A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden Markov models

Takuya Higuchi[1]*, Hirofumi Takeda*[1]*, Tomohiko Nakamura*[1] *and Hirokazu Kameoka*[1,2]

[1]Graduate School of Information Science and Technology, The University of Tokyo
[2]NTT Communication Science Laboratories, NTT Corporation
{higuchi, h-takeda, nakamura, kameoka}@hil.t.u-tokyo.ac.jp

Abstract

This paper proposes to introduce a new model called "the multichannel factorial hidden Markov Model (MFHMM)" for underdetermined blind signal separation (BSS). For monaural source separation, one successful approach involves applying non-negative matrix factorization (NMF) to the magnitude spectrogram of a mixture signal, interpreted as a non-negative matrix. Up to now, multichannel extensions of NMF, which allow for the use of spatial information as an additional clue for source separation, have been proposed by several authors and proven to be an effective approach for underdetermined BSS. This approach is based on the assumption that an observed signal is a mixture of a limited number of source signals each of which has a static power spectral density scaled by a time-varying amplitude. However, many source signals in real world are non-stationary in nature and the variations of the spectral densities are much richer in time. Moreover, many sources including speech tend to stay inactive for some while until they switch to an active mode, implying that the total power of a source may depend on its underlying state. To reasonably characterize such a non-stationary nature of source signals, this paper proposes to extend the multichannel NMF model by modeling the transition of the set consisting of the spectral densities and the total power of each source using a hidden Markov model (HMM). By letting each HMM contain states corresponding to active and inactive modes, we will show that voice activity detection and source separation can be solved simultaneously through parameter inference of the present model. The experiment showed that the proposed algorithm provided a 7.65 dB improvement compared with the conventional multichannel NMF in terms of the signal-to-distortion ratio.

Index Terms: blind signal separation, source activity detection, a hidden Markov model, non-negative matrix factorization

1. Introduction

Blind source separation (BSS) refers to a technique for separating out individual source signals from microphone array inputs when the transfer characteristics between the sources and microphones are unknown. The best-known commercial application of BSS techniques is their use in teleconferencing systems. To solve BSS problems, it is generally necessary to make some assumptions about the 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] by assuming that the sources are statistically independent of each other. However, in an underdetermined case, the independence assumption is too weak to allow us to determine a unique solution and so directly applying ICA will not work well.

For monaural source separation, one successful approach involves applying non-negative matrix factorization (NMF) to the magnitude spectrogram of a mixture signal, interpreted as a non-negative matrix [2, 3]. With this approach, the spectrogram of a mixture signal is factorized into the product of a basis matrix consisting of basis spectra and an activation matrix consisting of time-varying amplitudes associated with the basis spectra. An important feature of this approach is that it is capable of finding a finite set of basis spectra that are considered to be the dominant elements constituting the observed spectrogram, in an unsupervised manner. Up to now, several attempts have been made to extend this approach to a multichannel case in order to allow for the use of spatial information as an additional clue for separation, which have opened a door to a new promising approach for underdetermined BSS [4, 5]. This approach is based on the assumption that an observed signal is a mixture of a limited number of source signals each of which has a static power spectral density (i.e., the basis spectrum) scaled by a time-varying amplitude. However, many source signals in real world are non-stationary in nature and the variations of the spectral densities are much richer in time. For example, the sound of a piano note would be more accurately characterized by a succession of several basis spectra corresponding to "attack," "decay," "sustain" and "release" segments than only by a single basis spectrum. Another important fact is that many sources including speech tend to stay inactive for some while until they switch to an active mode. This implies that the total power of a source may depend on its underlying state. To reasonably characterize such a non-stationary nature of source signals, this paper proposes to extend the multichannel NMF model by modeling the transition of the set consisting of the spectral densities and the total power of each source using a hidden Markov model (HMM). With this model, we would be able to flexibly reflect the time-varying nature of each source by appropriately specifying or training the transition probabilities prior to analysis. We formulate an entire generative model of a multichannel mixture signal by incorporating a generative model of the spectral densities of each source, described using an HMM. Each hidden state of the HMM represents the state of the corresponding source. We call this model "the multichannel factorial hidden Markov model (MFHMM)".

In general, simultaneous estimation is preferable when several estimation problems are interdependent. If we knew when each source is active and inactive, source separation would become a relatively simple matter. On the other hand, if all the sources were already separated, voice activity detection would become a relatively simple matter. This simply implies that the problems of source separation and voice/source activity detection are interdependent of each other. It is important to note that through parameter inference of the present model, we would be able to simultaneously perform voice activity detection and source separation based on a unified maximum likelihood criterion.

The remainder of this paper is organized as follows. Sec. 2 formulates a generative model of a multichannel observed signal and source signals based on the factorial HMM, Sec. 3 describes a parameter inference algorithm for the present model and Sec. 4 presents some experimental results.

2. Multichannel factorial HMM

2.1. Mixing model

First we consider a situation where I source signals are recorded by M microphones. Here, let $y_m(\omega_k, t_l) \in \mathbb{C}$ be the short-time Fourier transform (STFT) component observed at the m-th microphone, and $s_i(\omega_k, t_l) \in \mathbb{C}$ be the STFT component of the

i-th source. $1 \leq k \leq K$ and $1 \leq l \leq L$ are the frequency and time indices, respectively. If we assume that the length of the impulse response from a source to microphones is sufficiently shorter than the frame length of the STFT, the observed signal can be approximated fairly well by an instantaneous mixture in the frequency domain:

$$\boldsymbol{y}(\omega_k, t_l) = \sum_{i=1}^{I} \boldsymbol{a}_i(\omega_k) s_i(\omega_k, t_l), \tag{1}$$

where $\boldsymbol{y}(\omega_k, t_l) = (y_1(\omega_k, t_l), \ldots, y_M(\omega_k, t_l))^{\mathsf{T}} \in \mathbb{C}^M$. $\boldsymbol{a}_i(\omega_k)$ denotes the frequency array response for source i at frequency ω_k. For convenience of notation, we hereafter use subscripts k and l to indicate ω_k and t_l respectively.

2.2. Generative process of observed signals

Here we describe the generative process of an observed signal based on Eq. (1). If we assume that each source signal follows a piecewise stationary Gaussian process, then $s_{i,k,l}$ follows a complex normal distribution with mean 0 and covariance $\sigma^2_{i,k,l}$,

$$s_{i,k,l}|\sigma^2_{i,k,l} \sim \mathcal{N}_{\mathbb{C}}(s_{i,k,l}; 0, \sigma^2_{i,k,l}), \tag{2}$$

where $\sigma^2_{i,k,l}$ denotes the power spectral density of i-th source at frequency k and time l. From Eq. (1) and Eq. (2), $\boldsymbol{y}_{k,l}$ is also normally distributed such that

$$\boldsymbol{y}_{k,l}|\boldsymbol{a}_{1:I,k}, \sigma_{1:I,k,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; 0, \sum_i \boldsymbol{C}_{i,k}\sigma^2_{i,k,l}), \tag{3}$$

conditioned on $\boldsymbol{a}_{1:I,k} = \{\boldsymbol{a}_{1,k}, \ldots, \boldsymbol{a}_{I,k}\}$, $\sigma^2_{1:I,k,l} = \{\sigma^2_{1,k,l}, \ldots, \sigma^2_{I,k,l}\}$ where $\boldsymbol{C}_{i,k} = \boldsymbol{a}_{i,k}\boldsymbol{a}^{\mathsf{H}}_{i,k}$ and $\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \propto \exp(-(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{H}}\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))$. $\boldsymbol{C}_{i,k}$ represents a spatial correlation matrix of the i-th source at frequency k.

2.3. Generative model for multichannel NMF [4, 5]

In the regular NMF model (see [6]), the power spectra of a source signal is assumed to be static up to a scale factor. We can incorporate this assumption into the above model by setting

$$\sigma^2_{i,k,l} = w_{i,k}h_{i,l}, \tag{4}$$

where $\sigma^2_{i,k,l}$ is assumed to be factorized into the product of the static power spectrum $w_{i,k}$ and the time-varying amplitude $h_{i,l}$. The generative model of $s_{i,k,l}$ is thus rewritten as

$$\boldsymbol{s}_{i,k,l}|w_{i,k}, h_{i,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_{i,k,l}; 0, w_{i,k}h_{i,l}), \tag{5}$$

conditioned on $w_{i,k}$ and $h_{i,l}$. Since the generative model of $\boldsymbol{Y} = \{\boldsymbol{y}_{k,l}\}_{k,l}$ under the assumption Eq. (5) can be viewed as a natural extension of NMF to a multichannel case, a BSS approach based on this model is called the multichannel NMF [4, 5].

2.4. Generative modeling of source signals using HMMs

As described above, the multichannel NMF model roughly assumes that the power spectra of each sound source is static up to a scale factor. However, many sound sources exhibit different spectra according to underlying "states" of the sources. For example, the spectra of the sound of a piano note would be different in "attack," "decay," "sustain" and "release" segments. Another important fact is that many sources including speech tend to stay inactive for some while until they switch to an active mode. This implies that the total power of a source also depends on its underlying state. To reasonably characterize such a non-stationary nature of source signals, here we propose to model the sequence of the power spectra and the total powers of each source using an HMM.

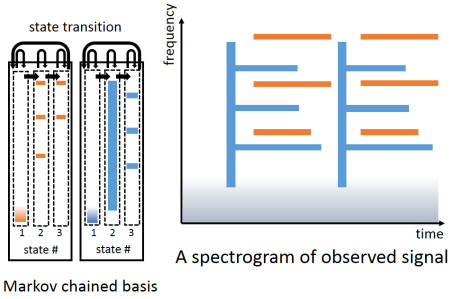

Figure 1: Illustration of the concept of the FHMM.

Now we introduce a latent variable $z_{i,l} \in \{1, \ldots, D\}$ to denote a state of the i-th source at time l. The state sequence $z_{i,1}, \ldots, z_{i,L}$ is assumed to follow a Markov chain:

$$z_{i,l}|z_{i,l-1} \sim \mathrm{Categorical}(z_{i,l}; \boldsymbol{\rho}_{z_{i,l-1}}), \tag{6}$$

where $\mathrm{Categorical}(x; \boldsymbol{y}) = y_x$, $\boldsymbol{\rho}_d = (\rho_{d,1}, \ldots, \rho_{d,D})$ denotes the transition probability of state d to each state $1, \ldots, D$, and $\boldsymbol{\rho} = (\rho_{d,d'})_{D\times D}$ denotes the transition matrix. Here, we assume that $h_{i,l}$ follows a gamma distribution with hyperparameters determined according to $z_{i,l}$,

$$h_{i,l}|z_{i,l} \sim \mathrm{Gamma}(h_{i,l}; \alpha_{z_{i,l}}\beta_{z_{i,l}}), \tag{7}$$

where $\alpha_{1:D}$ and $\beta_{1:D}$ are the shape and scale parameters of a gamma distribution, and $\mathrm{Gamma}(x; \alpha, \beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$. As we want $h_{i,l}$ to take a small value when $z_{i,l}$ is the "inactive" (i.e., silent) state, we set the hyperparameters of the gamma distribution of that state so that it becomes a sparsity-inducing distribution. As regards the gamma distributions of the remaining states, we consider setting the hyperparameters so that they become uniform distributions. In particular, we consider setting α and β at 1 and 10^{-2} respectively for the inactive state, and at 1 and 10^{20} respectively for the active states. We expect that this setting allows us to solve source separation and voice activity detection in a cooperative manner. Let us use $w_{i,k,d}$ to denote the power spectrum of the i-th source at state d. The power spectrum of the i-th source at time l is also assumed to be determined according to $z_{i,l}$. Thus, the generative model of $s_{i,k,l}$ is eventually written as

$$\begin{aligned} &s_{i,k,l}|w_{i,k,1:D}, h_{i,l}, z_{i,l} \\ &\quad \sim \mathcal{N}_{\mathbb{C}}(s_{i,k,l}; 0, w_{i,k,z_{i,l}}h_{i,l}). \end{aligned} \tag{8}$$

Since the generative of $\boldsymbol{y}_{k,l}$ contains multiple HMMs associated with the underlying sources, the overall model can be viewed as a Factorial HMM. Fig. 1 shows an illustration of the proposed FHMM. Note that we can integrate an assumption that the signal tend to stay non-active for some time by setting the transition probabilities of self-transitions at reasonably large values in a specific state of the HMM. If we assume the emission probabilities of the HMMs as uniform distributions, our overall generative model is given by Eqs. (6) and

$$\begin{aligned} &\boldsymbol{y}_{k,l}|\boldsymbol{a}_{1:I,k}, w_{1:I,k,1:D}, h_{1:I,l}, z_{1:I,l} \\ &\quad \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; 0, \sum_i \boldsymbol{C}_{i,k}w_{i,k,z_{i,l}}h_{i,l}), \end{aligned} \tag{9}$$

conditioned on $\boldsymbol{a}_{1:I,k}$,$w_{1:I,k,1:D}$,$h_{1:I,l}$ and $z_{1:I,l}$.

3. Algorithm for parameter estimation

3.1. Objective function

In this section, we describe a parameter estimation algorithm for our generative model based on an auxiliary function method. The random variables of interest in our model are $\boldsymbol{W} = w_{1:I,1:K,1:D}$, $\boldsymbol{H} = h_{1:I,1:L}$,$\boldsymbol{C} = \boldsymbol{C}_{1:I,1:K}$ and $\boldsymbol{Z} =$

$z_{1:I,1:L}$. We denote the entire set of the above parameters as Θ. In the following, $\boldsymbol{\rho}$ is constants that is determined experimentally. Our goal is to compute the posterior

$$p(\Theta|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y},\Theta)}{p(\boldsymbol{Y})}, \tag{10}$$

where $\boldsymbol{Y} = \boldsymbol{y}_{1:K,1:L}$ is a set consisting of the time-frequency components of observed multichannel signals. By using the conditional distributions defined in Sec. 2, we can write the joint distribution $p(\boldsymbol{Y},\Theta)$ as

$$p(\boldsymbol{Y},\boldsymbol{\Theta}) \propto p(\boldsymbol{Y}|\boldsymbol{\Theta})p(\boldsymbol{H}|\boldsymbol{Z})p(\boldsymbol{Z}). \tag{11}$$

The objective function is defined as $L(\Theta) = \log p(\Theta|\boldsymbol{Y})$. Our goal is to obtain $\hat{\Theta}$ such that

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \log p(\Theta|\boldsymbol{Y}). \tag{12}$$

By using Eqs. (10), (11) and (12), the current optimization problem can be rewritten as

$$\begin{aligned}\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\big(&\log p(\boldsymbol{Y}|\boldsymbol{\Theta}) \\ &+ \log p(\boldsymbol{H}|\boldsymbol{Z}) + \log p(\boldsymbol{Z})\big).\end{aligned} \tag{13}$$

According to the generative model defined in Sec. 2, $\log p(\boldsymbol{Y}|\boldsymbol{\Theta})$ is written as

$$\begin{aligned}&\log p(\boldsymbol{Y}|\boldsymbol{\Theta}) \\ &= \sum_{k,l}(-\frac{M}{2}\log 2\pi - \frac{1}{2}\log|\hat{\boldsymbol{X}}_{k,l}| - \frac{1}{2}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\hat{\boldsymbol{X}}_{k,l}^{-1}\boldsymbol{y}_{k,l}),\end{aligned} \tag{14}$$

where $\hat{\boldsymbol{X}}_{k,l} = \sum_i \boldsymbol{C}_{i,k} w_{i,k,z_{i,l}} h_{i,l}$.

3.2. Optimization algorithm based on an auxiliary function method

The optimization problem of maximizing $L(\Theta)$ with respect to Θ is difficult to solve analytically. However, we can invoke an auxiliary function approach to derive an iterative algorithm that searches for the estimate of Θ, as with [5]. To apply an auxiliary function approach to the current optimization problem, the first step is to construct an auxiliary function $L^+(\Theta,\Lambda)$ satisfying $L(\Theta) = \max_\Lambda L^+(\Theta,\Lambda)$. We refer to Λ as an auxiliary variable. It can then be shown that $L(\Theta)$ is non-increasing under the updates $\Theta \leftarrow \operatorname{argmax}_\Theta L^+(\Theta,\Lambda)$ and $\Lambda \leftarrow \operatorname{argmax}_\Lambda L^+(\Theta,\Lambda)$. The proof of this shall be omitted owing to space limitations. Thus, $L^+(\Theta,\Lambda)$ should be designed as a function that can be maximized analytically with respect to Θ and Λ. Such a function can be constructed as follows.

$$\begin{aligned}&L(\boldsymbol{\Theta}) \\ &\geq L^+(\boldsymbol{\Theta},\Lambda) \\ &= -\frac{1}{2}\sum_{k,l}\Bigg(\sum_i\Bigg(\frac{\operatorname{tr}(\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{y}_{k,l}\boldsymbol{R}_{i,k,l}\boldsymbol{C}_{i,k}^{-1}\boldsymbol{R}_{i,k,l})}{w_{i,k,z_{i,l}}h_{i,l}} \\ &\quad + \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{C}_{i,k})w_{i,k,z_{i,l}}h_{i,l}\Bigg) + \log|\boldsymbol{U}_{k,l}| - M\Bigg) \\ &\quad + \sum_{i,l}\Big((\alpha_{z_{i,l}}-1)\log h_{i,l} - h_{i,l}/\beta_{z_{i,l}} - \alpha_{z_{i,l}}\log\beta_{z_{i,l}}\Big) \\ &\quad + \log p(\boldsymbol{Z}),\end{aligned} \tag{15}$$

where $\boldsymbol{R}_{i,k,l}$ and $\boldsymbol{U}_{k,l}$ are auxiliary variables that satisfy Hermitian positive definiteness and $\sum_i \boldsymbol{R}_{i,k,l} = \boldsymbol{I}$. We denote the set of the auxiliary variables as Λ. $\operatorname{tr}(\cdot)$ is the trace of a matrix. The equality $L(\Theta) = L^+(\Theta,\Lambda)$ is satisfied when

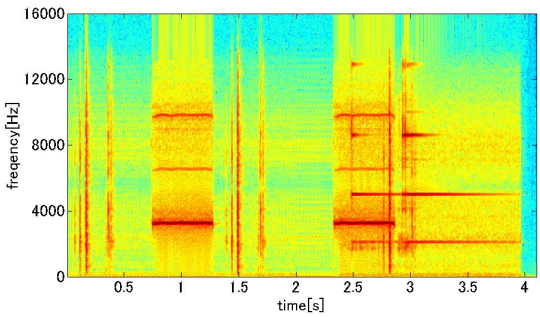

Figure 2: A spectrogram of the mixed signal.

$$\boldsymbol{R}_{i,k,l} = \boldsymbol{C}_{i,k} w_{i,k,z_{i,l}} h_{i,l} \hat{\boldsymbol{X}}_{k,l}^{-1}, \tag{16}$$

$$\boldsymbol{U}_{k,l} = \hat{\boldsymbol{X}}_{k,l}. \tag{17}$$

Therefore, we can monotonically increase L by repeating the following two steps.

1. Maximizing L^+ with respect to $\boldsymbol{R}$ and $\boldsymbol{U}$.
2. Maximizing L^+ with respect to $\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{C}$ and $\boldsymbol{Z}$.

Step 1 consists in updating $\boldsymbol{R}$ and $\boldsymbol{U}$ using Eqs. (16) and (17). In step 2, we can obtain update rules of $\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{C}$ by setting the partial derivative of L^+ with respect to each of the parameters at zero. The partial derivatives of L^+ with respect to $\boldsymbol{W}$ and $\boldsymbol{H}$ are given by

$$\begin{aligned}\frac{\partial L^+}{\partial w_{i,k,z_{i,l}}} = \sum_l\Bigg(&\frac{\operatorname{tr}(\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{y}_{k,l}\boldsymbol{R}_{i,k,l}\boldsymbol{C}_{i,k}^{-1}\boldsymbol{R}_{i,k,l})}{w_{i,k,z_{i,l}}^2 h_{i,l}} \\ &- \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{C}_{i,k})h_{i,l}\Bigg),\end{aligned} \tag{18}$$

$$\begin{aligned}\frac{\partial L^+}{\partial h_{i,l}} = \sum_k\Bigg(&\frac{\operatorname{tr}(\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{y}_{k,l}\boldsymbol{R}_{i,k,l}\boldsymbol{C}_{i,k}^{-1}\boldsymbol{R}_{i,k,l})}{w_{i,k,z_{i,l}} h_{i,l}^2} \\ &- \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{C}_{i,k})w_{i,k,z_{i,l}}\Bigg) \\ &+ (\alpha_{z_{i,l}}-1)/h_{i,l} - 1/\beta_{z_{i,l}},\end{aligned} \tag{19}$$

respectively. By setting them at zero, we obtain the following update rules:

$$w_{i,k,z_{i,l}} \leftarrow \sqrt{\frac{\sum_l \frac{\operatorname{tr}(\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{y}_{k,l}\boldsymbol{R}_{i,k,l}\boldsymbol{C}_{i,k}^{-1}\boldsymbol{R}_{i,k,l})}{h_{i,l}}}{\sum_l \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{C}_{i,k})h_{i,l}}}, \tag{20}$$

$$h_{i,l} \leftarrow \frac{(\alpha_{z_{i,l}}-1) + \sqrt{(\alpha_{z_{i,l}}-1) + 4\mu_{i,l}\nu_{i,l}}}{2\mu_{i,l}}, \tag{21}$$

where

$$\mu_{i,l} = \sum_k \frac{\operatorname{tr}(\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{y}_{k,l}\boldsymbol{R}_{i,k,l}\boldsymbol{C}_{i,k}^{-1}\boldsymbol{R}_{i,k,l})}{w_{i,k,z_{i,l}}}, \tag{22}$$

$$\nu_{i,l} = \sum_k \operatorname{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{C}_{i,k})w_{i,k,z_{i,l}} + 1/\beta_{z_{i,l}}. \tag{23}$$

The partial derivatives of L^+ with respect to $\boldsymbol{C}$ is given by

$$\frac{\partial L^+}{\partial \boldsymbol{C}_{i,k}} = \sum_l\Bigg(\frac{\boldsymbol{C}_{i,k}^{-1}\boldsymbol{R}_{i,k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{y}_{k,l}\boldsymbol{R}_{i,k,l}\boldsymbol{C}_{i,k}^{-1}}{w_{i,k,z_{i,l}}h_{i,l}}$$

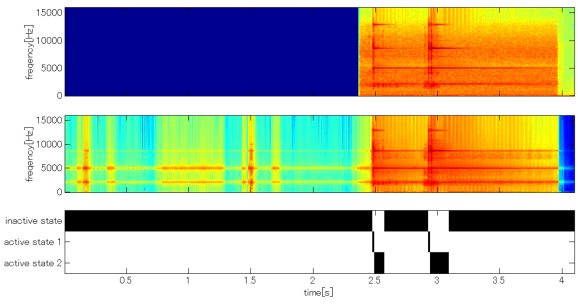

Figure 3: Examples of spectrogram of signal of a bell (top), that of separated signal (middle), and its activity detection result (bottom). Black indicates the state is assigned at the time.

Table 1: The average output SDRs and standard deviations of the three sources by the conventional and proposed methods.

SDR(±SD) [dB]	bell	whistle	stapler
Proposed	19.92(±11.50)	31.21(±6.69)	16.81(±10.09)
Conventional	13.33(±8.22)	23.37(±4.57)	8.28(±5.98)

$$- \boldsymbol{U}_{k,l}^{-1} w_{i,k,z_{i,l}} h_{i,l} \Big). \tag{24}$$

By setting this at zero, we obtain an algebraic Riccati equation:

$$\boldsymbol{C}_{i,k} \boldsymbol{A}_{i,k} \boldsymbol{C}_{i,k} = \boldsymbol{B}_{i,k}, \tag{25}$$

where

$$\boldsymbol{A}_{i,k} = \sum_{l} w_{i,k,z_{i,l}} h_{i,l} \hat{\boldsymbol{X}}_{k,l}^{-1}, \boldsymbol{B}_{i,k} \quad = \boldsymbol{C}_{i,k} (\sum_{l} w_{i,k,z_{i,l}} h_{i,l} \hat{\boldsymbol{X}}_{k,l}^{-1} \boldsymbol{y}_{k,l}{}^{\mathsf{H}} \boldsymbol{y}_{k,l} \hat{\boldsymbol{X}}_{k,l}^{-1}) \boldsymbol{C}_{i,k}. \tag{26}$$

We can solve this equation by using a method in [5]. We perform an eigenvalue decomposition of a $2M \times 2M$ matrix

$$\begin{bmatrix} 0 & -\boldsymbol{A}_{i,k} \\ -\boldsymbol{B}_{i,k} & 0 \end{bmatrix}, \tag{27}$$

and let $\boldsymbol{e}_{1,i,k} \dots \boldsymbol{e}_{M,i,k}$ be eigenvectors with negative eigenvalues. It is guaranteed that there are exactly M negative eigenvalues. Then, let us decompose the $2M$-dimensional eigenvectors as

$$\boldsymbol{e}_{m,i,k} = \begin{bmatrix} \boldsymbol{f}_{m,i,k} \\ \boldsymbol{g}_{m,i,k}, \end{bmatrix} \tag{28}$$

for $m = 1 \dots M$ where $\boldsymbol{f}_{m,i,k}$ and $\boldsymbol{g}_{m,i,k}$ are M-dimensional vectors. We obtain the update rule for $\boldsymbol{C}_{i,k}$ as

$$\boldsymbol{C}_{i,k} \leftarrow \boldsymbol{G}_{i,k} \boldsymbol{F}_{i,k}^{-1}, \tag{29}$$

where $\boldsymbol{F}_{i,k} = [\boldsymbol{f}_{1,i,k}, \dots, \boldsymbol{f}_{M,i,k}]$ and $\boldsymbol{G}_{i,k} = [\boldsymbol{g}_{1,i,k}, \dots, \boldsymbol{g}_{M,i,k}]$.

L^{+} is equal up to constant terms to the sum of the log posteriors of HMMs, when viewed as a function of Z. Thus, we can invoke the Viterbi algorithm to search for the optimal path $z_{i,1}, \dots, z_{i,L}$ for each i individually. Note that updating $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{C}$ and $\boldsymbol{Z}$ corresponds to solving the problems of blind signal separation and source activity detection based on a unified objective function.

4. Experimental evaluation

We evaluated the performance of the present method in terms of the ability to separate sources and detect their activity. We used a mixed stereo signal as the experimental data, each of which we obtained by mixing the non-speech signals (sounds of a whistle, a bell and a stapler) from the RWCP database[8] and was convolved with the measured room impulse response from the RWCP database [8] (in which the distance between the microphones was 5.85 cm and the reverberation time was 0 ms). Thus, the three signals were artificially located 20°, 60° and 100° from the microphones respectively. The sampling rate was 32 kHz. To compute the STFT components of the observed signal, the STFT frame length was set at 16 ms and a Hamming window was used with an overlap length of 8 ms. Fig. 2 shows a spectrogram of the mixed signal. We set the number of states of HMMs D as 3. We expected that $d = 1$ is an inactive state and $d = 2$ and 3 are active states, by setting α_1 and β_1 as 1 and 10^{-2} respectively, and $\alpha_{2:3}$ and $\beta_{2:3}$ 1 and 10^{20} respectively. the transition probability $\boldsymbol{\rho}$ as $\boldsymbol{\rho}_1 = (0.9, 0.1, 0)$, $\boldsymbol{\rho}_2 = (0, 0.5, 0.5)$ and $\boldsymbol{\rho}_3 = (0.5, 0, 0.5)$ based on the assumption that a sound source tends to stay inactive. The diagonal elements of $\boldsymbol{C}$ were initially all set to $1/\sqrt{M}$, and the off-diagonal elements were initially set to zero. We first performed the single channel NMF based on IS-divergence starting from different random initial matrices W and H, then we set the results of $\boldsymbol{W}$, $\boldsymbol{H}$ as the initial parameters of $\boldsymbol{W}$ and $\boldsymbol{H}$. The parameter estimation algorithm was run for 100 iterations. We chose the method proposed in [5] as a comparison. The separated signal $\hat{\boldsymbol{y}}_{i,k,l}$ was obtained by Wiener filtering

$$\hat{\boldsymbol{y}}_{i,k,l} = w_{i,k,z_{i,l}} h_{i,l} \boldsymbol{C}_{i,k} \hat{\boldsymbol{X}}_{k,l}^{-1} \boldsymbol{y}_{k,l}. \tag{30}$$

As an evaluation measure, we used the signal-to-distortion ratio (SDR) [9]. The SDR is expressed in decibels (dB), and a higher SDR indicates superior quality. The input SDRs [dB] of the bell, the whistle and the stapler were -8.36, 6.64 and -12.78 [dB], respectively.

Table 1 shows the average SDRs and standard deviations for the ten trials obtained by the conventional and proposed methods. The average SDRs of the proposed method were superior to those of the conventional method for each signal. The total average of the SDRs obtained with the proposed method was 7.65 dB more than that obtained with the conventional approach. These results show the effectiveness of the proposed method for BSS. Examples of separated signals are available at `http://www.hil.t.u-tokyo.ac.jp/~higuchi/demo/Examples_interspeech.htm`. Fig. 3 shows examples of spectrogram of signal of a bell (top), that of separated signal (middle), and its activity detection result (bottom). The activities of the bell were almost all estimated correctly. However, our results also showed that abilities of signal separation and source activity detection by our method depended on the first condition of $\boldsymbol{W}$ and $\boldsymbol{H}$ to some extent. As future work, we plan to solve this kind of problem by learning $\boldsymbol{W}$ and $\boldsymbol{\rho}$ from a clean signal and apply our proposed method to acoustic event detection.

5. Conclusion

In this paper we extend our previous approach to a multichannel observed signal and proposes a comprehensive approach to deal with an underdetermined BSS problem and source activity detection. Specifically, we describe a generative model of mixture signals by incorporating a generative model of spectra for each source, using MFHMMs. Each hidden state of the present HMM represents states of each source such as its activities. Through the estimation of the parameters of the overall generative model, we can simultaneously the underdetermined blind signal separation and source activity detection. The experiment showed that the proposed algorithm provided a 7.65 dB improvement compared with the conventional method in terms of the signal-to-distortion ratio.

6. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 26730100.

7. References

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

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

[3] P. Smaragdis, and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," *Proc. WASPAA 2003*, Oct. 2003, pp. 177–180.

[4] A. Ozerov, and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550-563, Mar.2010.

[5] H. Sawada, H. Kameoka, S. Araki and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012*, pp. 261-264, 2012.

[6] C. Févotte, N. Bertin, and J.L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[7] Masahiro Nakano, Le Roux Jonathan, Hirokazu Kameoka, Yu Kitano, Nobutaka Ono, and Shigeki Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," *Latent Variable Analysis and Signal Separation*, vol. 6365, pp. 149–156, 2010.

[8] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. 2nd International Conference on Language Resources & Evaluation (LREC 2000)*, pp. 965–968, 2000.

[9] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1462–1469, 2006.