrix} \rho(v_1^f, v_1^g) & \rho(v_1^f, v_2^g) & \rho(v_1^f, v_3^g) \\ \rho(v_2^f, v_1^g) & \rho(v_2^f, v_2^g) & \rho(v_2^f, v_3^g) \\ \rho(v_3^f, v_1^g) & \rho(v_3^f, v_2^g) & \rho(v_3^f, v_3^g) \end{bmatrix} = \begin{bmatrix} 0.01 & 0.07 & -0.10 \\ -0.09 & 0.10 & -0.15 \\ -0.07 & -0.07 & 0.44 \end{bmatrix},$$

and those calculated with the posterior probability sequences are

$$\begin{bmatrix} \rho(v_1^f, v_1^g) & \rho(v_1^f, v_2^g) & \rho(v_1^f, v_3^g) \\ \rho(v_2^f, v_1^g) & \rho(v_2^f, v_2^g) & \rho(v_2^f, v_3^g) \\ \rho(v_3^f, v_1^g) & \rho(v_3^f, v_2^g) & \rho(v_3^f, v_3^g) \end{bmatrix} = \begin{bmatrix} 0.51 & -0.20 & -0.29 \\ -0.25 & 0.46 & -0.23 \\ -0.28 & -0.28 & 0.55 \end{bmatrix}.$$

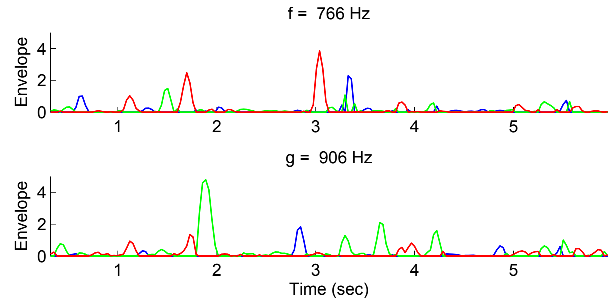

Figure 3: *Envelopes of three separated signals*

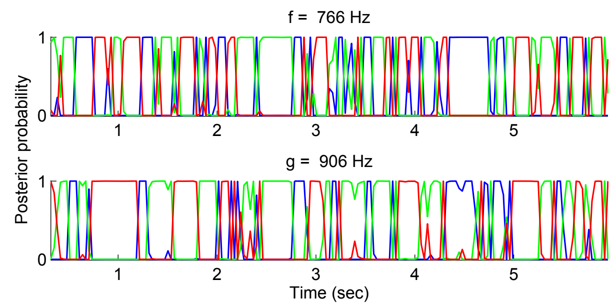

Figure 4: *Posterior probability sequences for three classes*

The posterior probability sequences resulted in higher correlation coefficients for the same source. This is because the posterior probabilities represent active signals with values uniformly close to 1, whereas envelopes represent active signals with various values. In this sense, a posterior probability has similar characteristics to a dominance measure developed in [10] for determined BSS.

To determine permutations Π_f, we cluster such activity sequences $v_i^f(t)$ for each source. For computational efficiency, we first employ a global rough optimization that maximizes

$$\mathcal{J}(\{c_k\}, \{\Pi_f\}) = \sum_{f \in \mathcal{F}} \sum_{k=1}^N \rho(v_i^f, c_k) \,\big|_{i=\Pi_f(k)}, \tag{11}$$

where $\mathcal{F}$ represents the set of all frequency bins. The centroid c_k is calculated for each source as the average of activity sequences with the current permutation Π_f:

$$c_k(t) \leftarrow \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} v_i^f(t) \,\big|_{i=\Pi_f(k)}, \quad \forall \, k, t \, , \tag{12}$$

where $|\mathcal{F}|$ is the number of elements in the set $\mathcal{F}$. The permutations Π_f are optimized to maximize the correlation coefficients ρ between activity sequences and the current centroid:

$$\Pi_f \leftarrow \text{argmax}_\Pi \sum_{k=1}^N \rho(v_i^f, c_k) \,\big|_{i=\Pi(k)}, \quad \forall \, f \in \mathcal{F} \, . \tag{13}$$

These two operations (12) and (13) are iterated until convergence.

Then we employ a local fine optimization to improve the separation further. It maximizes the sum of the correlation coefficients over a set of selected frequencies $\mathcal{R}(f)$ for a frequency f:

$$\Pi_f \leftarrow \text{argmax}_\Pi \sum_{g \in \mathcal{R}(f)} \sum_{k=1}^N \rho(v_i^f, v_{i'}^g) \,\big|_{i=\Pi(k), i'=\Pi_g(k)} \, . \tag{14}$$

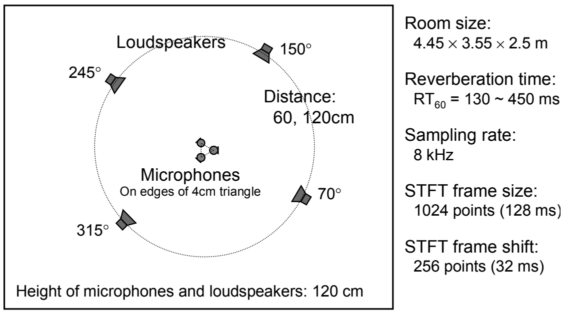

Figure 5: *Experimental setup*

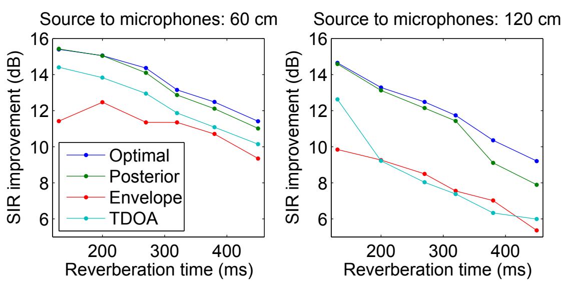

Figure 6: *Experimental results*

The set $\mathcal{R}(f)$ preferably consists of frequencies g where a high correlation coefficient $\rho(v_i^f, v_{i'}^g)$ would be attained for v_i^f and $v_{i'}^g$ corresponding to the same source. We typically select adjacent frequencies and harmonic frequencies [9]. The fine local optimization (14) is performed for one selected frequency f at a time, and repeated until no improvement is found for any frequency f.

5. EXPERIMENTAL RESULTS

We performed experiments to separate four speech sources with three microphones. We measured impulse responses $\mathrm{h}_{jk}(l)$ under the conditions shown in Fig. 5. Mixtures at the microphones were made by convolving the impulse responses and 6-second English speeches. The computation time was around 8 seconds for 6-second speech mixtures. The program was coded in Matlab and run on an AMD 2.4 GHz Athlon 64 processor. The separation performance was evaluated in terms of signal-to-interference ratio (SIR) improvement. The improvement was calculated by $\mathsf{OutputSIR}_i - \mathsf{InputSIR}_i$ for each output i, and we took the average over all outputs. These two types of SIRs are defined as

$$\mathsf{InputSIR}_i = 10\log_{10}\frac{\sum_t |\sum_l \mathrm{h}_{Ji}(l)\mathrm{s}_i(t-l)|^2}{\sum_t |\sum_{k\neq i}\sum_l \mathrm{h}_{Jk}(l)\mathrm{s}_k(t-l)|^2} \quad \text{(dB)},$$

$$\mathsf{OutputSIR}_i = 10\log_{10}\frac{\sum_t |\mathrm{y}_{ii}(t)|^2}{\sum_t |\sum_{k\neq i}\mathrm{y}_{ik}(t)|^2} \quad \text{(dB)},$$

where $J \in \{1,\ldots,M\}$ is the index of one selected reference sensor, and $\mathrm{y}_{ik}(t)$ is the component of s_k that appears at output $\mathrm{y}_i(t)$, i.e. $\mathrm{y}_i(t) = \sum_{k=1}^N \mathrm{y}_{ik}(t)$.

Figure 6 shows the average SIR improvements for 8 combinations of 4 speeches. We had two source-to-microphones distances (60 and 120 cm) and six reverberation times (130, 200, 270, 320, 380 and 450 ms). We compared four strategies for permutation alignment. "TDOA" corresponds to the existing techniques mentioned in the Introduction. "Envelope" and "Posterior" use signal envelopes $|y_i|$ and posterior probability sequences $P(C_i|\mathbf{x})$, respectively, for the permutation alignment procedure described in Sec. 4. "Optimal" corresponds to optimal permutations calculated with knowledge of the original source signals. This is not a practical solution but used to determine the upper bound of the results.

We observed the following tendencies. "TDOA" performed moderately well in short distance (60 cm) or low reverberant (130 ms) cases, but did not perform well under reverberant conditions with a distance of 120 cm. "Envelope" did not perform so well in many cases. The proposed "Posterior" performed the best among the practical solutions (except "Optimal") in all cases.

We also made recordings in a room using a portable audio recorder with two microphones, and separated the mixtures of three speeches. Sound examples can be found on our web site [11].

6. CONCLUSIONS

This paper presented a method for underdetermined convolutive BSS. The method has a two-stage hierarchical structure that classifies observation vectors $\mathbf{x}(f,t)$ into N sources. Frequency bin-wise samples along the time axis are classified in the first stage, and then the posterior probability sequences are clustered along the frequency axis in the second stage. The experimental results show the effectiveness and robustness of the proposed method even under reverberant conditions.

7. REFERENCES

[1] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. John Wiley & Sons, 2000.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

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

[4] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[6] P. D. O'Grady and B. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. ICA 2004 (LNCS 3195)*. Springer, Sept. 2004, pp. 430–436.

[7] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and L1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, 2007, Article ID 24717.

[8] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, Oct. 2001.

[9] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[10] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. ISCAS 2007*, 2007, pp. 3247–3250.

[11] [Online]. http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/ubssconv/