Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment

Hiroshi Sawada, *Senior Member, IEEE*, Shoko Araki, *Member, IEEE*, and Shoji Makino, *Fellow, IEEE*

***Abstract*—This paper presents a blind source separation method for convolutive mixtures of speech/audio sources. The method can even be applied to an underdetermined case where there are fewer microphones than sources. The separation operation is performed in the frequency domain and consists of two stages. In the first stage, frequency-domain mixture samples are clustered into each source by an expectation–maximization (EM) algorithm. Since the clustering is performed in a frequency bin-wise manner, the permutation ambiguities of the bin-wise clustered samples should be aligned. This is solved in the second stage by using the probability on how likely each sample belongs to the assigned class. This two-stage structure makes it possible to attain a good separation even under reverberant conditions. Experimental results for separating four speech signals with three microphones under reverberant conditions show the superiority of the new method over existing methods. We also report separation results for a benchmark data set and live recordings of speech mixtures.**

***Index Terms*—Blind source separation (BSS), convolutive mixture, expectation–maximization (EM) algorithm, permutation problem, short-time Fourier transform (STFT), sparseness, time–frequency (T–F) masking.**

I. Introduction

THE technique for estimating individual source components from their mixtures at multiple sensors is known as blind source separation (BSS) [1]–[5]. With acoustic applications of BSS, such as solving a cocktail party problem, signals are mixed in a **convolutive** manner with reverberation. Since a typical room reverberation time is about 300 ms, we need thousands of coefficients estimated for the separation filters even with an 8-kHz sampling rate. This makes the convolutive BSS problem much more difficult than the BSS of simple **instantaneous** mixtures. Various attempts have been made to solve the convolutive BSS problem. Among them, frequency-domain approaches [6]–[13] are popular ones where time-domain observation signals are converted into frequency-domain time-series signals by a short-time Fourier transform (STFT).

Another difficulty stems from the fact that there may be more source signals of interest than sensors (or microphones in acoustic applications). If we have a sufficient number of microphones, i.e., a **determined** case, linear filters that are estimated for example by independent component analysis (ICA) [1]–[4] effectively separate the mixtures. However, if the number of microphones is insufficient, i.e., an **underdetermined** case, such linear filters do not work well. Instead, time–frequency (T–F) masking [14]–[23] or a maximum *a posteriori* (MAP) estimator [24]–[27] is widely used to separate such underdetermined mixtures. For underdetermined cases, frequency-domain approaches are also popular. This is because most interesting acoustic sources, such as speech and music, exhibit a sparseness property in the time–frequency representation, and this sparseness property helps the design of T–F masking or MAP estimation.

Therefore, **underdetermined convolutive** BSS has been recognized as a challenging task, and a lot of research effort has been devoted to it [14]–[25]. The majority of the existing techniques [14]–[21] rely on time-difference-of-arrival (TDOA) estimations for each source at multiple microphones, or interaural time difference (ITD) estimations for a two-microphone stereo case and a human/animal auditory system. A nice simplicity of these techniques is that clustering frequency components for each source is conducted in a full-band manner as shown in Fig. 3(a). Such techniques work effectively under low reverberant conditions, where the assumed anechoic model is satisfied to a certain degree. However, under severe reverberant conditions, TDOA estimations become unreliable and such techniques do not work well.

The main goal of this paper is to develop an underdetermined convolutive BSS method that realizes good separation performance even under reverberant conditions. The method employs a widely used T–F masking scheme to separate the mixtures. We adopt a two-stage approach where the first stage is responsible for frequency bin-wise clustering as shown in Fig. 3(b). Since the clustering is conducted in a frequency bin-wise manner rather than a full-band manner, it is robust as regards room reverberations as long as the frame length of the STFT analysis window is long enough to cover the main part of the impulse responses. Moreover, the method is immune to the spatial aliasing problem [28], [29] encountered when

Manuscript received November 23, 2009; revised March 11, 2010; accepted May 10, 2010. Date of publication May 27, 2010; date of current version December 03, 2010. Earlier versions of this work were presented at the 2007 IEEE International Symposium on Circuits and Systems (ISCAS 2007) and the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007) as symposium/workshop papers. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dan Ellis.

H. Sawada and S. Araki are with NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan (e-mail: sawada@cslab.kecl.ntt.co.jp; shoko@cslab.kecl.ntt.co.jp).

S. Makino is with Tsukuba University, Ibaraki 305-8577, Japan (e-mail: maki@tara.tsukuba.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2051355

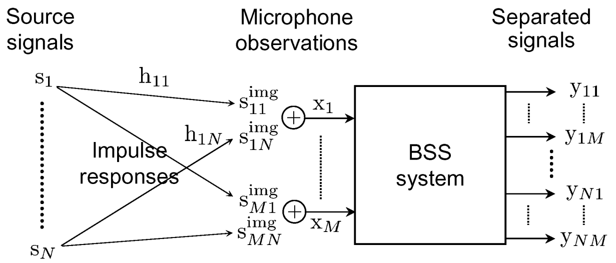

Fig. 1. Signal notations.

TDOAs/ITDs are estimated with widely spaced microphones (e.g., spatial aliasing occurs for frequencies $f > 850$ Hz with 20-cm spacing microphones).

With such a two-stage approach, an additional task is performed in the second stage to group together bin-wise separated frequency components coming from the same source. This task is almost identical to the permutation problem of frequency-domain ICA-based BSS [6]–[10], [13]. A few methods [24], [25] that employ such a two-stage structure for underdetermined convolutive BSS have already been proposed. With these methods, permutation alignment is performed by maximizing the correlation coefficients of **amplitude envelopes**, which basically represent sound source activity, of the same source. As also presented in this paper, the correlation coefficient of the amplitude envelopes is not always a good criterion with which to judge whether two sets of separated frequency components come from the same source or not.

In the proposed method, the bin-wise clustering results of the first stage are represented by a set of **posterior probabilities** $P(C_i|\mathbf{x}(\tau,f))$, the probability that the observation vector $\mathbf{x}$ at time τ and frequency f belongs to the $\imath$th class. The permutation alignment procedure in the second stage utilizes these posterior probabilities instead of traditionally used amplitude envelopes. Posterior probabilities also represent sound source activity. We observed that the time sequences of posterior probabilities exhibited a much clearer contrast between a same-source pair and a different-source pair when we calculated their the correlation coefficients, as long as different sources were not synchronized. As a result, the permutation alignment capability has been considerably improved compared to previous methods using amplitude envelopes.

This paper is organized as follows. Section II provides a system overview of the proposed method. Sections III and IV present detailed explanations of the first and second stages of the proposed method, respectively. Section V reports experimental results. Section VI concludes this paper.

II. System Overview

This section provides a system overview of the proposed BSS method. Fig. 1 shows our signal notations for the convolutive BSS problem. Fig. 2 shows a processing flow for T–F masking based BSS. Fig. 3 details the **Clustering** part by comparing widely used methods and our proposed method. The example spectrograms in Fig. 4 help us to understand intuitively how signals are processed.

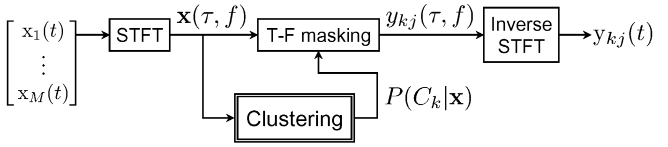

Fig. 2. Generic processing flow for BSS with time–frequency (T–F) masking.

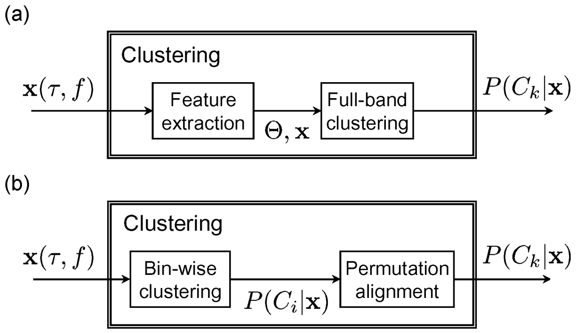

Fig. 3. Comparison of the Clustering part shown in Fig. 2 for widely used methods and the proposed method. (a) Widely used methods based on an anechoic model. (b) The method proposed in this paper.

A. Signal Notations

As shown in Fig. 1, let $\mathrm{s}_1,\ldots,\mathrm{s}_N$ be source signals and $\mathrm{x}_1,\ldots,\mathrm{x}_M$ be microphone observations. The numbers of sources and microphones are denoted by N and M, respectively. A case where $N > M$ is called an underdetermined BSS (our focus here), and alternatively a case where $N \leq M$ is called a determined BSS. The observation x_j at microphone j is described by a mixture

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \mathrm{s}_{jk}^{\mathrm{img}}(t) \tag{1}$$

of source s_k images at the microphone j

$$\mathrm{s}_{jk}^{\mathrm{img}}(t) = \sum_{l} \mathrm{h}_{jk}(l)\mathrm{s}_k(t-l) \tag{2}$$

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

Our goal for the BSS task is to obtain sets of separated signals $\{\mathrm{y}_{11},\ldots,\mathrm{y}_{1M}\},\ldots,\{\mathrm{y}_{N1},\ldots,\mathrm{y}_{NM}\}$, where each set corresponds to each of the source signals $\mathrm{s}_1,\ldots,\mathrm{s}_N$. More specifically, y_{kj} is an estimated source k image $\mathrm{s}_{jk}^{\mathrm{img}}$ at the jth microphone. The task should be performed only with M observed mixtures $\mathrm{x}_1,\ldots,\mathrm{x}_M$, and without information on the sources s_k, the impulse responses h_{jk}, and the source images $\mathrm{s}_{jk}^{\mathrm{img}}$.

B. Short-Time Fourier Transform (STFT)

The rest of this section explains the processing parts shown in Fig. 2, starting with STFT. The microphone observations (1) sampled at a sampling frequency f_s, or with a sampling period $t_s = 1/f_s$, are converted into frequency-domain time-series signals $x_j(\tau,f)$ by an STFT with an L-sample frame and its S-sample shift

$$x_j(\tau,f) \leftarrow \sum_{t'=0,t_s,\cdots,(L-1)t_s} \mathrm{win}_a(t')\mathrm{x}_j(t'+\tau)e^{-\imath 2\pi f t'} \tag{3}$$

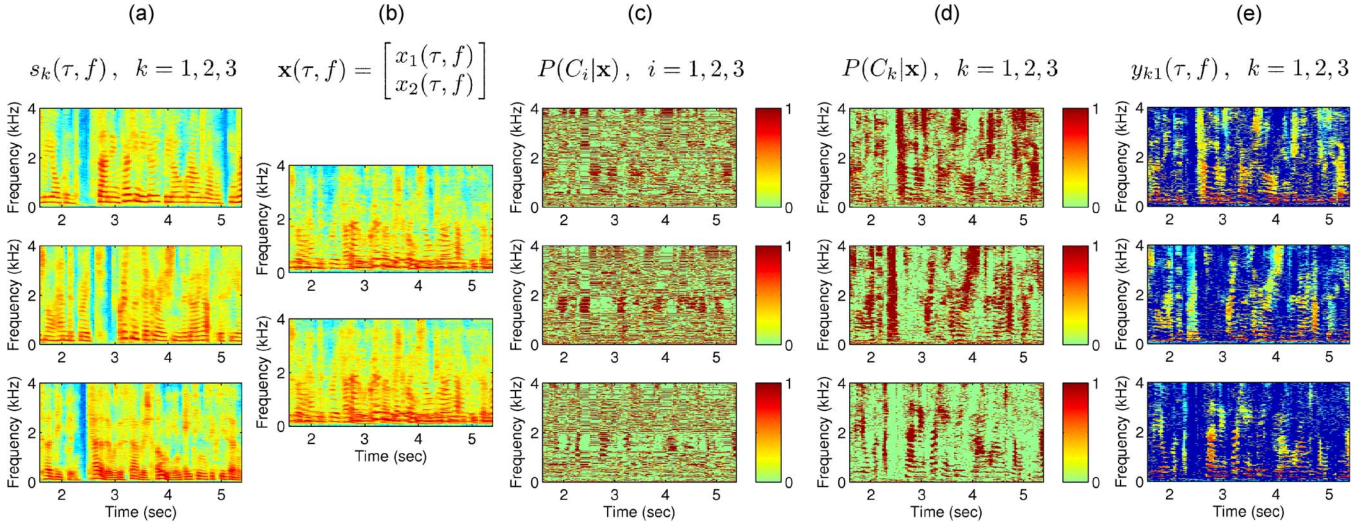

Fig. 4. Spectrogram examples: a case with three speech sources and two microphones. (a) Sources. (b) Mixtures. (c) Bin-wise classification. (d) Permutation aligned classification. (e) Separated signals.

for frame time indices $\tau = 0, St_s, \ldots, T-1$ and frequencies $f = 0, (1/L)f_s, \ldots, ((L-1)/L)f_s$. Note that τ represents the starting time of the corresponding frame. We typically use an analysis window $\mathrm{win}_a(t)$ that tapers smoothly to zero at each end, such as a Hanning window $\mathrm{win}_a(t) = (1/2)(1 - \cos(2\pi t/Lt_s))$.

If the frame size L is long enough to cover the main part[1] of the impulse responses h_{jk}, the convolutive mixture model (1) and (2) can be approximated as an instantaneous mixture model [6], [9] at each frequency

$$x_j(\tau, f) = \sum_{k=1}^{N} h_{jk}(f) s_k(\tau, f) + n_j(\tau, f) \tag{4}$$

where $h_{jk}(f)$ is the frequency response from source k to microphone j, $s_k(\tau, f)$ is a frequency-domain time-series signal of $\mathrm{s}_k(t)$ obtained by an STFT similar to (3), and $n_j(\tau, f)$ is a noise term that consists of additive background noise and reverberant components outside the analysis window. We also use a vector notation

$$\mathbf{x}(\tau, f) = \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(\tau, f) + \mathbf{n}(\tau, f) \tag{5}$$

where $\mathbf{h}_k = [h_{1k}, \ldots, h_{Mk}]^T$, $\mathbf{n} = [n_1, \ldots, n_M]^T$, and $\mathbf{x} = [x_1, \ldots, x_M]^T$.

C. Time-Frequency (T–F) Masking

Separated signals $\{y_{11}, \ldots, y_{1M}\}, \ldots, \{y_{N1}, \ldots, y_{NM}\}$ in the frequency domain are constructed by time-frequency (T–F) masking

$$y_{kj}(\tau, f) = \mathcal{M}_k(\tau, f) x_j(\tau, f) \tag{6}$$

where $0 \leq \mathcal{M}_k(\tau, f) \leq 1$ is a mask specified for each separated signal y_k and each time-frequency slot (τ, f).

For the design of masks $\mathcal{M}_k(\tau, f)$, we rely on the sparseness property of source signals [17]. A sparse source can be characterized by the fact that the source amplitude is close to zero most of the time. A time-frequency-domain speech source is a good example of a sparse source. Based on this property, it is likely that at most only one source signal has a large contribution to each time-frequency observation $\mathbf{x}(\tau, f)$. Thus, the mixture model (5) can be further approximated as

$$\mathbf{x}(\tau, f) = \mathbf{h}_{k^\star}(f) s_{k^\star}(\tau, f) + \tilde{\mathbf{n}}(\tau, f), \quad k^\star \in \{1, \ldots, N\} \tag{7}$$

for sparse sources. The subscript $k^\star = k^\star(\tau, f)$ depends on each time-frequency slot (τ, f), and represents the index of the most dominant source for the corresponding T–F slot. The noise term now becomes $\tilde{\mathbf{n}} = \mathbf{n} + \sum_{k' \neq k^\star} \mathbf{h}_{k'} s_{k'}$. The index $k^\star$ should be identified or estimated for each (τ, f) to separate the sources by T–F masking.

For that purpose, observation vectors $\mathbf{x}(\tau, f)$ for all time-frequency slots (τ, f) are clustered into N classes $C_1, \ldots, C_N$, each of which corresponds to a source signal s_k. A vector $\mathbf{x}(\tau, f)$ should belong to class C_k if the source s_k is the most dominant in the observation $\mathbf{x}(\tau, f)$. We perform the clustering in a soft sense. A posterior probability $P(C_k|\mathbf{x})$, which represents how likely the vector $\mathbf{x}$ belongs to the kth class, is calculated in the "Clustering" part shown in Fig. 2. Then, the T–F masks that are required in (6) are specified by

$$\mathcal{M}_k(\tau, f) = \begin{cases} 1, & \text{if } P(C_k|\mathbf{x}) \geq P(C_{k'}|\mathbf{x}),\ \forall k' \neq k \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

In other words, the kth mask $\mathcal{M}_k$ at a time-frequency slot (τ, f) is specified as 1 if and only if the kth source is estimated as the most dominant source in the observation $\mathbf{x}$ at the T–F slot.

D. Inverse STFT

At the end of the processing flow, time-domain separated signals $\mathrm{y}_{kj}(t)$, $k = 1, \ldots, N$, $j = 1, \ldots, M$ are calculated with

[1] The definition of the main part of the impulse responses is not rigorous, and in general the frame size L is determined empirically. An experimental analysis of the relationship between frame sizes and separation performance is presented in [30].

an inverse STFT applied to the separated frequency components $y_{kj}(\tau, f)$

$$\mathrm{y}_{kj}(t) \leftarrow \sum_{\tau} \mathrm{win}_s(t-\tau) \left[\frac{1}{L} \sum_{f} y_{kj}(\tau, f) e^{\imath 2\pi f (t-\tau)} \right] \quad (9)$$

where the summation over frequencies f is with $f = 0, (1/L)f_s, \cdots, ((L-1)/L)f_s$, and the summation over frame time indices τ is with those that satisfy $0 \leq t - \tau \leq (L-1)t_s$. We use a synthesis window win_s that is defined as nonzero only in the L-sample interval $[0, (L-1)t_s]$ and tapers smoothly to zero at each end to mitigate the edge effect. To realize a perfect reconstruction, the analysis and synthesis windows should satisfy the condition

$$\sum_{\tau} \mathrm{win_s}(t-\tau)\mathrm{win_a}(t-\tau) = 1.$$

Again, the summation over frame time indices τ is with those that satisfy $0 \leq t - \tau \leq (L-1)t_s$.

E. Comparison With Widely Used Methods

This subsection compares the proposed method with widely used methods [14]–[21] by focusing on the Clustering procedure shown in Fig. 2 and detailed in Fig. 3.

With the widely used methods, a set Θ of features is extracted from an observation vector $\mathbf{x}$ for each T–F slot (τ, f). A typical feature is the time-difference-of-arrival (TDOA) that occurs at microphone pairs. Based on an anechoic assumption, the features of all times τ and all frequencies f (full-band) are expected to form several clusters, each of which corresponds to a source signal located at a specific position. Although such methods perform well under low reverberant conditions, the separation performance degrades as the reverberation becomes heavy. This is because the anechoic assumption imposes a linear phase constraint on the vector $\mathbf{h}_k(f)$ in the mixture model (7), and the constraint contradicts the observations affected by reverberations. Some improvement for highly reverberant conditions could be gained by modeling TDOA variations with a mixture of Gaussians [18] or gradually making the parameters frequency dependent [19].

The Clustering procedure of the method proposed in this paper has a two-stage structure. The first stage performs frequency bin-wise clustering, and the second stage performs permutation alignment. Example spectrograms corresponding to these two stages are shown in Fig. 4(c) and (d). The purpose of the two-stage structure is to tackle the reverberation problem mentioned above. The proposed method has no assumption as regards the vector $\mathbf{h}_k(f)$ in (7). It can be adapted to various impulse responses $\mathrm{h}_{jk}(l)$ caused typically by reverberations, as long as the STFT analysis window $\mathrm{win}_a(t)$ covers the main part of the impulse responses.

The next two sections explain how to calculate in the proposed method the posterior probability $P(C_k|\mathbf{x})$ that the kth source is the most dominant source in the observation $\mathbf{x}$. The procedure consists of two stages, "Bin-wise clustering" and "Permutation alignment."

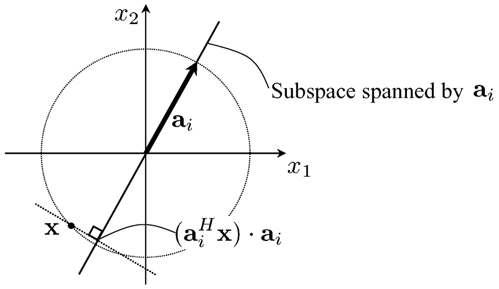
x_2
Subspace spanned by $\mathbf{a}_i$
$\mathbf{a}_i$
x_1
$\mathbf{x}$
$(\mathbf{a}_i^H\mathbf{x}) \cdot \mathbf{a}_i$

Fig. 5. Illustration of the line orientation idea. Two-dimensional real vector space is presented for simplicity.

III. Bin-Wise Clustering

This section describes the first stage "Bin-wise clustering" in detail.

A. Model

Since the operation is performed in a frequency bin-wise manner, let us omit the frequency dependence in (5) and (7) for simplicity in this section

$$\mathbf{x}(\tau) = \sum_{i=1}^{N} \mathbf{h}_i s_i(\tau) + \mathbf{n}(\tau) = \mathbf{h}_{i^\star} s_{i^\star}(\tau) + \tilde{\mathbf{n}}(\tau). \quad (10)$$

The subscript $i^\star = i^\star(\tau)$ is the index of the most dominant source for each time τ. We changed the use of the source subscript from k to i, intending to clarify that there are permutation ambiguities in the frequency bin-wise clustering. Such permutation ambiguities will be aligned in the second stage, which is detailed in the next section.

We see in (10) that clustering can be performed according to the information on the vectors $\mathbf{h}_1, \ldots, \mathbf{h}_N$. To eliminate the effect of source amplitude $s_{i^\star}(\tau)$ from $\mathbf{x}$, we normalize them so that they have a unit norm

$$\mathbf{x}(\tau) \leftarrow \frac{\mathbf{x}(\tau)}{||\mathbf{x}(\tau)||} = \frac{\mathbf{h}_{i^\star}}{||\mathbf{h}_{i^\star}||} \cdot \frac{s_{i^\star}(\tau)}{|s_{i^\star}(\tau)|}. \quad (11)$$

An unknown phase $s_{i^\star}(\tau)/|s_{i^\star}(\tau)|$ ambiguity still remains in $\mathbf{x}(\tau)$. To model such a vector for each source, we follow the line orientation idea in [26], [27] and employ a complex Gaussian density function of the form

$$p(\mathbf{x}|\mathbf{a}_i, \sigma_i) = \frac{1}{(\pi\sigma_i^2)^{M-1}} \exp\left(-\frac{\left\|\mathbf{x} - \left(\mathbf{a}_i^H \mathbf{x}\right) \cdot \mathbf{a}_i\right\|^2}{\sigma_i^2} \right) \quad (12)$$

where $\mathbf{a}_i$ is the centroid with unit norm $||\mathbf{a}_i||^2 = 1$, 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 $\mathbf{a}_i$, the distance $||\mathbf{x} - (\mathbf{a}_i^H\mathbf{x}) \cdot \mathbf{a}_i||$ represents the minimum distance between the point $\mathbf{x}$ and the subspace, which implies how probable $\mathbf{x}$ belongs to the ith class (Fig. 5).

Since the observation vector $\mathbf{x}$ is modeled as (10), the density function $p(\mathbf{x})$ can be described by a mixture model

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{N} \alpha_i p(\mathbf{x}|\mathbf{a}_i, \sigma_i) \tag{13}$$

with a parameter set

$$\theta = \{\mathbf{a}_1, \sigma_1, \alpha_1, \ldots, \mathbf{a}_N, \sigma_N, \alpha_N\}. \tag{14}$$

The mixture ratios α_i should satisfy $\alpha_1 + \cdots + \alpha_N = 1$ and $0 \leq \alpha_i \leq 1$, and are modeled by a Dirichlet distribution as

$$p(\alpha_1, \ldots, \alpha_N) = \frac{\Gamma(N \cdot \phi)}{\Gamma(\phi)^N} \prod_{i=1}^{N} \alpha_i^{(\phi-1)} \tag{15}$$

where ϕ is a hyper-parameter.

B. EM Algorithm

We employ the EM algorithm [31], [32] to estimate the parameters in the set θ and posterior probabilities $P(C_i|\mathbf{x}(\tau))$ for all times τ and $i = 1, \ldots, N$. The EM algorithm iterates the E-step and the M-step until convergence.

In the E-step, posterior probabilities are calculated by

$$P(C_i|\mathbf{x}, \theta') = \frac{\alpha'_i p\left(\mathbf{x}|\mathbf{a}'_i, \sigma'_i\right)}{p(\mathbf{x}|\theta')} = \frac{\alpha'_i p\left(\mathbf{x}|\mathbf{a}'_i, \sigma'_i\right)}{\sum_{i=1}^{N} \alpha'_i p\left(\mathbf{x}|\mathbf{a}'_i, \sigma'_i\right)} \tag{16}$$

with the current parameter set

$$\theta' = \left\{\mathbf{a}'_1, \sigma'_1, \alpha'_1, \ldots, \mathbf{a}'_N, \sigma'_N, \alpha'_N\right\}.$$

In the M-step, the parameter set θ is updated by maximizing

$$Q(\theta, \theta') + \log p(\theta) \tag{17}$$

where $Q(\theta, \theta')$ is an auxiliary function defined by

$$Q(\theta, \theta') = \sum_{\tau}^{T} \sum_{i=1}^{N} P\left(C_i|\mathbf{x}(\tau), \theta'\right) \log \alpha_i p\left(\mathbf{x}(\tau)|\mathbf{a}_i, \sigma_i\right)$$

and $p(\theta)$ is a prior distribution for the parameters. We consider the prior (15) for the mixture ratios α_i but no prior for the Gaussian parameters $\mathbf{a}_i$ and σ_i. Thus, we have

$$\log p(\theta) = (\phi - 1) \sum_{i=1}^{N} \log \alpha_i + \text{const.}$$

As described in detail in the Appendix, each parameter is updated as follows. The new centroid $\mathbf{a}_i$ is given by the eigenvector corresponding to the maximum eigenvalue of

$$\mathbf{R} = \sum_{\tau}^{T} P\left(C_i|\mathbf{x}(\tau), \theta'\right) \cdot \mathbf{x}(\tau)\mathbf{x}^H(\tau). \tag{18}$$

The variance σ_i^2 and the mixture ratio α_i are updated by

$$\sigma_i^2 = \frac{\sum_{\tau}^{T} P\left(C_i|\mathbf{x}(\tau), \theta'\right) \cdot \left\|\mathbf{x}(\tau) - \left(\mathbf{a}_i^H \mathbf{x}(\tau)\right) \cdot \mathbf{a}_i\right\|^2}{(M-1) \cdot \sum_{\tau}^{T} P\left(C_i|\mathbf{x}(\tau), \theta'\right)} \tag{19}$$

and

$$\alpha_i = \frac{\sum_{\tau}^{T} P\left(C_i|\mathbf{x}(\tau), \theta'\right) + \phi - 1}{T + N \cdot (\phi - 1)} \tag{20}$$

respectively.

After convergence, the clustering results are represented by the posterior probabilities $P(C_i|\mathbf{x}, \theta)$ shown in (16).

C. Practical Issues

Pre-whitening [3] the observation vectors $\mathbf{x}(\tau)$ is effective for a robust execution of the clustering procedure, and can be simply performed by

$$\mathbf{x}(\tau) \leftarrow \mathbf{V}\mathbf{x}(\tau)$$

where the whitening matrix $\mathbf{V}$ is calculated by $\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^H$ with an eigenvalue decomposition $\mathrm{E}\{\mathbf{x}\mathbf{x}^H\} = \mathbf{E}\mathbf{D}\mathbf{E}^H$ of the correlation matrix. The unit-norm procedure (11) must be employed again after the pre-whitening process.

In the experiments shown in Section V, we assumed that the information on the number N of sources was given *a priori*. For such a case, it is advantageous to choose a large number for the hyper-parameter ϕ in (15) so that each cluster has almost the same weight α_i based on (20).

We confirmed empirically that the EM algorithm presented in the previous subsection generally exhibits satisfactory convergence behaviors as long as the initial parameters are set appropriately, for instance as follows. We choose the initial centroids from the samples in such a way that we specify N time points $\tau_1, \ldots, \tau_N$ beforehand and then set them by $\mathbf{a}_i \leftarrow \mathbf{x}(\tau_i)$ for $i = 1, \ldots, N$. The other parameters are initially set as $\sigma_i^2 = 0.1$ and $\alpha_i = 1/N$.

IV. Permutation Alignment

This section describes the second stage "Permutation Alignment" in detail.

A. Purpose

After the first stage, we have posterior probabilities $P(C_i|\mathbf{x}(\tau, f))$ according to (16) for $i = 1, \ldots, N$ and all time–frequency slots (τ, f). However, since the class order $C_1, \ldots, C_N$ may be different from one frequency to another [Fig. 4(c)], we need to reorder the indices so that the same index corresponds to the same source over all frequencies [Fig. 4(d)]. In other words, we need to determine a permutation

$$\Pi_f : \{1, \ldots, N\} \rightarrow \{1, \ldots, N\}$$

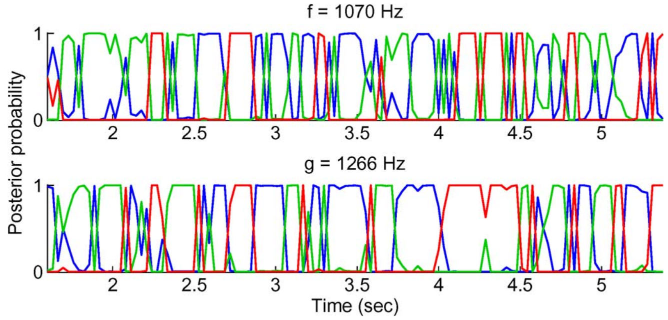

Fig. 6. Posterior probability sequences v_1^f, v_2^f, v_3^f at frequency $f = 1070$ Hz and v_1^g, v_2^g, v_3^g at frequency $g = 1266$ Hz. Permutations are aligned and the sequences originating from the same sound source are shown in the same color for ease of interpretation.

for all frequencies f, and then update the posterior probabilities by

$$P(C_k|\mathbf{x}) \leftarrow P(C_i|\mathbf{x})|_{i=\Pi_f(k)}, \quad k = 1, \ldots, N \tag{21}$$

to construct proper separated signals. Such a permutation problem has been extensively studied for frequency-domain ICA-based BSS applied to a determined case, e.g., [6]–[10], [13].

B. Posterior Probability Sequence

In this paper, we propose utilizing the sequence of posterior probabilities $P(C_k|\mathbf{x})$ along the time axis at a frequency. Let us define a posterior probability sequence[2]

$$v_i^f(\tau) = P\left(C_i|\mathbf{x}(\tau, f)\right) \tag{22}$$

for the ith class (separated components) at frequency f. As Fig. 6 shows intuitively, posterior probability sequences that belong to the same source generally have similar patterns among different frequencies. This is because a sound source has a specific activity pattern along the time axis, and more specifically, it has common silence periods, onsets and offsets. Inversely with different sound sources, posterior probability sequences have dissimilar patterns.

Such similarity and dissimilarity can be calculated by a correlation coefficient defined for two sequences v_i and v_j

$$\rho(v_i, v_j) = \frac{\mathrm{E}\left\{(v_i - \mu_i)(v_j - \mu_j)\right\}}{\sigma_i \sigma_j}$$

where $\mu_i = \mathrm{E}\{v_i\}$ is the mean and $\sigma_i = \sqrt{\mathrm{E}\{v_i^2\} - \mu_i^2}$ is the standard deviation of v_i.[3] The correlation coefficient of any two sequences is bounded by $-1 \leq \rho(v_i, v_j) \leq 1$, and becomes 1 if the two sequences are identical up to a positive scaling and an additive offset.

Let us calculate the correlation coefficients $\rho(v_i^f, v_j^g)$ for the posterior probability sequences shown in Fig. 6, i.e., v_i^f and v_j^g for output indices $i, j = 1, 2, 3$, and frequencies $f = 1070$ and $g = 1266$

$$\begin{bmatrix} \rho\left(v_1^f, v_1^g\right) & \rho\left(v_1^f, v_2^g\right) & \rho\left(v_1^f, v_3^g\right) \\ \rho\left(v_2^f, v_1^g\right) & \rho\left(v_2^f, v_2^g\right) & \rho\left(v_2^f, v_3^g\right) \\ \rho\left(v_3^f, v_1^g\right) & \rho\left(v_3^f, v_2^g\right) & \rho\left(v_3^f, v_3^g\right) \end{bmatrix} = \begin{bmatrix} 0.37 & -0.02 & -0.33 \\ -0.06 & 0.39 & -0.29 \\ -0.30 & -0.32 & 0.57 \end{bmatrix}. \tag{23}$$

We observe that $\rho(v_i^f, v_j^g)$ is positive for two sequences originating from the same sound source, and inversely $\rho(v_i^f, v_j^g)$ is negative for those originating from different two sources. Therefore, permutation alignment should be conducted so that $\rho(v_i^f, v_j^g)$ is positive for $i = j$ and is negative or close to zero for $i \neq j$.

C. Score Value Optimized by Permutation

To describe our permutation alignment procedure in a more formal manner, we introduce certain notations. Let $\{v_i^f\} = [v_1^f, \ldots, v_N^f]$ be an ordered list of sequences v_i^f, and let $\{v_i^f\}|_{\Pi_f} = [v_{\Pi_f(1)}^f, \ldots, v_{\Pi_f(N)}^f]$ be a permuted list of sequences with a permutation Π_f. Also, let $\mathbf{Q}(\{v_i^f\}, \{v_j^g\})$ be an $N \times N$ matrix whose (i, j)-element is $\rho(v_i^f, v_j^g)$. For example if $N = 3$

$$\mathbf{Q}\left(\left\{v_i^f\right\}, \left\{v_j^g\right\}\right) = \begin{bmatrix} \rho\left(v_1^f, v_1^g\right) & \rho\left(v_1^f, v_2^g\right) & \rho\left(v_1^f, v_3^g\right) \\ \rho\left(v_2^f, v_1^g\right) & \rho\left(v_2^f, v_2^g\right) & \rho\left(v_2^f, v_3^g\right) \\ \rho\left(v_3^f, v_1^g\right) & \rho\left(v_3^f, v_2^g\right) & \rho\left(v_3^f, v_3^g\right) \end{bmatrix} \tag{24}$$

like (23). Then, let us define a scalar

$$\mathit{score}[\mathbf{Q}] = \mathrm{sum}\left(\mathrm{diag}(\mathbf{Q})\right) - \mathrm{sum}\left(\mathrm{offdiag}(\mathbf{Q})\right) \tag{25}$$

where diag() and offdiag() take the diagonal and off-diagonal elements of a matrix, respectively, and sum() calculates the sum of the elements. For (23), the score value is 2.66.

A primitive operation in the permutation alignment procedure is to maximize the $\mathit{score}[\mathbf{Q}]$ value by a permutation Π_f. For example, if

$$\mathbf{Q}\left(\left\{v_i^f\right\}, \left\{v_j^g\right\}\right) = \begin{bmatrix} -0.06 & 0.39 & -0.29 \\ 0.37 & -0.02 & -0.33 \\ -0.30 & -0.32 & 0.57 \end{bmatrix}$$

is given, we employ a permutation $\Pi_f : [1, 2, 3] \rightarrow [2, 1, 3]$ that converts the ordered list $\{v_i^f\}$ into a permuted list $\{v_i^f\}|_{\Pi_f}$ to obtain the maximum score value with

$$\mathbf{Q}\left(\left\{v_i^f\right\}\Big|_{\Pi_f}, \left\{v_j^g\right\}\right) = \begin{bmatrix} 0.37 & -0.02 & -0.33 \\ -0.06 & 0.39 & -0.29 \\ -0.30 & -0.32 & 0.57 \end{bmatrix}.$$

[2] A similar sequence defined for ICA-based determined BSS is presented by (15) in our previous work [13].

[3] Here, σ_i is used differently from that used in Section III.

D. Permutation Optimization

This subsection describes the procedure for permutation optimization. The permutations Π_f in (21) of all frequency bins f should be optimized so that

$$\sum_{f,g\in\mathcal{F}} score\left[\mathbf{Q}\left(\left\{v_i^f\right\}\Big|_{\Pi_f}, \{v_j^g\}|_{\Pi_g}\right)\right]$$

is maximized, where the set $\mathcal{F}$ consists of all frequency bins. However, considering all the possible pair-wise frequencies is computationally heavy in that even one sweep needs $O(|\mathcal{F}|^2)$ $score$ value calculations. Thus, we employ a strategy where we first perform a rough global optimization followed by a fine local optimization. These optimization procedures are explained in this subsection. With this strategy, the number of $score$ value calculations is reduced down to $O(|\mathcal{F}|)$ for one sweep.

1) Global Optimization With Single Centroid per Source: First, we perform a rough global optimization, where a centroid c_k is explicitly identified for each k and accordingly the goal function

$$\mathcal{J}(\{c_k\},\{\Pi_f\}) = \sum_{f\in\mathcal{F}} score\left[\mathbf{Q}\left(\left\{v_i^f\right\}\Big|_{\Pi_f},\{c_k\}\right)\right] \tag{26}$$

is maximized. The centroid c_k is calculated for each source as the average of the posterior probability sequences with the current permutations Π_f

$$c_k(\tau) \leftarrow \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} v_i^f(\tau)|_{i=\Pi_f(k)}, \quad \forall k,\tau \tag{27}$$

where $|\mathcal{F}|$ is the number of elements in the set $\mathcal{F}$. Note that the sequences v_i^f are normalized to zero-mean and unit-variance. On the other hand, the permutation Π_f is optimized to maximize the correlation coefficients ρ between posterior probability sequences v_i^f and the current centroid

$$\Pi_f \leftarrow \arg\max\nolimits_{\Pi} score\left[\mathbf{Q}\left(\left\{v_i^f\right\}\Big|_{\Pi},\{c_k\}\right)\right]. \tag{28}$$

The two operations (27) and (28) are iterated until convergence.

In (28), an exhaustive search through $N!$ permutations for the best one is feasible only with a very small N. Thus, we apply a simple yet effective heuristic method that reduces the size of $\mathbf{Q}$ one by one until it becomes very small: the mapping $i = \Pi(k)$ related to the maximum correlation coefficient ρ is decided immediately, and the ith row and the kth column are eliminated in the next step.

2) Global Optimization With Multiple Centroids per Source: According to the goal function (26), one centroid c_k is identified for each source k. This means that we expect similar posterior probability sequences for all the frequencies. However, if we increase the sampling rate, for example up to 16 kHz, the sequences are significantly different for the low and high frequency ranges. To model such source signals precisely, we introduce multiple centroids for a source, and modify the goal function (26) to

$$\mathcal{J}(\{c_{k,m}\},\{\Pi_f\}) = \sum_{f\in\mathcal{F}} \max\nolimits_m score\left[\mathbf{Q}\left(\left\{v_i^f\right\}\Big|_{\Pi_f},\{c_{k,m}\}\right)\right] \tag{29}$$

where $c_{k,m}$ is the mth centroid for source k. In practice, each source has two or three centroids (m =1, 2 or m =1, 2, 3).

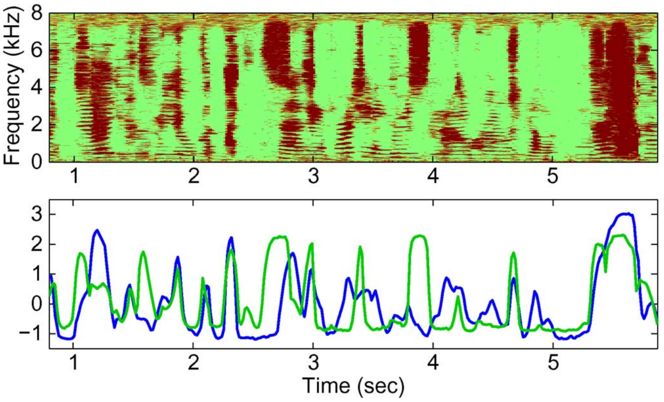

Fig. 7. Permutation aligned posterior probabilities $P(C_k|\mathbf{x})$ for separation of speech signals sampled at 16 kHz (above). And, two centroids $c_{k,1}$ and $c_{k,2}$ for the kth source obtained after the goal function (29) is maximized (below). Note that the centroids are normalized to zero-mean and unit-variance.

Fig. 7 shows an example. The upper plot shows permutation aligned posterior probabilities $P(C_k|\mathbf{x})$ for the separation of speech signals sampled at 16 kHz. The lower plot shows two centroids $c_{k,1}$ and $c_{k,2}$ obtained after the goal function (29) had been maximized. We observe that the blue line corresponds to most of the lower half frequencies and the green line corresponds to most of the higher half frequencies. In this way, multiple centroids model the activity pattern of a sound source more accurately than a single centroid.

The optimization procedure for the multiple-centroid goal function (29) is slightly complicated but not seriously so. Instead of using the simple average (27), the centroids $c_{k,m}$ are obtained through another level of clustering, where posterior probability sequences $v_i^f(\tau)|_{i=\Pi_f(k)}$ that belong to the kth source of all frequencies f are clustered. We employ the k-means algorithm [33] for the clustering. Then, $c_{k,m}$ is obtained as the average sequence of the mth cluster in the k-means algorithm. As regards the permutation optimization at each frequency, the (28) is slightly modified to

$$\Pi_f \leftarrow \arg\max\nolimits_{\Pi} \max\nolimits_m score\left[\mathbf{Q}\left(\left\{v_i^f\right\}\Big|_{\Pi},\{c_{k,m}\}\right)\right] \tag{30}$$

in the multiple-centroid version. As with the single centroid version, the calculation of multiple centroids by k-means and the permutation optimization by (30) are iterated until convergence.

3) Local Optimization: After completing the rough global optimization described above, we perform a fine local optimization for better permutation alignment. This maximizes the $score$ values over a set of selected frequencies $\mathcal{R}(f)$ for a frequency f

$$\Pi_f \leftarrow \arg\max\nolimits_{\Pi} \sum_{g\in\mathcal{R}(f)} score\left[\mathbf{Q}\left(\left\{v_i^f\right\}\Big|_{\Pi}, \{v_j^g\}|_{\Pi_g}\right)\right]. \tag{31}$$

The set $\mathcal{R}(f)$ preferably consists of frequencies g where a high correlation coefficient $\rho(v_i^f, v_j^g)$ would be attained for v_i^f and v_j^g corresponding to the same source. We typically select adjacent frequencies $\mathcal{A}(f)$ and harmonic frequencies $\mathcal{H}(f)$ so that $\mathcal{R}(f) = \mathcal{A}(f) \cup \mathcal{H}(f)$. For example, $\mathcal{A}$ is given by

$$\mathcal{A}(f)=\{f-3\Delta f, f-2\Delta f, f-\Delta f, f+\Delta f, f+2\Delta f, f+3\Delta f\}$$

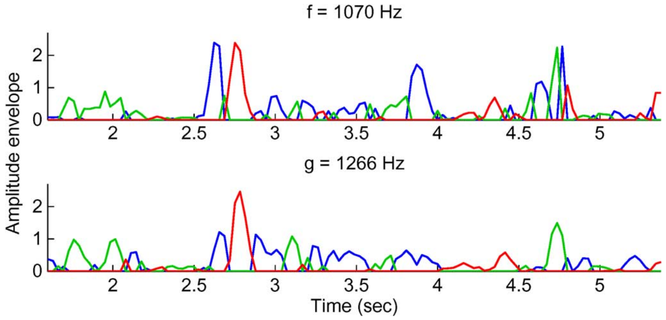

Fig. 8. Amplitude envelopes v_1^f, v_2^f, v_3^f at frequency $f = 1070$ Hz and v_1^g, v_2^g, v_3^g at frequency $g = 1266$ Hz. Permutations are aligned and the sequences originating from the same sound source are shown in the same color for ease of interpretation.

where $\Delta f = (1/L)f_s$, and $\mathcal{H}$ is given by

$$\mathcal{H}(f) = \{round(f/2) - \Delta f, round(f/2), \\ round(f/2) + \Delta f, 2f - \Delta f, 2f, 2f + \Delta f\}$$

where $round(\cdot)$ selects the nearest frequency to $\cdot$ from the set $\mathcal{F}$. The fine local optimization (31) is performed for one selected frequency f at a time, and repeated until no improvement is found for any frequency f.

E. Comparison to Amplitude Envelope

So far this section has described the procedure embodied in the Permutation Alignment stage. This subsection is devoted to a comparison of a posterior probability sequence and an amplitude envelope, used in the context of permutation alignment. Amplitude envelopes are widely used [9], [10], [24], [25] to represent the activity of separated signals and thus for permutation alignment.

An amplitude envelope is a sequence of the absolute values of separated frequency components

$$v_i^f(\tau) = |y_{ij}(\tau, f)|$$

defined along the time axis at a frequency. Here, the microphone index j is arbitrarily specified, but it should be the same over all frequencies f. Even before permutation alignment is conducted, $y_{ij}(\tau, f)$ can be temporarily calculated using (6) and (8).

Fig. 8 shows example amplitude envelopes. They are calculated from the separated frequency components in the same BSS execution and at the same frequencies as those shown in Fig. 6. We see some pattern similarity for the same source. The correlation coefficients $\rho(v_i^f, v_j^g)$ for these amplitude envelopes are

$$\begin{bmatrix} \rho\left(v_1^f, v_1^g\right) & \rho\left(v_1^f, v_2^g\right) & \rho\left(v_1^f, v_3^g\right) \\ \rho\left(v_2^f, v_1^g\right) & \rho\left(v_2^f, v_2^g\right) & \rho\left(v_2^f, v_3^g\right) \\ \rho\left(v_3^f, v_1^g\right) & \rho\left(v_3^f, v_2^g\right) & \rho\left(v_3^f, v_3^g\right) \end{bmatrix} \\ = \begin{bmatrix} 0.29 & 0.05 & -0.14 \\ 0.11 & 0.55 & -0.11 \\ -0.14 & -0.12 & 0.66 \end{bmatrix}. \quad (32)$$

We observe that $\rho(v_i^f, v_j^g)$ is positive for two sequences originating from the same sound source, and $\rho(v_i^f, v_j^g)$ has a small value around zero for those originating from two different sources. For (32), the $score$ value is 1.85, which is smaller than 2.66 that (23) has.

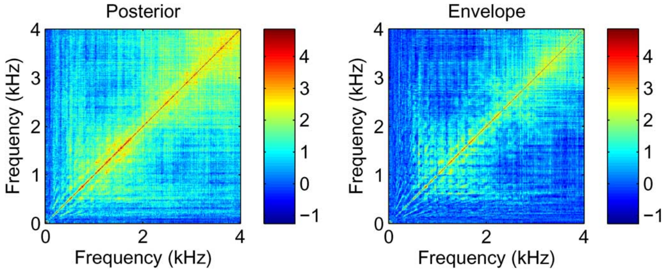

Fig. 9. $score[\mathbf{Q}]$ values defined in (25) calculated for every pair of frequencies. A case of the separation of three sources with two microphones. A larger number indicates a higher confidence in the permutation alignment between the corresponding two frequencies. Posterior probability sequences generally yield higher $score[\mathbf{Q}]$ values (1.11 in average) than amplitude envelopes (0.54 in average).

TABLE I
EXPERIMENTAL CONDITIONS

Number of microphones	$M = 3$
Number of sources	$N = 4$
Source signals	Speeches of 6 s
Reverberation time	$\mathrm{RT}_{60} = 130 \sim 450$ ms
Sampling rate	f_s = 8 kHz or 16 kHz
STFT frame size	$L = 1024$ (8 kHz) or 2048 (16 kHz) 128 ms
STFT frame shift	$S = 256$ (8 kHz) or 512 (16 kHz) 32 ms

Fig. 9 shows $score$ values for every pair of frequencies. We can see that posterior probability sequences generally exhibit higher $score$ values, i.e., there is a clearer contrast between same-source pairs and different-source pairs. This means that a posterior probability sequence has an advantage over an amplitude envelope in that permutation alignment is performed correctly and with more confidence.

A major difference between posterior probability sequences and amplitude envelopes can be found in the off-diagonal elements of a permutation aligned $\mathbf{Q}$ matrix (24), i.e., the correlation coefficients of two sequences from different sound sources. For posterior probability sequences, those correlations tend to be negative. This is because of the exclusiveness of a posterior probability. Namely, if the posterior probability for a class is high, that probability for another class is automatically low. The tendency helps in deciding permutations: pairing two sequences originating from different sources can clearly be avoided with a negative correlation.

V. EXPERIMENTS

A. Experimental Setups and Evaluation Measure

To verify the effectiveness of the proposed method, we conducted experiments designed to separate four speech sources with three microphones. The experimental conditions are summarized in Table I. We measured impulse responses $\mathrm{h}_{jk}(l)$ in a real room under the conditions shown in Fig. 10. The mixtures at the microphones were constructed by convolving the impulse responses and 6-s English speech sources.

The separation performance was evaluated in terms of the signal-to-distortion ratio (SDR) defined in [34]. To calculate

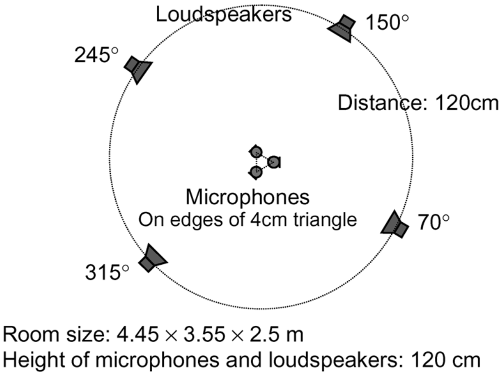

Fig. 10. Experimental setup.

SDR_k for output k, we first decompose the separated signals $\mathrm{y}_{k1}, \ldots, \mathrm{y}_{kM}$ as

$$\mathrm{y}_{kj}(t) = \mathrm{s}_{jk}^{\mathrm{img}}(t) + \mathrm{y}_{kj}^{\mathrm{spat}}(t) + \mathrm{y}_{kj}^{\mathrm{int}}(t) + \mathrm{y}_{kj}^{\mathrm{artif}}(t) \qquad (33)$$

where $\mathrm{y}_{kj}^{\mathrm{spat}}(t)$, $\mathrm{y}_{kj}^{\mathrm{int}}(t)$, and $\mathrm{y}_{kj}^{\mathrm{artif}}(t)$ are unwanted error components that correspond to spatial (filtering) distortion, interferences, and artifacts, respectively. These can be calculated by using a least-squares projection if we know all the source images $\mathrm{s}_{jk}^{\mathrm{img}}$ for all j and k. Then, SDR_k is calculated by the power ratio between the wanted and unwanted components

$$\mathrm{SDR}_k = 10\log_{10} \frac{\sum_{j=1}^{M} \sum_t \mathrm{s}_{jk}^{\mathrm{img}}(t)^2}{\sum_{j=1}^{M} \sum_t \left[\mathrm{y}_{kj}^{\mathrm{spat}}(t) + \mathrm{y}_{kj}^{\mathrm{int}}(t) + \mathrm{y}_{kj}^{\mathrm{artif}}(t)\right]^2}.$$

B. Separation Results With Various Reverberation Times

This subsection reports experimental results when the room reverberation time was varied from 130 to 450 ms by keeping/detaching some of the cushion walls in the experiment room. Fig. 11 shows the results. We examined six methods as shown in the figure. The first three methods were actual BSS methods. "Posterior" corresponds to the proposed method. "TDOA" and "Envelope" correspond to existing methods based on TDOA estimation [20] (compared in Section II-E), and based on amplitude envelope-based permutation alignment [10] (compared in Section IV-E), respectively. The other three methods were cheating methods that utilized source information. They were introduced to reveal the upper limit of the T–F masking separation performance and also to reveal the cause of separation performance degradation in the proposed BSS method. For "Ideal mask," we designed ideal T–F masks by

$$\mathcal{M}_k(\tau, f) = \begin{cases} 1, & \text{if } \sum_j \left|s_{jk}^{\mathrm{img}}\right|^2 \geq \sum_j \left|s_{jk'}^{\mathrm{img}}\right|^2, \ \forall k' \neq k \\ 0, & \text{otherwise.} \end{cases}$$

For "Ideal bin-wisemask," ideal frequency bin-wise T–F masks were designed in the same way as above, but permutation alignment were conducted by the proposed method using posterior probabilities, which were confined to 0 or 1 because of the ideal masks. With "Ideal permutation," T–F masks were designed by the method proposed in Section III, and then permutation ambiguities were ideally aligned by using the information on the source images s_{jk}^{img}. More specifically, true posterior probability sequences $\{u_k^f\}$ were calculated by using the source information, and then the permutation Π_f for each frequency f was calculated so that $score[\mathbf{Q}(\{v_i^f\}, \{u_k^f\})]$ was maximized.

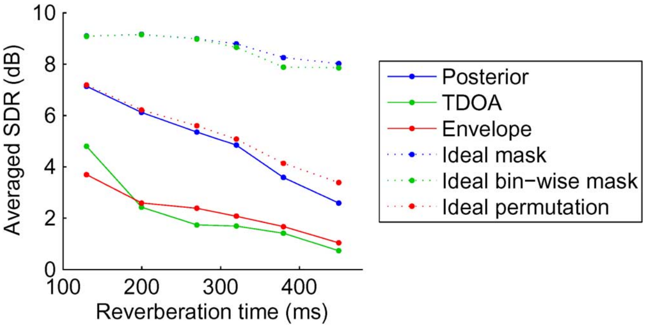

Fig. 11. Experimental results with various room reverberation times. Each point shows the averaged SDR over eight combinations of speeches under a specific experimental condition, which was defined by the reverberation time, the T–F mask design methodology and the permutation alignment method (detailed explanations are provided in the main text). The sampling rate was 8 kHz for the TDOA-based method to work properly without being affected by spatial aliasing.

We observe the following tendencies from the results. Our proposed method "Posterior" performed the best among the three actual BSS methods. "TDOA" performed moderately well only in the low reverberant (130 ms) condition. "Envelope" did not perform very well in many cases. We found that there was little difference between the separation performance of "Posterior" and "Ideal permutation," or "Ideal mask" and "Ideal bin-wise mask". This means that the proposed permutation alignment method utilizing posterior probabilities provided close to optimal performance. On the other hand, there was a large difference between "Ideal mask" and "Ideal permutation," especially with long reverberations.

The program was coded in Matlab and run on an Intel Core i7 965 (3.2-GHz) processor. The computational time was around 5 s for a set of 6-s speech mixtures. For permutation alignment by "Posterior" and "Envelope," we employed two centroids in the multiple-centroid cost function (29).

C. Effect of Permutation Alignment With Multiple Centroids

In the experiments described above, we used two centroids for modeling a source activity, where the sampling rate was 8 kHz. Even with a single centroid, the proposed permutation alignment method "Posterior" worked well, and the SDR numbers were almost the same with two centroids.

However, when we increased the sampling rate to 16 kHz, the effect of multiple centroids became prominent. Fig. 12 shows the SDR numbers for the separation of speech mixtures sampled at 16 kHz. We see that increasing the number of centroids from one or two to three had a great impact on the stable realization of good separation performance, whereas further increases in the number of centroids had little effect. These results support the discussion in Section IV-D2 numerically.

D. SiSEC 2008 Data

This subsection reports experimental results for publicly available benchmark data. We applied the proposed method to a set of data organized in the Signal Separation Evaluation

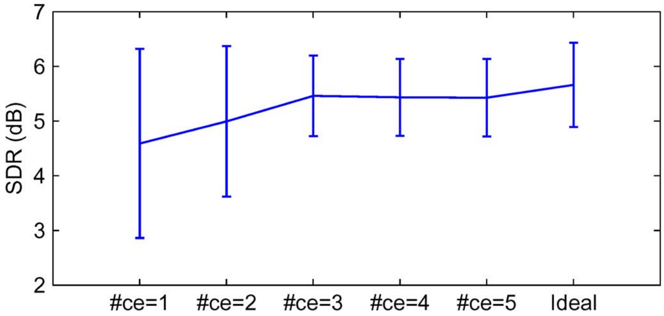

Fig. 12. Separation performance measured in SDR when employing multiple centroids in permutation alignment. The number of centroids varies from 1 to 5. Results with ideal permutations are also reported. A case with 270-ms room reverberation time, and 16-kHz sampling frequency. Separation runs of eight combinations of speech sources were evaluated. The error bars represent one standard deviation.

TABLE II
SEPARATION RESULTS FOR SISEC 2008 RECORDED DATA (IN SDR)

	RT_{60} = 130 ms		RT_{60} = 250 ms	
mic. spacing	5cm	1m	5cm	1m
`male3`	5.73 dB	6.46 dB	4.17 dB	5.95 dB
`female3`	6.45 dB	8.69 dB	5.91 dB	7.45 dB
`male4`	3.31 dB	4.44 dB	2.62 dB	3.41 dB
`female4`	3.92 dB	5.82 dB	3.49 dB	4.59 dB
`wdrums`	—	—	0.04 dB	-0.69 dB
`nodrums`	—	—	2.52 dB	1.20 dB
average	5.60 dB		3.39 dB	

Campaign (SiSEC 2008) [35]. We used the first development data (`dev1.zip`) in "Under-determined speech and music mixtures" data sets. Only live recording "`liverec`" data were used. Table II shows separation results measured in SDR. We found that the results for speech mixtures were substantially good compared to those reported in [35]. However, for music mixtures (`wdrums` and `nodrums`), the separation performance was not good. This is because the instrumental components, which were to be separated in the task, were often synchronized to each other. This situation was very difficult for the proposed permutation alignment method to deal with, because it is based on source activity sequences. An effective alternative way [36] is to employ nonnegative matrix factorization [37] in the context of convolutive BSS.

E. Live Recording

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 [38].

VI. CONCLUSION

This paper presented a method for underdetermined convolutive blind source separation. The two stage structure of the Clustering part considerably improves the separation performance compared with widely used methods based on TDOA. Permutation ambiguities that occur in the first stage are aligned by utilizing the information on posterior probabilities obtained in the first stage. This permutation alignment method performs better than a traditional method based on amplitude envelopes. For mixtures sampled at 16-kHz rate, the use of multiple centroids effectively models the source activities and yields better permutation alignment than a single centroid. Experimental results support these arguments very well. By comparing the separation performance in Fig. 11 with certain cheating methods (utilizing source information), we can see that there is room for improvement as regards frequency bin-wise clustering and separation. This could constitute future work.

APPENDIX
DERIVATION OF THE M-STEP UPDATE RULES

In the M-step shown in Section III-B, $Q(\theta,\theta')+\log p(\theta)$ by (17) is maximized with the parameter set θ by (14). This appendix shows the derivation of the parameter update rules.

As regards $\mathbf{a}_i$, it has the unit-norm constraint $||\mathbf{a}_i||^2 = 1$. Thus, with a Lagrange multiplier λ, we consider a function

$$L_1(\mathbf{a}_i,\lambda)=Q(\theta,\theta')+\log p(\theta)+\lambda\left(||\mathbf{a}_i||^2-1\right).$$

Setting the derivative of $L_1(\mathbf{a}_i,\lambda)$ with respect to $\mathbf{a}_i$, we obtain

$$\mathbf{R}\mathbf{a}_i=-\frac{\lambda}{\sigma_i^2}\mathbf{a}_i$$

with $\mathbf{R}$ defined by (18). Therefore, at stationary points, $\mathbf{a}_i$ should be an eigenvector of $\mathbf{R}$. By going back to the density function (12), we see that the eigenvector corresponding to the maximum eigenvalue gives the maximum of $L_1(\mathbf{a}_i,\lambda)$.

The update rule (19) is easily obtained by the derivative of $Q(\theta,\theta')$ with respect to σ_i^2.

As regards α_i, the property of mixture ratios $\sum_{i=1}^{N}\alpha_i = 1$ should be satisfied. Thus, again with a Lagrange multiplier λ, we consider a function

$$L_2(\alpha_i,\lambda)=Q(\theta,\theta')+\log p(\theta)+\lambda\left(\sum_{i=1}^{N}\alpha_i-1\right).$$

Setting the derivative of $L_2(\alpha_i,\lambda)$ with respect to α_i for $i = 1,\ldots,N$, we obtain

$$\sum_{\tau}^{T}P\left(C_i|\mathbf{x}(\tau),\theta'\right)+\phi-1+\alpha_i\lambda=0$$

for $i = 1,\ldots,N$. Summing these up with $i = 1,\ldots,N$, we have

$$\lambda=-\left[T+N\cdot(\phi-1)\right].$$

Then, we have (20).

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers who provided many valuable comments that helped us to improve the quality of this paper.

REFERENCES

[1] T.-W. Lee, *Independent Component Analysis—Theory and Applications*. Norwell, MA: Kluwer, 1998.

[2] *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*, S. Haykin, Ed. New York: Wiley, 2000.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[5] *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. New York: Springer, 2007.

[6] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[7] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[8] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. ICA 2000*, Jun. 2000, pp. 215–220.

[9] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, Oct. 2001.

[10] 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 Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[11] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA (LNCS 3889)*, Mar. 2006, pp. 601–608, Springer.

[12] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, Jan. 2007.

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

[14] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. ICASSP*, 2000, vol. 5, pp. 2985–2988.

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

[16] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.

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

[18] M. I. Mandel, D. P. W. 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.

[19] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[20] 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.

[21] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. WASPAA*, 2007, pp. 147–150.

[22] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. WASPAA*, Oct. 2007, pp. 139–142.

[23] Z. E. Chami, A. Pham, C. Servière, and A. Guerin, "A new model based underdetermined source separation," in *Proc. IWAENC*, 2008, pp. 147–150.

[24] 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 J. Adv. Signal Process.*, 2007, pp. Article ID 24 717, 12 pp..

[25] R. Olsson and L. Hansen, "Blind separation of more sources than sensors in convolutive mixtures," in *Proc. ICASSP'06*, May 2006, vol. V, pp. 657–660.

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

[27] P. D. O'Grady and B. A. Pearlmutter, "The LOST algorithm: Finding lines and separating speech mixtures," *EURASIP J. Adv. Signal Process.*, 2008, pp. Article ID 784 296, 17 pp..

[28] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[29] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1592–1604, Jul. 2007.

[30] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Evaluation of separation and dereverberation performance in frequency domain blind source separation," *Acoust. Sci. Technol.*, vol. 25, no. 2, pp. 119–126, 2004.

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[32] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[33] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley Interscience, 2000.

[34] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. ICA'07*, 2007, pp. 552–559 [Online]. Available: http://www.irisa.fr/metiss/SASSEC07/

[35] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. ICA'09*, 2009 [Online]. Available: http://sisec2008.wiki.irisa.fr/tiki-index.php

[36] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[37] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

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

Hiroshi Sawada (M'02–SM'04) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1991, 1993, and 2001, respectively.

He joined NTT Corporation in 1993. He is now the Group Leader of Learning and Intelligent Systems Research Group at the NTT Communication Science Laboratories, Kyoto, Japan. His research interests include statistical signal processing, audio source separation, array signal processing, machine learning, latent variable model, graph-based data structure, and computer architecture.

From 2006 to 2009, he served as an associate editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He is a member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE Signal Processing Society. He received the Ninth TELECOM System Technology Award for Student from the Telecommunications Advancement Foundation in 1994, the Best Paper Award of the IEEE Circuits and System Society in 2000, and the MLSP Data Analysis Competition Award in 2007. Dr. Sawada is a member of the IEICE and the ASJ.

Shoko Araki (M'01) received the B.E. and M.E. degrees from the University of Tokyo, Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D. degree from Hokkaido University, Sapporo, Japan, in 2007.

She is with NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. Since she joined NTT in 2000, she has been engaged in research on acoustic signal processing, array signal processing, blind source separation (BSS) applied to speech signals, meeting diarization, and auditory scene analysis.

Dr. Araki was a member of the organizing committee of the ICA 2003, the finance chair of IWAENC 2003, the registration chair of WASPAA 2007, and the evaluation co-chair of SiSEC2010. She received the 19th Awaya Prize from Acoustical Society of Japan (ASJ) in 2001, the Best Paper Award of the IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Academic Encouraging Prize from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2006, and the Itakura Prize Innovative Young Researcher Award from (ASJ) in 2008. She is a member of the IEICE and the ASJ.

Shoji Makino (A'89–M'90–SM'99–F'04) received B. E., M. E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1979, 1981, and 1993, respectively.

He joined NTT Corporation in 1981. He is now a Professor at University of Tsukuba, Ibaraki, Japan. His research interests include adaptive filtering technologies, the realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech, and acoustic signal processing for speech and audio applications. He is the author or coauthor of more than 200 articles in journals and conference proceedings and is responsible for more than 150 patents.

Prof. Makino received the ICA Unsupervised Learning Pioneer Award in 2006, the IEEE MLSP Competition Award in 2007, the TELECOM System Technology Award in 2004, the Achievement Award of the Institute of Electronics, Information, and Communication Engineers (IEICE) in 1997, and the Outstanding Technological Development Award of the Acoustical Society of Japan (ASJ) in 1995, the Paper Award of the IEICE in 2005 and 2002, the Paper Award of the ASJ in 2005 and 2002. He was a Keynote Speaker at ICA2007 and a Tutorial speaker at ICASSP2007. He has served on IEEE SPS Awards Board (2006–2008) and IEEE SPS Conference Board (2002–2004). He is a member of the James L. Flanagan Speech and Audio Processing Award Committee. He was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2002–2005) and is an Associate Editor of the *EURASIP Journal on Advances in Signal Processing*. He is a member of SPS Audio and Electroacoustics Technical Committee and the Chair of the Blind Signal Processing Technical Committee of the IEEE Circuits and Systems Society. He was the Vice President of the Engineering Sciences Society of the IEICE (2007–2008), and the Chair of the Engineering Acoustics Technical Committee of the IEICE (2006–2008). He is a member of the International IWAENC Standing committee and a member of the International ICA Steering Committee. He was the General Chair of WASPAA2007, the General Chair of IWAENC2003, the Organizing Chair of ICA2003, and is the designated Plenary Chair of ICASSP2012. He is an IEEE SPS Distinguished Lecturer (2009–2010), an IEICE Fellow, a council member of the ASJ, and a member of EURASIP.