Audio source separation with multiple microphones on time-frequency representations

Hiroshi Sawada

NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper presents various source separation methods that utilize multiple microphones. We classify them into two classes. Methods that fall into the first class apply independent component analysis (ICA) or Gaussian mixture model (GMM) to frequency bin-wise observations, and then solve the permutation problem to reconstruct separated signals. The second type of method extends non-negative matrix factorization (NMF) to a multi-microphone situation, in which NMF bases are clustered according to their spatial properties. We have a unified understanding that all methods analyze a time-frequency representation with an additional microphone axis.

Keywords: Source separation, Short-time Fourier transform, Time-frequency representation, Independent component analysis, Gaussian mixture model, Non-negative matrix factorization, Permutation problem

1. INTRODUCTION

Many research efforts have been devoted to separate audio sources mixed in a real-room situation.[1,2] If the mixed signals are captured with multiple microphones, we can exploit the spatial diversity of the sources, and therefore expect good separation.

In this paper, we explain various audio source separation methods that utilize multiple microphones. Particularly, we focus on time-frequency representations that are obtained by applying a short-time Fourier transform (STFT) to the observed signal. Then, observed signals at multiple microphones are represented as a tensor that has frequency, time, and microphone axes, as shown in Fig. 1.

The methods described in this paper fall into two categories depending on along which axis the tensor of the observed signals is sliced in order to apply a basic statistical tool. The methods in the first category (shown in the top part of Fig. 1) slice the tensor in a frequency bin-wise manner and apply independent component analysis (ICA)[3,4] or a Gaussian mixture model (GMM)[5] to the sliced matrices with a time frame axis and a microphone axis. The methods in the second category (shown in the bottom part of Fig. 1) slice the tensor in a microphone-wise manner and apply non-negative matrix factorization (NMF)[6,7] to the sliced matrices with a frequency bin axis and a time frame axis.

The methods in the first category need to resolve the permutation ambiguities[8] that inherent in the ICA or GMM solutions. There have been methods designed to resolve the ambiguities as a post-processing.[9,10] As more advanced approaches, independent vector analysis (IVA)[11–13] and unified statistical models[14,15] have been proposed for resolving the permutation ambiguities simultaneously with ICA and GMM, respectively.

The methods in the second category need to cluster frequency sound patterns extracted as NMF bases for each source. Multichannel NMF[16,17] is a way of estimating the information related to the source location for each NMF basis, and clustering the NMF bases according to the estimated information.

This paper is organized as follows. The next section formulates the source separation problem. Sections 3 and 4 describe the two categorized methods. Section 5 concludes this paper.

Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XI,
edited by Harold H. Szu, Proc. of SPIE Vol. 8750, 875007 · © 2013 SPIE
CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2018632

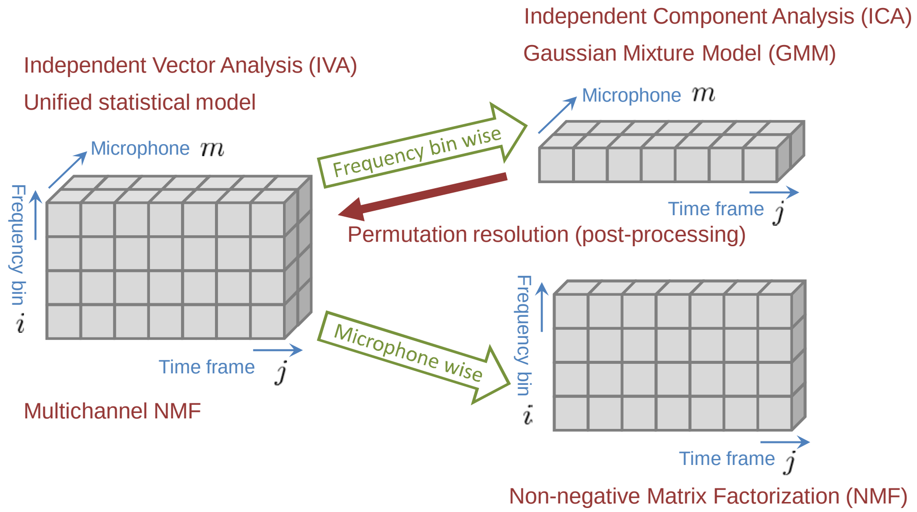
Independent Component Analysis (ICA)
Gaussian Mixture Model (GMM)
Independent Vector Analysis (IVA)
Unified statistical model
Microphone m
Frequency bin wise
Microphone m
Frequency bin i
Time frame j
Permutation resolution (post-processing)
Time frame j
Microphone wise
Frequency bin i
Multichannel NMF
Time frame j
Non-negative Matrix Factorization (NMF)

Figure 1. Time-frequency-microphone representation (left) and its slices (right) with corresponding audio source separation methods.

2. PROBLEM FORMULATION

2.1 Mixing Process

Let $\mathrm{s}_1, \ldots, \mathrm{s}_N$ be N source signals and $\mathrm{x}_1, \ldots, \mathrm{x}_M$ be M microphone observations. The convolutive mixture model is formulated as

$$\mathrm{x}_m(t) = \sum_{n=1}^{N} \mathrm{h}_{mn} * \mathrm{s}_n(t)\,, \quad m = 1, \ldots, M\,, \tag{1}$$

where t represents time, and h_{mn} represents the impulse response from source s_n to microphone x_m. If we apply a short-time Fourier transform (STFT) with a sufficiently long window to cover the main part of the impulse responses, the convolutive mixture model (1) can be approximated as an instantaneous mixture model:[18, 19]

$$x_{ijm} = \sum_{n=1}^{N} h_{imn} s_{ijn}\,, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,\ m = 1, \ldots, M\,, \tag{2}$$

where i and j represent the frequency-bin index and time frame index, respectively. The number I of frequency bins is defined according to the sampling frequency and the STFT window length. The number J of time frames is defined according to the amount of STFT shift and the length of the microphone observation signals.

In this way, we obtain the **time-frequency representations**, Eq. (2), of the microphone observations. They can be compactly represented with an $I \times J \times M$ complex-valued tensor $\mathbf{X}$, $[\mathbf{X}]_{ijm} = x_{ijm}$, as the left side of Fig. 1 shows. Here we define some vector notation for conciseness: $\mathbf{x}_{ij} = [x_{ij1}, \ldots, x_{ijM}]^T \in \mathbb{C}^M$ for M microphone observations, and $\mathbf{h}_{in} = [h_{i1n}, \ldots, h_{iMn}]^T \in \mathbb{C}^M$ for the frequency response from source n to all the microphones at frequency bin i. Then, Eq. (2) becomes

$$\mathbf{x}_{ij} = \sum_{n=1}^{N} \mathbf{h}_{in} s_{ijn}\,, \quad i = 1, \ldots, I,\ j = 1, \ldots, J\,. \tag{3}$$

Further author information:
E-mail: sawada.hiroshi@lab.ntt.co.jp, Telephone: +81-774-93-5272

2.2 Source Separation

The purpose of audio source separation is to obtain N separated signals $\mathrm{y}_m^{(n)}, n = 1, \dots, N$ that correspond to N source signal components observed at a microphone m. In this paper, the separation process is conducted for a time-frequency representation to obtain y_{ijmn}, which is a separated signal component for source n at frequency bin i, time frame j, and microphone m. By applying inverse STFT to y_{ijmn} for all i, j, we obtain a time domain separated signal $\mathrm{y}_m^{(n)}(t)$. Let us define a vector $\mathbf{y}_{ij}^{(n)} = [y_{ij1n}, \dots, y_{ijMn}]^T \in \mathbb{C}^M$ again for conciseness.

3. METHODS BASED ON FREQUENCY BIN-WISE SEPARATION

This section describes source separation methods based on frequency bin-wise operations. As shown in the top half of Fig. 1, in such methods, we obtain an $M \times J$ matrix $\mathbf{X}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{iJ}]$ for each frequency bin i by slicing the observation tensor $\mathbf{X}$ of a time-frequency-microphone representation.

3.1 Linear Filters

First, we describe the source separation method that involves constructing linear filters in a frequency bin-wise manner. If we obtain prior knowledge about the locations of sources or the time periods when a particular source is absent, beamforming techniques can be utilized. However, if such prior knowledge is unavailable, we generally employ ICA[3,4] to construct the linear filters.

ICA obtains separated signals $\breve{\mathbf{y}}_{ij} = [\breve{y}_{ij1}, \dots, \breve{y}_{ijN}]^T \in \mathbb{C}^N$ by a linear transformation

$$\breve{\mathbf{y}}_{ij} = \mathbf{W}_i \mathbf{x}_{ij}\,, \;\; j = 1, \dots, J \tag{4}$$

with a separation matrix $\mathbf{W}_i$ of size $N \times M$. The matrix $\mathbf{W}_i$ is optimized so that the distribution of the vector elements $\breve{y}_{ijn}$, $j = 1, \dots, J$ is far from a Gaussian distribution. Various optimization learning rules have been proposed, for example FastICA[3] or one based on natural gradient.[4,20] In the learning rule, a non-linear function is utilized to evaluate how similar or different the distribution of each element $\breve{y}_{ijn}$ is compared with the Gaussian distribution. For a complex-valued variable obtained as a result of STFT, a polar coordinate based non-linear function[21] is effective.

There is scaling ambiguity in an ICA solution. For an audio source separation task, the scaling ambiguity is resolved by trying to represent the observed signals at microphones with scaled separated signals. For that purpose, we calculate the inverse matrix $\mathbf{A}_i = [\mathbf{a}_{i1}, \dots, \mathbf{a}_{iN}] = \mathbf{W}_i^{-1}$ or the Moore-Penrose pseudo inverse matrix $\mathbf{A}_i = [\mathbf{a}_{i1}, \dots, \mathbf{a}_{iN}] = \mathbf{W}_i^{+}$ of the separated matrix. And by multiplying it on both sides of (4), we have

$$\mathbf{A}_i \breve{\mathbf{y}}_{ij} = \sum_{n=1}^{N} \mathbf{a}_{in} \breve{y}_{ijn} = \mathbf{x}_{ij}\,. \tag{5}$$

Then we obtain a vector $\mathbf{y}_{ij}^{(n)}$ of scaled (ambiguity-resolved) separated signals as

$$\mathbf{y}_{ij}^{(n)} = \mathbf{a}_{in} \breve{y}_{ijn}\,. \tag{6}$$

3.2 Time-Frequency Masking

Second, we describe a source separation method that employs time-frequency masking, in which sparseness is assumed for the frequency-domain sources s_{ijn}.[22,23] The sparseness of a source can be characterized by the fact that the source amplitude is rarely large and is close to zero most of the time. Frequency-domain speech or music sources are good examples of sparse sources. When such sources are mixed, we can assume that at most only one source signal makes a large contribution to each time-frequency observation. Thus, the mixture model (2) can be further approximated as

$$x_{ijm} \approx h_{im\bar{n}} s_{ij\bar{n}} \tag{7}$$

where $\bar{n} = \bar{n}(i, j) \in \{1, \dots, N\}$ is the index of the source that has the largest amplitude, and depends on each time-frequency slot (i, j).

In the method based on time-frequency masking, we need to estimate which source has the largest amplitude for each time frequency slot (i, j). For that purpose, we employ a clustering method for observation vectors $\mathbf{x}_{ij}$ to calculate the posterior probability $P(C_n|\mathbf{x}_{ij})$ that a vector $\mathbf{x}_{ij}$ belongs to a cluster C_n. Then, time frequency masks are made by

$$\mathcal{M}_{ijn} = \begin{cases} 1 & p(C_n|\mathbf{x}_{ij}) \geq p(C_{n'}|\mathbf{x}_{ij}), \ {}^{\forall}n' \neq n \\ 0 & otherwise, \end{cases} \tag{8}$$

and separated signals are obtained by

$$\mathbf{y}_{ij}^{(n)} = \mathcal{M}_{ijn}\mathbf{x}_{ij} \, . \tag{9}$$

In terms of clustering methods, one based on an anechoic propagation model[24, 25] is easy and simple, and works well under low reverberant conditions. However, to cope with more complicated real-room sound propagation, frequency bin-wise clustering has been proposed.[10] Specifically, a Gaussian mixture model (GMM)[5]

$$p(\mathbf{x}_{ij}|\theta) = \sum_{n=1}^{N} \alpha_{in}\, p(\mathbf{x}_{ij}|\mathbf{a}_{in}, \sigma_{in}) \tag{10}$$

with a complex Gaussian density function of the form[10]

$$p(\mathbf{x}|\mathbf{a}_{in}, \sigma_{in}) = \frac{1}{(\pi\sigma_{in}^2)^{M-1}} \exp\left(-\frac{||\mathbf{x} - (\mathbf{a}_{in}^H\mathbf{x}) \cdot \mathbf{a}_{in}||^2}{\sigma_{in}^2}\right) \tag{11}$$

is assumed for each frequency bin i, and we estimate a parameter set $\theta = \{\mathbf{a}_{i1}, \sigma_{i1}, \alpha_{i1}, \ldots, \mathbf{a}_{iN}, \sigma_{iN}, \alpha_{iN}\}$ that maximizes the likelihood $p(\mathbf{X}_i|\theta) = \prod_{j=1}^{J} p(\mathbf{x}_{ij}|\theta)$ of the matrix $\mathbf{X}_i$. In (10), $\mathbf{a}_{in}$ is the mean vector, σ_{in} is the variance, and α_{in} is the mixture ratio of the n-th cluster. After the parameter set is estimated, the posterior probabilities used in (8) is given by

$$p(C_n|\mathbf{x}_{ij}) = \frac{\alpha_{in}p(\mathbf{x}_{ij}|\mathbf{a}_{in}, \sigma_{in})}{\sum_{n=1}^{N} \alpha_{in}p(\mathbf{x}_{ij}|\mathbf{a}_{in}, \sigma_{in})} \, . \tag{12}$$

3.3 Post-processing for Permutation Alignment

The method based on ICA or GMM, described in Subsections 3.1 or 3.2, performs a source separation task in a frequency bin-wise manner. Therefore, we need to align the permutation ambiguity of the ICA or GMM results in each frequency bin so that a separated signal in the time domain contains frequency components from the same source signal. This problem is well known as the permutation problem of frequency-domain BSS.[8] Although various approaches to the permutation problem have been proposed,[26] the following approach based on dominance measures[9, 10] performs very well.

When using ICA to construct linear filters, we employ the power ratio

$$r_i^{(n)}(j) = \frac{||\mathbf{y}_{ij}^{(n)}||^2}{\sum_{n=1}^{N} ||\mathbf{y}_{ij}^{(n)}||^2} \tag{13}$$

of scaled separated signals (6) as a dominance measure.[9] On the other hand, when using a GMM for time-frequency masking, we employ the posterior probability (12)

$$r_i^{(n)}(j) = p(C_n|\mathbf{x}_{ij}) \tag{14}$$

as a dominance measure.[10] After calculating the dominance measure, we basically interchange the indices n of separated signals so that the correlation coefficient $\rho(r_i^{(n)}, r_{i'}^{(n)})$ between the dominance measures at different frequency bins i and i' is maximized for the same source. The optimization procedure is described in detail in a reference.[10]

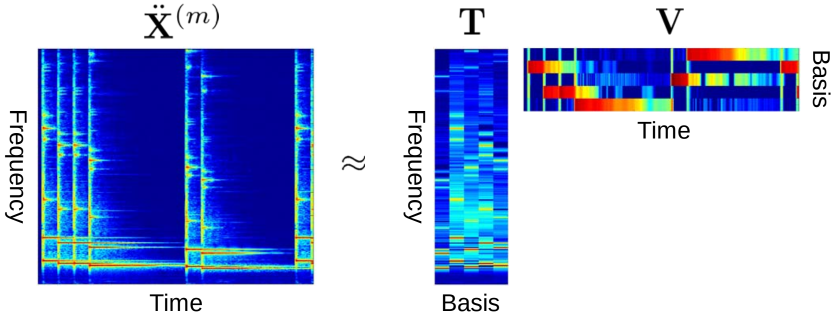

Figure 2. Frequent sound patterns are extracted by NMF as its bases.

3.4 Simultaneous Permutation Alignment

The last subsection describes a way to solve the permutation problem by post-processing after ICA or GMM. This subsection, on the other hand, introduces some methods in which the permutation problem is solved simultaneously with the source separation procedure.

Regarding the linear filters (i.e., ICA) described in Subsection 3.1, independent vector analysis (IVA)[11–13] is the corresponding method. With IVA, basically the same linear filter is constructed as with ICA (4). However, the learning rule for the separation matrix $\mathbf{W}_i$ is different from that of ICA, and the non-linear function used in the rule depends on the separated signals of all the frequency bins $\breve{y}_{1jn}, \ldots, \breve{y}_{Ijn}$.

Regarding the time-frequency masking described in Subsection 3.2, methods with unified statistical models were recently proposed.[14,15] More specifically, in addition to frequency bin-wise clustering based on a GMM, those methods employ other statistical models for the source signal activities and/or the direction of arrival of sources, and have derived inference algorithms that optimize all the model parameters simultaneously.

4. METHOD BASED ON FREQUENT SOUND PATTERN EXTRACTION

This section describes a source separation method based on frequent sound pattern extraction. More specifically, it is based on non-negative matrix factorization (NMF),[6,7] as depicted in the bottom half of Fig. 1.

4.1 Standard Single-channel NMF

This subsection reviews standard single-channel NMF. Let the observation tensor $\mathbf{X}$ be sliced to obtain an $I \times J$ matrix $\mathbf{X}^{(m)}$, $[\mathbf{X}^{(m)}]_{ij} = x_{ijm}$, for each microphone m.

There are several distance/divergence metrics proposed for NMF. In this paper, we describe NMF with Itakura-Saito (IS) divergence, IS-NMF,[27] which is recognized as being highly suited to audio modeling and separation. The NMF input should be a matrix with non-negative elements. Thus, we calculate the squared absolute value

$$\ddot{x}_{ijm} = |x_{ijm}|^2 = x_{ijm} x^*_{ijm} \tag{15}$$

of x_{ijm}, and construct a non-negative matrix $\ddot{\mathbf{X}}^{(m)}$, $[\ddot{\mathbf{X}}^{(m)}]_{ijm} = \ddot{x}_{ijm}$. Then, we factorize this $I \times J$ matrix into the product

$$\ddot{\mathbf{X}}^{(m)} \approx \mathbf{TV} \tag{16}$$

of an $I \times K$ matrix $\mathbf{T}$ and a $K \times J$ matrix $\mathbf{V}$. Here, K is the number of NMF bases. The matrix elements $t_{ik} = [\mathbf{T}]_{ik}$, $v_{kj} = [\mathbf{V}]_{kj}$ of the factored form should also be non-negative. Figure 2 shows an example of NMF applied to an audio signal. Five sound patterns are extracted as five NMF bases shown in matrix $\mathbf{T}$. Matrix $\mathbf{V}$ indicates when each basis becomes active with its intensity.

Algorithms for NMF minimize the divergence between the given matrix $\ddot{\mathbf{X}}^{(m)}$ and its factored form $\mathbf{TV}$. Let us define

$$\hat{x}_{ij} = \sum_{k=1}^{K} t_{ik} v_{kj} \tag{17}$$

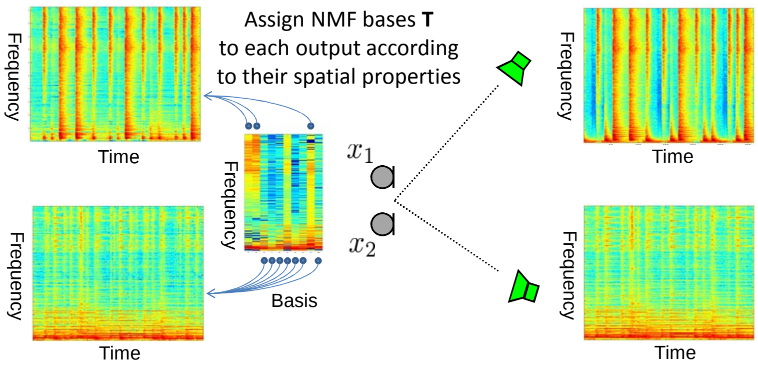

Figure 3. Multichannel extension of NMF achieves audio source separation by clustering NMF bases according to their spatial properties.

as the element that should be equal to $\ddot{x}_{ijm}$. Then, the IS divergence between $\ddot{\mathbf{X}}^{(m)}$ and $\mathbf{TV}$ is defined as

$$D_{IS}(\ddot{\mathbf{X}}^{(m)}, \{\mathbf{T}, \mathbf{V}\}) = \sum_{i=1}^{I}\sum_{j=1}^{J} d_{IS}(\ddot{x}_{ijm}, \hat{x}_{ij})\,, \tag{18}$$

$$d_{IS}(\ddot{x}_{ijm}, \hat{x}_{ij}) = \frac{\ddot{x}_{ijm}}{\hat{x}_{ij}} - \log\frac{\ddot{x}_{ijm}}{\hat{x}_{ij}} - 1\,. \tag{19}$$

The IS divergence defined by (18) can be minimized by the following manner. First, we initialize the elements of matrices $\mathbf{T}$ and $\mathbf{V}$ with random non-negative numbers. Then, we apply the following update rules until convergence:[28]

$$t_{ik} \leftarrow t_{ik}\sqrt{\frac{\sum_j \frac{\ddot{x}_{ijm}}{\hat{x}_{ij}}\frac{v_{kj}}{\hat{x}_{ij}}}{\sum_j \frac{v_{kj}}{\hat{x}_{ij}}}}\,, \tag{20}$$

$$v_{kj} \leftarrow v_{kj}\sqrt{\frac{\sum_i \frac{\ddot{x}_{ijm}}{\hat{x}_{ij}}\frac{t_{ik}}{\hat{x}_{ij}}}{\sum_i \frac{t_{ik}}{\hat{x}_{ij}}}}\,. \tag{21}$$

These update rules are called *multiplicative*, since another non-negative value is multiplied with the element before the update.

For the multichannel extension described in the next section, here we explain a generative model related to the formulation of IS-NMF. Suppose that each of the elements x_{ijm} of the observation $\mathbf{X}^{(m)}$ is generated from a zero-mean complex Gaussian distribution $\mathcal{N}_c(x_{ijm}|0, \hat{x}_{ij})$ with variance $\hat{x}_{ij}$. Then, the maximization of the log-likelihood of the observation

$$\log p(\mathbf{X}^{(m)}|\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} \log \mathcal{N}_c(x_{ijm}|0, \hat{x}_{ij}) \tag{22}$$

is equivalent to the minimization of the IS divergence defined by (18) with $\theta = \{\mathbf{T}, \mathbf{V}\}$.[17]

4.2 Multi-channel NMF (without Basis Clustering)

This subsection and the next discuss multichannel extensions of NMF. The purpose of the extension is to cluster NMF bases and to obtain source separation results as shown in Fig. 3. We begin with the simplest form where each NMF basis is attached to its own spatial property. In the next subsection, we will cluster the bases by sharing spatial properties.

For the multichannel extension of IS-NMF, we extend the generative model explained in the last subsection to a multichannel case. Suppose that each observation vector $\mathbf{x}_{ij} = [x_{ij1}, \ldots, x_{ijM}]^T \in \mathbb{C}^M$ for M microphones is generated from a multivariate complex Gaussian distribution $\mathcal{N}_{\mathbf{c}}(\mathbf{x}_{ij}|\mathbf{0}, \hat{\mathsf{X}}_{ij})$ with mean $\mathbf{0}$ and covariance matrix $\hat{\mathsf{X}}_{ij}$. Then, as with (22), we consider the log-likelihood

$$\log p(\mathbf{X}|\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} \log \mathcal{N}_{\mathbf{c}}(\mathbf{x}_{ij}|\mathbf{0}, \hat{\mathsf{X}}_{ij}) \tag{23}$$

of the time-frequency-microphone tensor $\mathbf{X}$. Now, let us consider a factored form, similar to (17), for this multichannel extension.

As Fig. 3 shows, we associate spatial properties with the NMF bases $\mathbf{T}$. More specifically, we introduce an $M \times M$ matrix H_{ik} for an element t_{ik} of $\mathbf{T}$. Then, the covariance matrix $\hat{\mathsf{X}}_{ij}$ is modeled as

$$\hat{\mathsf{X}}_{ij} = \sum_{k=1}^{K} \mathsf{H}_{ik} t_{ik} v_{kj} \,. \tag{24}$$

Matrix H_{ik} should be Hermitian positive semidefinite to guarantee non-negativity in a multichannel sense. The maximization of the log-likelihood (23) is equivalent to the minimization of the multichannel IS divergence

$$D_{IS}(\ddot{\mathbf{X}}, \{\mathbf{T}, \mathbf{V}, \mathbf{H}\}) = \sum_{i=1}^{I}\sum_{j=1}^{J} d_{IS}(\mathsf{X}_{ij}, \hat{\mathsf{X}}_{ij})\,, \tag{25}$$

$$d_{IS}(\mathsf{X}_{ij}, \hat{\mathsf{X}}_{ij}) = \mathrm{tr}(\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}) - \log\det \mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1} - M \tag{26}$$

where $\mathrm{tr}(\mathsf{B}) = \sum_{m=1}^{M} b_{mm}$ is the trace of a matrix B, det is the determinant, and $\mathsf{X}_{ij} = \mathbf{x}_{ij}\mathbf{x}_{ij}^{H}$ is the outer product of the observation vector. And $\ddot{\mathbf{X}}$ in (25) is a hierarchical matrix that has such outer products as its elements $[\ddot{\mathbf{X}}]_{ij} = \mathsf{X}_{ij}$.

The divergence defined by (25) and (26) is minimized by repeating the following update rules.[17]

$$t_{ik} \leftarrow t_{ik}\sqrt{\frac{\sum_j v_{kj}\mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{ik})}{\sum_j v_{kj}\mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{ik})}} \tag{27}$$

$$v_{kj} \leftarrow v_{kj}\sqrt{\frac{\sum_i t_{ik}\mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{ik})}{\sum_i t_{ik}\mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{ik})}} \tag{28}$$

These update rules reduce to those (20)-(21) of the single-channel counterpart if we assume $M = 1$, $\mathsf{X}_{ij} = \ddot{x}_{ijm}$, $\mathsf{H}_{ik} = 1$. For updating H_{ik}, the algebraic Riccati equation

$$\mathsf{H}_{ik}^{H}\mathsf{A}\mathsf{H}_{ik} = \mathsf{B} \tag{29}$$

is solved with

$$\mathsf{A} = \sum_{j} v_{kj}\hat{\mathsf{X}}_{ij}^{-1} \tag{30}$$

$$\mathsf{B} = \mathsf{H}'_{ik}\left(\sum_{j} v_{kj}\hat{\mathsf{X}}_{ij}^{-1}\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}\right)\mathsf{H}'_{ik} \tag{31}$$

where H'_{ik} is the matrix before the update.

4.3 Multi-channel NMF (with Basis Clustering)

This subsection modifies the formulation of the multichannel NMF in order to cluster K matrices $\mathsf{H}_{i1}, \ldots, \mathsf{H}_{iK}$ into N classes. Let us introduce a new matrix $\mathbf{Z}$ of size $N \times K$. The elements z_{nk} indicate whether the k-th matrix belongs to the n-th class ($z_{nk} = 1$) or not ($z_{nk} = 0$). Replacing H_{ik} in (24) with $\sum_{n=1}^{N} \mathsf{H}_{in} z_{nk}$, we have

$$\hat{\mathsf{X}}_{ij} = \sum_{k=1}^{K} \left(\sum_{n=1}^{N} \mathsf{H}_{in} z_{nk} \right) t_{ik} v_{kj} \,. \tag{32}$$

The definition of the multichannel IS divergence (26) is still valid even if we change the formulation of $\hat{\mathsf{X}}_{ij}$ as above.

Now, to optimize the newly introduced matrix $\mathbf{Z}$, $[\mathbf{Z}]_{nk} = z_{nk}$, in the same manner as that employed for matrices $\mathbf{T}$ and $\mathbf{V}$, let us redefine z_{nk} as continuous values to satisfy $z_{nk} \geq 0$ and $\sum_{n=1}^{N} z_{nk} = 1$. We consider this redefinition corresponds to estimating the expectation of z_{nk} according to the posterior probability $p(z_{nk} = 1 | \mathbf{X}, \mathbf{T}, \mathbf{V}, \mathbf{H})$ instead of the value z_{nk} itself.

The multichannel NMF after replacing (24) with (32) is performed by the following update rules:[17]

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_n z_{nk} \sum_j v_{kj} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{in})}{\sum_n z_{nk} \sum_j v_{kj} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{in})}} \tag{33}$$

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_n z_{nk} \sum_i t_{ik} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{in})}{\sum_n z_{nk} \sum_i t_{ik} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{in})}} \tag{34}$$

$$z_{nk} \leftarrow z_{nk} \sqrt{\frac{\sum_{i,j} t_{ik} v_{kj} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{in})}{\sum_{i,j} t_{ik} v_{kj} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{in})}} \tag{35}$$

For updating H_{in}, the algebraic Riccati equation

$$\mathsf{H}_{in}^{H} \mathsf{A} \mathsf{H}_{in} = \mathsf{B} \tag{36}$$

is solved with

$$\mathsf{A} = \sum_{k} z_{nk} t_{ik} \sum_{j} v_{kj} \hat{\mathsf{X}}_{ij}^{-1} \tag{37}$$

$$\mathsf{B} = \mathsf{H}_{in} \left(\sum_{k} z_{nk} t_{ik} \sum_{j} v_{kj} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \right) \mathsf{H}_{in} \,. \tag{38}$$

For z_{nk}, unit-norm normalization $z_{nk} \leftarrow z_{nk} \,/\, (\sum_n z_{nk})$ should follow (35) to satisfy the constraint $\sum_{n=1}^{N} z_{nk} = 1$.

Source separation results are obtained with the multichannel Wiener filter

$$\mathbf{y}_{ij}^{(n)} = \left(\sum_{k=1}^{K} z_{nk} t_{ik} v_{kj} \right) \mathsf{H}_{in} \hat{\mathsf{X}}_{ij}^{-1} \mathbf{x}_{ij} \tag{39}$$

with $\hat{\mathsf{X}}_{ij}$ being defined in (32), after $\mathbf{T}$, $\mathbf{V}$, $\mathbf{H}$, $\mathbf{Z}$ are optimized by the above update rules.

5. CONCLUSION

In this paper, we tried to provide a unified view of audio source separation with multiple microphones. Regarding experimental results related to the described methods, please refer to the corresponding references. In general, the methods in Section 3 are good at separating speech sources, whereas the methods in Section 4 are good at separating music sources. The multichannel NMF described in Subsections 4.2 and 4.3 is a more recent research result, and still requires work on, for example, reducing the computational time. We hope that many researchers are interested in audio source separation and will develop many effective methods.

REFERENCES

[1] Makino, S., Lee, T.-W., and Sawada, H., eds., [*Blind Speech Separation*], Springer (2007).

[2] Pedersen, M. S., Larsen, J., Kjems, U., and Parra, L. C., "A survey of convolutive blind source separation methods," *Multichannel Speech Processing Handbook* (2007).

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

[4] Cichocki, A. and Amari, S., [*Adaptive Blind Signal and Image Processing*], John Wiley & Sons (2002).

[5] Bishop, C. M., [*Pattern Recognition and Machine Learning*], Springer (2006).

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

[7] Cichocki, A., Zdunek, R., Phan, A., and Amari, S., [*Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*], Wiley (2009).

[8] Sawada, H., Mukai, R., Araki, S., and Makino, S., "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Processing* **12**, 530–538 (Sept. 2004).

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

[10] Sawada, H., Araki, S., and Makino, S., "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, and Language Processing* **19**, 516–527 (Mar. 2011).

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

[12] Kim, T., Attias, H. T., Lee, S.-Y., and Lee, T.-W., "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech and Language Processing* , 70–79 (Jan. 2007).

[13] Ono, N., "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in [*Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*], 189–192 (Oct. 2011).

[14] Araki, S., Nakatani, T., and Sawada, H., "Simultaneous clustering of mixing and spectral model parameters for blind sparse source separation," in [*Proc. ICASSP 2010*], 5–8 (2010).

[15] Otsuka, T., Ishiguro, K., Sawada, H., and Okuno, H., "Bayesian unification of sound source localization and separation with permutation resolution," in [*Twenty-Sixth AAAI Conference on Artificial Intelligence*], 2038–2045 (2012).

[16] Ozerov, A. and Févotte, C., "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech and Language Processing* **18**, 550–563 (Mar. 2010).

[17] Sawada, H., Kameoka, H., Araki, S., and Ueda, N., "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, and Language Processing* **21**, 971–982 (May 2013).

[18] Smaragdis, P., "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing* **22**, 21–34 (1998).

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

[20] Amari, S., "Natural gradient works efficiently in learning," *Neural Computation* **10**(2), 251–276 (1998).

[21] Sawada, H., Mukai, R., Araki, S., and Makino, S., "Polar coordinate based nonlinear function for frequency domain blind source separation," *IEICE Trans. Fundamentals* **E86-A**, 590–596 (Mar. 2003).

[22] Aoki, M., Okamoto, M., Aoki, S., Matsui, H., Sakurai, T., and Kaneda, Y., "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology* **22**(2), 149–157 (2001).

[23] Yilmaz, O. and Rickard, S., "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing* **52**, 1830–1847 (July 2004).

[24] Araki, S., Sawada, H., Mukai, R., and Makino, S., "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.* **87**(8), 1833–1847 (2007).

[25] Sawada, H., Araki, S., Mukai, R., and Makino, S., "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech, and Language Processing* **15**, 1592–1604 (July 2007).

[26] Sawada, H., Makino, S., and Araki, S., "Frequency-domain blind source separation," in [*Blind Speech Separation*], Makino, S., Lee, T.-W., and Sawada, H., eds., 47–78, Springer (2007).

[27] Févotte, C., Bertin, N., and Durrieu, J.-L., "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation* **21**(3), 793–830 (2009).

[28] Nakano, M., Kameoka, H., Roux, J. L., Kitano, Y., Ono, N., and Sagayama, S., "Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence," in [*Proc. MLSP 2010*], 283–288 (Aug. 2010).