Underdetermined BSS With Multichannel Complex NMF Assuming W-Disjoint Orthogonality of Source

Kazuma Takeda[†‡], Hirokazu Kameoka[∀‡], Hiroshi Sawada[‡],
Shoko Araki[‡], Shigeki Miyabe[†], Takeshi Yamada[†], Shoji Makino[†]

[†] Graduate School of Systems and Information Engineering, University of Tsukuba

[‡] NTT Communication Science Laboratories, NTT Corporation

[∀] Graduate School of Information Science and Technology, The University of Tokyo

***Abstract*—This paper presents a new method for underdetermined Blind Source Separation (BSS), based on a concept called multichannel complex non-negative matrix factorization (NMF). The method assumes (1) that the time-frequency representations of sources have disjoint support (W-disjoint orthogonality of sources), and (2) that each source is modeled as a superposition of components whose amplitudes vary over time coherently across all frequencies (amplitude coherence of frequency components) in order to jointly solve the indeterminacy involved in the frequency domain underdetermined BSS problem. We confirmed experimentally that the present method performed reasonably well in terms of the signal-to-interference ratio when the mixing process was known.**

***Index Terms*—Underdetermined blind source separation(BSS), non-negative matrix factorization(NMF), multichannel audio, W-disjoint orthogonality(W-DO), Expectation-maximization (EM) algorithm**

I. Introduction

BLIND Source Separation (BSS) is a technique for reconstructing source signals from microphone inputs on the assumption that the source signals and transfer characteristics between sources and microphones are unknown.

In the following, we consider a situation where K source signals are captured by M microphones. Here, let $y_m(\omega, t)$ be the short-time Fourier transform (STFT) component observed at the m-th microphone, and $s_k(\omega, t)$ be the STFT component of the k-th source. $1 \leq \omega \leq \Omega$ and $1 \leq t \leq T$ are the frequency and time indices, respectively. We assume that the length of the impulse response from a source to a microphone is sufficiently shorter than the frame length of the STFT, in which case the observed signals can be approximated fairly well by an instantaneous mixture in the frequency domain:

$$\boldsymbol{y}(\omega, t) = \boldsymbol{A}(\omega)\boldsymbol{s}(\omega, t) + \boldsymbol{n}(\omega, t), \tag{1}$$

where $\boldsymbol{y}(\omega, t) = (y_1(\omega, t), \ldots, y_M(\omega, t))^{\mathrm{T}} \in \mathbb{C}^M$ and $\boldsymbol{s}(\omega, t) = (s_1(\omega, t), \ldots, s_K(\omega, t))^{\mathrm{T}} \in \mathbb{C}^K$. $\boldsymbol{A}(\omega) = (a_{m,k}(\omega))_{M\times K} = (\boldsymbol{a}_1(\omega), \ldots, \boldsymbol{a}_K(\omega)) \in \mathbb{C}^{M\times K}$ denotes a mixing matrix that consists of the frequency response $a_{m,k}(\omega)$ of the transfer characteristics, which is assumed to be time-invariant. $\boldsymbol{n}(\omega, t)$ is assumed to comprise all kinds of components that cannot be expressed by the instantaneous mixture representation (e.g., background noise and reverberant components). BSS based on this observation model is called frequency domain BSS, which allows for a fast implementation compared with BSS that use a time domain convolutive mixture model.

To estimate the unknown mixing matrix and source signals solely from observed signals, we must make some assumptions about sources, and formulate an appropriate optimization problem based on criteria designed according to those assumptions. For example, if the observed signals outnumber the sources, we can employ independent component analysis (ICA) [1], which estimates the separation matrix (the inverse of the mixing matrix) such that the independence of the source estimates is maximized. For an underdetermined case where there are fewer observations than sources, there are an infinite number of solutions for source signals even if a mixed process is known. Hence, the independence assumption is too weak to allow us to determine a unique solution. For this reason, in underdetermined situations, we need a stronger assumption than independence such as sparseness [2]–[6]. Moreover, even if separation is achieved in each frequency bin ω so that we obtain K components $s_1(\omega, t), \ldots, s_K(\omega, t)$, which component belongs to which source remains ambiguous. Therefore, some additional clue is required to make it possible to group together the components considered to originate from the same source. This kind of problem is usually called a permutation problem. Various approaches to tackling the permutation problem have already been proposed. They include approaches that use the direction-of-arrival information of sources [7], the correlation of amplitudes in sources between different frequency bins [8], and the combination of the above two pieces of information and a harmonicity constraint [9].

As mentioned above, the frequency domain underdetermined BSS problem simultaneously involves indeterminacy in terms of the inverse mixing process and permutation ambiguity. To effectively avoid the indeterminacy, we find it natural to introduce a unified framework incorporating (1) an observation model (mixed process) that assumes sparseness of sources in the time-frequency domain, and (2) a source model consisting of components each of whose amplitudes varies over time coherently across frequencies. We have introduced a framework called "multichannel complex NMF" based on this concept (which is an extension of the complex NMF framework proposed in [10] to a multichannel case), and have already proposed several different methods [11], [12]. Ozerov introduces a similar concept in [13], where an attempt is made to extend non-negative matrix factorization (NMF) to a multichannel case. It is important to note here that the multichannel complex NMF framework differs from Ozerov's framework in that it uses a complex factor representation introduced in [10] to model the means of Gaussian source components whereas Ozerov's uses a non-negative factor representation to model the variances.

As a contribution to this series of approaches, this paper proposes a multichannel complex NMF method assuming the W-disjoint orthogonality of the source components, which can be thought of as one example of the notion of sparseness.

II. Multichannel Complex NMF

A. Assumptions on sources

We henceforth assume the use of the observation model given by (1) and also assume that $\boldsymbol{n}(\omega, t)$ is a complex

Gaussian random variable that is uncorrelated in time, space and frequency:

$$\boldsymbol{n}(\omega,t) \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2(\omega)\boldsymbol{I}). \tag{2}$$

Moreover, we assume

$$\|\boldsymbol{a}_k(\omega)\|_2^2 = 1,\ a_{1,k}(\omega) \geq 0,\ 1 \leq k \leq K, \tag{3}$$

to remove the scaling indeterminacy and an ambiguity in the sign between $\boldsymbol{A}(\omega)$ and $\boldsymbol{s}(\omega,t)$. We assume the following in relation to the source signals:

Assumption 1 (Amplitude coherence of sources)**.** *Each source is expressed as the sum of I elements each of whose amplitudes varies over time coherently across frequencies.*

Let us now consider expressing the STFT component of the i-th element of the k-th source using a time-invariant spectral shape $H_k^i(\omega)$, a time-varying gain $U_k^i(t)$ and a time-varying phase spectrum $\phi_k^i(\omega,t)$, such that

$$x_k^i(\omega,t) = H_k^i(\omega)U_k^i(t)e^{\mathrm{j}\phi_k^i(\omega,t)}. \tag{4}$$

We assume for convenience

$$\sum_{\omega} H_k^i(\omega) = 1, \tag{5}$$

to eliminate the scaling indeterminacy between $H_k^i(\omega)$ and $U_k^i(t)$. We note that since $H_k^i(\omega)$ is time-invariant and $U_k^i(t)$ is frequency-invariant, using the above representation amounts to assuming that the amplitude $|x_k^i(\omega,t)|$ is constrained to vary over time coherently across frequencies. By using (4), we can now construct a source model incorporating Assumption 1:

$$s_k(\omega,t) = \sum_{i=1}^{I_k} x_k^i(\omega,t) + \epsilon_k(\omega,t). \tag{6}$$

$\epsilon_k(\omega,t)$ is used to denote a modeling error for each source. Since a signal model expressed as the sum of (4) is based on the complex factor model employed in the complex NMF framework [10], we call a BSS framework based on the source model given by (6) a "multichannel complex NMF".

Next, we make an assumption about the mixing process.

Assumption 2 (Sparseness of sources)**.** *Only one source is active in each time-frequency bin.*

This assumption is called the W-disjoint orthogonality (W-DO) [2], which is known to help solve underdetermined BSS problems effectively, especially for sparse sources. Now, let us use $\hat{k}(\omega,t)$ to denote the (unknown) active source index at time-frequency bin (ω,t). Then, with this assumption, (1) can be rewritten as:

$$\boldsymbol{y}(\omega,t) = \boldsymbol{a}_{\hat{k}(\omega,t)}(\omega)s_{\hat{k}(\omega,t)}(\omega,t) + \boldsymbol{n}(\omega,t) \tag{7}$$

$$= \boldsymbol{A}(\omega)\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{s}(\omega,t) + \boldsymbol{n}(\omega,t). \tag{8}$$

Here, we have used $\boldsymbol{Z}_{\hat{k}(\omega,t)}$ to denote a $K \times K$ matrix whose component at row n and column m is given by $\delta_{n,\hat{k}(\omega,t)} \cdot \delta_{m,\hat{k}(\omega,t)}$ (where $\delta_{n,m}$ is the Kronecker delta).

B. Probability density function of observed signals

Based on the above assumptions, we now derive the probability density function of the observed signals. Let $\epsilon_k(\omega,t)$ be a Gaussian noise with variance ν_k^2:

$$\epsilon_k(\omega,t) \sim \mathcal{N}_{\mathbb{C}}(0, \nu_k^2). \tag{9}$$

Hence, $\boldsymbol{s}(\omega,t)$ is normally distributed such that

$$\boldsymbol{s}(\omega,t) \sim \mathcal{N}_{\mathbb{C}}\big(\boldsymbol{\mu}(\omega,t), \boldsymbol{\Lambda}\big), \tag{10}$$

where

$$\boldsymbol{\mu}(\omega,t) := \big(\mu_1(\omega,t), \ldots, \mu_K(\omega,t)\big)^{\mathrm{T}}, \tag{11}$$

$$\mu_k(\omega,t) := \sum_{i=1}^{I_k} H_k^i(\omega)U_k^i(t)e^{\mathrm{j}\phi_k^i(\omega,t)}, \tag{12}$$

$$\boldsymbol{\Lambda} := \mathrm{diag}\big(\nu_1^2, \ldots, \nu_K^2\big). \tag{13}$$

According to (2), (8) and (10), the conditional probability density function of $\boldsymbol{y}(\omega,t)$ given values of

$$\theta := \{\boldsymbol{a}_k(\omega), \pi_k, \sigma^2(\omega), \nu_k^2\}_{1\leq k\leq K, 1\leq\omega\leq\Omega}, \tag{14}$$

$$\pi_k := \{H_k^i(\omega), U_k^i(t), \phi_k^i(\omega,t)\}_{1\leq i\leq I_k, 1\leq\omega\leq\Omega, 1\leq t\leq T},$$

is described as:

$$p(\boldsymbol{y}(\omega,t)|\hat{k}(\omega,t),\theta) = \mathcal{N}_{\mathbb{C}}\big(\boldsymbol{A}(\omega)\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{\mu}(\omega,t),$$
$$\boldsymbol{A}(\omega)\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{\Lambda}\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{A}(\omega)^{\mathrm{H}} + \sigma^2(\omega)\boldsymbol{I}\big). \tag{15}$$

According to (8), the conditional probability density function of $\boldsymbol{y}(\omega,t)$ given the values of θ, $\hat{k}(\omega,t)$ and $\boldsymbol{s}(\omega,t)$ is:

$$p(\boldsymbol{y}(\omega,t)|\boldsymbol{s}(\omega,t),\hat{k}(\omega,t),\theta)$$
$$= \mathcal{N}_{\mathbb{C}}\big(\boldsymbol{A}(\omega)\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{s}(\omega,t), \sigma^2(\omega)\boldsymbol{I}\big). \tag{16}$$

III. Maximum Likelihood Estimation of $\boldsymbol{\theta}$ Using the EM Algorithm

Here we derive an iterative algorithm that searches for the maximum likelihood (ML) estimate of θ given the observed data $y := \{\boldsymbol{y}(\omega,t)\}_{1\leq\omega\leq\Omega,1\leq t\leq T}$. By treating y as incomplete data and a set consisting of y, $s = \{\boldsymbol{s}(\omega,t)\}_{1\leq\omega\leq\Omega,1\leq t\leq T}$ and $\hat{k} = \{\hat{k}(\omega,t)\}_{1\leq\omega\leq\Omega,1\leq t\leq T}$ as complete data, the ML problem under consideration can be viewed as an incomplete data problem, which can be dealt with using the Expectation-Maximization (EM) algorithm. The first step when devising an EM algorithm is to define the Q function. The Q function is defined as the conditional expectation of the complete data log-likelihood $\log p(c|\theta)$ where $c := \{y, s, \hat{k}\}$ with respect to the latent variables $s(\omega,t)$ and $\hat{k}(\omega,t)$, given the values of the observed data $y(\omega,t)$ and the model parameters $\theta = \theta'$. Hence, the Q function $Q(\theta,\theta')$ is

$$Q(\theta,\theta') = \big\langle \log p(c|\theta)\big\rangle_{p(s,\hat{k}|y,\theta=\theta')} \tag{17}$$

$$= \big\langle\big\langle \log p(c|\theta)\big\rangle_{p(s|\hat{k},y,\theta=\theta')}\big\rangle_{p(\hat{k}|y,\theta=\theta')} \tag{18}$$

$$\stackrel{\theta}{=} \big\langle\big\langle \log P(y|s,\hat{k},\theta)\big\rangle_{p(s|\hat{k},y,\theta=\theta')}\big\rangle_{p(\hat{k}|y,\theta=\theta')}$$
$$+ \big\langle\big\langle \log P(s|\theta)\big\rangle_{p(s|\hat{k},y,\theta=\theta')}\big\rangle_{p(\hat{k}|y,\theta=\theta')}. \tag{19}$$

The notation $\langle f(\xi)\rangle_{p(\xi)}$ is used to indicate the expectation of $f(\xi)$ with respect to ξ that is assumed have the distribution $p(\xi)$. $\stackrel{\xi}{=}$ denotes equality up to terms not depending on ξ. We give the concrete form of each term of (19) below.

Firstly, $\big\langle\big\langle \log p(y|s,\hat{k},\theta)\big\rangle_{p(s|\hat{k},y,\theta=\theta')}\big\rangle_{p(\hat{k}|y,\theta=\theta')}$ can be written as:

$$\big\langle\big\langle \log p(y|s,\hat{k},\theta)\big\rangle_{p(s|\hat{k},y,\theta=\theta')}\big\rangle_{p(\hat{k}|y,\theta=\theta')}$$
$$\stackrel{\theta}{=} -TM\sum_{\omega}\log\sigma^2(\omega)$$
$$- \sum_{\omega,t}\sum_{\hat{k}(\omega,t)=1}^{K}\frac{m_{\hat{k}(\omega,t)}}{\sigma^2(\omega)}\mathrm{tr}\big(\boldsymbol{F}^y_{\hat{k}(\omega,t)}(\omega,t)\big), \tag{20}$$

where $\boldsymbol{F}^{y}_{\hat{k}(\omega,t)}(\omega,t)$ is given by

$$\begin{aligned}\boldsymbol{F}^{y}_{\hat{k}(\omega,t)}(\omega,t) :=& \boldsymbol{y}(\omega,t)\boldsymbol{y}(\omega,t)^{\mathrm{H}} \\ &-\boldsymbol{A}(\omega)\boldsymbol{Z}_{\hat{k}(\omega,t)}\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t)\boldsymbol{y}(\omega,t)^{\mathrm{H}} \\ &-\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{A}(\omega)^{\mathrm{H}}\boldsymbol{y}(\omega,t)\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t)^{\mathrm{H}} \\ &+\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{A}(\omega)^{\mathrm{H}}\boldsymbol{A}(\omega)\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{R}_{\hat{k}(\omega,t)}(\omega,t),\end{aligned} \tag{21}$$

with

$$\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t) := \boldsymbol{\mu}'(\omega,t) + \boldsymbol{G}_{\hat{k}(\omega,t)}(\omega)\boldsymbol{y}(\omega,t), \tag{22}$$

$$\begin{aligned}\boldsymbol{R}_{\hat{k}(\omega,t)}(\omega,t) :=& \bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t)\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t)^{\mathrm{H}} + \boldsymbol{\Lambda}' \\ &- \boldsymbol{G}_{\hat{k}(\omega,t)}(\omega)\boldsymbol{A}'(\omega)\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{\Lambda}',\end{aligned} \tag{23}$$

$$\begin{aligned}\boldsymbol{G}_{\hat{k}(\omega,t)}(\omega) :=& \boldsymbol{\Lambda}'\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{A}'(\omega)^{\mathrm{H}} \\ &\big(\boldsymbol{A}'(\omega)\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{\Lambda}'\boldsymbol{Z}_{\hat{k}(\omega,t)}\boldsymbol{A}'(\omega)^{\mathrm{H}} + \sigma'^{2}(\omega)\boldsymbol{I}\big)^{-1}.\end{aligned} \tag{24}$$

In the above, we have used $\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t)$ and $\boldsymbol{R}_{\hat{k}(\omega,t)}(\omega,t)$ to denote $\langle \boldsymbol{s}(\omega,t)\rangle_{p(s|y,\hat{k},\theta=\theta')}$ and $\langle \boldsymbol{s}(\omega,t)\boldsymbol{s}(\omega,t)^{\mathrm{H}}\rangle_{p(s|y,\hat{k},\theta=\theta')}$, respectively. We note that $\boldsymbol{G}_{\hat{k}(\omega,t)}(\omega)$ has the same form as the Wiener filter. $m_{\hat{k}(\omega,t)}$ denotes the probability that the k-th source is active at time-frequency bin (ω,t):

$$m_{\hat{k}(\omega,t)} := p(\hat{k}(\omega,t)|\boldsymbol{y}(\omega,t),\theta=\theta'). \tag{25}$$

ξ' corresponds to the value of ξ when $\theta=\theta'$. On the other hand, $\langle\langle \log P(s|\theta)\rangle_{p(s|\hat{k},y,\theta=\theta')}\rangle_{p(\hat{k}|y,\theta=\theta')}$ is written as:

$$\begin{aligned}&\big\langle\big\langle \log P(s|\theta)\big\rangle_{p(s|\hat{k},y,\theta=\theta')}\big\rangle_{p(\hat{k}|y,\theta=\theta')} \\ &\overset{\theta}{=} -T\sum_{\omega}\log|\boldsymbol{\Lambda}| \\ &\quad -\sum_{\omega,t}\sum_{\hat{k}(\omega,t)=1}^{K} m_{\hat{k}(\omega,t)}\mathrm{tr}\big(\boldsymbol{\Lambda}^{-1}\boldsymbol{F}^{s}_{\hat{k}(\omega,t)}(\omega,t)\big),\end{aligned} \tag{26}$$

where $\boldsymbol{F}^{s}_{\hat{k}(\omega,t)}(\omega,t)$ is given by

$$\begin{aligned}\boldsymbol{F}^{s}_{\hat{k}(\omega,t)}(\omega,t) :=& \boldsymbol{R}_{\hat{k}(\omega,t)}(\omega,t) \\ &-\boldsymbol{\mu}(\omega,t)\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t)^{\mathrm{H}} \\ &-\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t)\boldsymbol{\mu}(\omega,t)^{\mathrm{H}} \\ &+\boldsymbol{\mu}(\omega,t)\boldsymbol{\mu}(\omega,t)^{\mathrm{H}}.\end{aligned} \tag{27}$$

Altogether, the term involving θ in the Q function is expressed as the sum of (20) and (26).

A. E step

In the E step, we substitute the newest value of θ, updated at the previous M step, into θ'. Specifically, we update the values of $\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t)$, $\boldsymbol{R}_{\hat{k}(\omega,t)}(\omega,t)$ and $m_{\hat{k}(\omega,t)}$ using (22), (23) and (25). If we assume $p(\hat{k}(\omega,t)|\theta)=1/K$, (25) can be written in a more concise form

$$\begin{aligned}m_{\hat{k}(\omega,t)} :=& p(\hat{k}(\omega,t)|\boldsymbol{y}(\omega,t),\theta=\theta') \\ =& \frac{p(\boldsymbol{y}(\omega,t)|\hat{k}(\omega,t),\theta=\theta')}{\sum_{k'=1}^{K} p(\boldsymbol{y}(\omega,t)|\hat{k}(\omega,t)=k',\theta=\theta')},\end{aligned} \tag{28}$$

which can be computed using the distribution given by (15).

B. M Step

Now, we consider updating θ to values such that $Q(\theta,\theta')$ is guaranteed to be non-decreasing. First, we consider fixing all the parameters in θ except for $\boldsymbol{A}(\omega)$ at constant values. Then, the Q function is maximized with respect to $\boldsymbol{A}(\omega)$ when

$$\boldsymbol{A}(\omega) = \boldsymbol{\Gamma}(\omega)\boldsymbol{\Sigma}(\omega)^{-1}, \tag{29}$$

$$\boldsymbol{\Gamma}(\omega) := \sum_{t}\boldsymbol{y}(\omega,t)\sum_{\hat{k}} m_{\hat{k}}\bar{\boldsymbol{s}}_{\hat{k}}(\omega,t)^{\mathrm{H}}\boldsymbol{Z}_{\hat{k}}, \tag{30}$$

$$\boldsymbol{\Sigma}(\omega) := \sum_{t}\sum_{\hat{k}} m_{\hat{k}}\boldsymbol{Z}_{\hat{k}}\boldsymbol{R}_{\hat{k}}(\omega,t)\boldsymbol{Z}_{\hat{k}}. \tag{31}$$

Care must be taken that when $\boldsymbol{A}(\omega)$ is updated by (29), the condition (3) is not necessarily satisfied and so it is necessary to perform an appropriate normalization. Next, we consider updating π_k to values such that $Q(\theta,\theta')$ is guaranteed to be non-decreasing. The term related to π_k in the Q function is

$$Q(\theta,\theta') \overset{\pi_k}{=} \frac{1}{\nu_k^2}\sum_{\omega,t}\big|[\hat{\boldsymbol{s}}(\omega,t)]_k - \mu_k(\omega,t)\big|^2, \tag{32}$$

$$\hat{\boldsymbol{s}}(\omega,t) := \sum_{\hat{k}(\omega,t)=1}^{K} m_{\hat{k}(\omega,t)}\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t), \tag{33}$$

where $[\cdot]_k$ indicates the k-th element of a vector. We note that (32) has exactly the same form as the objective function of Complex NMF [10]. Thus, we can locally maximize $Q(\theta,\theta')$ with respect to π_k through an iterative algorithm described in [10], which consists of performing the following updates:

$$H_k^i(\omega) \leftarrow H_k^i(\omega)\frac{\sum_t U_k^i(t)|[\hat{\boldsymbol{s}}(\omega,t)]_k|}{\sum_t U_k^i(t)\sum_j H_k^j(\omega)U_k^j(t)}, \tag{34}$$

$$U_k^i(t) \leftarrow U_k^i(t)\frac{\sum_\omega H_k^i(\omega)|[\hat{\boldsymbol{s}}(\omega,t)]_k|}{\sum_\omega H_k^i(\omega)\sum_j H_k^j(\omega)U_k^j(t)}, \tag{35}$$

$$\phi_k^i(\omega,t) \leftarrow \arg([\hat{\boldsymbol{s}}(\omega,t)]_k). \tag{36}$$

Readers are referred to [10] for the derivation process of (34), (35) and (36). Note that the update rules for $\sigma^2(\omega)$ and $\nu_k^2(\omega)$ can also be obtained analytically by setting the partial derivatives of $Q(\theta,\theta')$ at zero, but for convenience, we fixed $\sigma^2(\omega)$ and $\nu_k^2(\omega)$ at constant values in the following experiments.

C. Obtaining separated signals

One natural choice for an estimate of separated signals would be $\langle \boldsymbol{s}(\omega,t)\rangle_{p(\boldsymbol{s}(\omega,t)|\boldsymbol{y}(\omega,t),\theta)}$. We can notice from

$$\begin{aligned}&\langle \boldsymbol{s}(\omega,t)\rangle_{p(\boldsymbol{s}(\omega,t)|\boldsymbol{y}(\omega,t),\theta)} \\ &\quad= \big\langle\langle \boldsymbol{s}(\omega,t)\rangle_{p(\boldsymbol{s}(\omega,t)|\boldsymbol{y}(\omega,t),\hat{k}(\omega,t),\theta)}\big\rangle_{p(\hat{k}(\omega,t)|\boldsymbol{y}(\omega,t),\theta)} \\ &\quad= \langle \bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t)\rangle_{p(\hat{k}(\omega,t)|\boldsymbol{y}(\omega,t),\theta)} \\ &\quad= \sum_{\hat{k}(\omega,t)=1}^{K} m_{\hat{k}(\omega,t)}\bar{\boldsymbol{s}}_{\hat{k}(\omega,t)}(\omega,t) = \hat{\boldsymbol{s}}(\omega,t),\end{aligned} \tag{37}$$

that $\langle \boldsymbol{s}(\omega,t)\rangle_{p(\boldsymbol{s}(\omega,t)|\boldsymbol{y}(\omega,t),\theta)}$ amounts exactly to (33).

IV. EXPERIMENTS

We evaluated the performance of the proposed method in terms of source separation ability. Although the present method is intended to solve all the indeterminacy involved in the frequency-domain underdetermined BSS problem, we decided to preliminarily confirm the effect of the present method only in terms of the indeterminacy with respect to the inverse mixing process. Thus, we assumed an oracle situation such that the true mixing matrix $\boldsymbol{A}(\omega)$ is given.

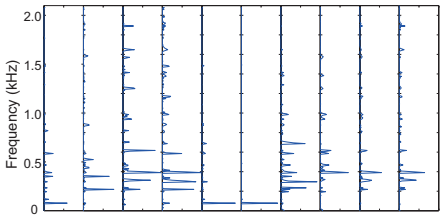

Figure. 1. Basis spectrum $H_1^i(\omega)$ of Source 1 (corresponding to the piano sound). $H_1^1(\omega), \cdots, H_1^{10}(\omega)$ are each displayed in a column.

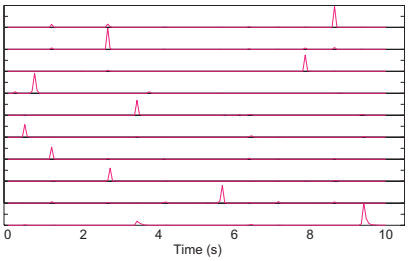

Figure. 2. Elementwise time-varying gain $U_1^i(t)$ of Source 1. $U_1^1(t)$, $\cdots, U_1^{10}(t)$ are each displayed in a row.

Table. I
SIRs OF EACH SOURCE BEFORE AND AFTER SEPARATION [dB]

	Before	After
Source1: Piano	-6.8	17.7
Source2: Vocal	-6.1	15.4
Source3: Drum	2.9	14.0

We used a 10-second stereo musical signal with a sampling rate of 16kHz as a test signal, which we obtained by mixing 3 sound sources (piano, vocal, and drum) using a measured room impulse response (in which the distance between the microphones and the reverberation time were assumed to be 1m and 100ms, respectively). To compute the STFT components of the observed signal, the STFT frame length was set at 128ms and a Hanning window was used with an overlap length of 64ms. The EM algorithm was run for 50 iterations. The mixing matrix $\boldsymbol{A}(\omega)$ was set at the true value, and all other parameters were initialized randomly. The number of basis spectra was set at $I_1 = \cdots = I_K = 10$. The separated signals were defined as (36). For an objective performance criterion, we used the Source-to-Interference Ratio (SIR) [14] for a quantitative evaluation of the distortion caused by other signals. The SIR is expressed in decibels (dB), and a higher SIR indicates superior quality. We compared the SIRs of the signals prior to separation with that of the separated signals obtained with the present method. The result can be seen in Tab I. The basis spectrum of the piano sound, $H_k^i(\omega)$, are shown in Fig 1, and the time-varying gain for each basis spectrum, $U_k^i(t)$, is shown in Fig 2.

Tab I confirms that there were significant improvements in the SIRs. From Fig 1, 2, we observe a harmonic structure in each basis spectrum and a fast rise followed by a fast decay in each gain function. This should imply that the element signals are successfully estimated such that each one is associated with a single piano note. We also confirmed through a number of evaluations with different $I_1, \ldots, I_K$ settings that $I_1 = \cdots = I_K = 10$ gave the best performance. Overall, we confirmed that reasonably good separation was obtained with the present method if the mixing matrix $\boldsymbol{A}(\omega)$ was known.

V. CONCLUSION

This paper proposed a multichannel complex NMF method based on the W-disjoint orthogonality of sources, which is intended to solve the indeterminacy involved in the underdetermined BSS problem. An experimental evaluation showed that reasonably good separation was obtained with the present method if the mixing matrix was known.

REFERENCES

[1] A. Hyvarinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.

[2] 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, 2004.

[3] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in Adv. Neural Info. Process. Syst., 2006, pp. 953-960.

[4] 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.

[5] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano,T. Hiekata, and T. Morita, "Real-time implementation of twostage blind source separation combining SIMO-ICA and binary masking," in Proc. Intl. Workshop Acoust., Echo, NoiseControl (IWAENC '05), 2005, pp. 229-232.

[6] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in Proc. 2007 IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA '07), 2007, pp. 147-150.

[7] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura,"Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in Proc. 2000 IEEE Intl.Conf. Acoust. Speech, Signal Process. (ICASSP '00), 2000, pp. 3140-3143.

[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, 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 Process., vol. 12, pp. 530-538, 2004.

[10] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals,"in Proc. 2009 IEEE Intl. Conf. Acoust., Signal, Speech Process. (ICASSP '09), 2009, pp. 3437-3440.

[11] Y. Kitano, H. Kameoka, Y. Izumi, N. Ono, and S. Sagayama,"A sparse component model of source signals and its application to blind source separation," in Proc. 2010 IEEE Intl. Conf. Acoust. Speech, Signal Process., 2010, pp. 4122-4125.

[12] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Formulations and algorithms for multichannel complex NMF," in Proc. 2011 IEEE Intl. Conf. Acoust. Speech, Signal Process., 2011, to appear.

[13] 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, 2010.

[14] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign:data, algorithms and results," in Proc. Intl. Conf. on Independent Component Analysis and Signal Separation, 2007, pp. 552-559.