2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2014, REIMS, FRANCE

JOINT AUDIO SOURCE SEPARATION AND DEREVERBERATION BASED ON MULTICHANNEL FACTORIAL HIDDEN MARKOV MODEL

Takuya Higuchi[1)] *and Hirokazu Kameoka*[1),2)]

[1)]Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[2)]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
3-1 Morinosatowakamiya, Atsugi, Kanagawa 243-0198, Japan
{higuchi,kameoka}@hil.t.u-tokyo.ac.jp

ABSTRACT

This paper proposes a unified approach for jointly solving underdetermined source separation, audio event detection and dereverberation of convolutive mixtures. 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. Several attempts have recently been made to extend this approach to a multichannel case in order to utilize the spatial correlation of the multichannel inputs as an additional clue for source separation. The multichannel NMF assumes 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. We have previously proposed an extension of this approach, in which the variations over time of the spectral density and the total power of each source is modeled by a hidden Markov model (HMM). This has allowed us to solve source activity detection and source separation simultaneously through model parameter inference. While this method was based on an anechoic mixing model, the aim of this paper is to further extend the above approach to deal with reverberation by incorporating an echoic mixing model into the generative model of observed signals. Through an experiment of underdetermined source separation under reverberant conditions, we confirmed that the proposed method provided a 9.61 dB improvement compared with the conventional method in terms of the signal-to-interference ratio.

Index Terms: source separation, dereverberation, audio event detection, non-negative matrix factorization, multichannel factorial hidden Markov model,

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. Several attempts have recently been made to extend this approach to a multichannel case in order to allow for the use of the spatial correlation of multichannel inputs 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. 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, we previously extended 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). We call this model the "multichannel factorial hidden Markov model (MFHMM)" [6]. With this model, we are able to perform source separation and source activity detection simultaneously. However, in the models mentioned above, the length of the impulse response from a source to microphones is assumed to be sufficiently shorter than the frame length of the STFT so that the observed signal can be approximately modeled by an instantaneous mixture in the time-frequency domain. To deal with reverberation, this

978-1-4799-3694-6/14/$31.00

paper aims to further extend the above approach by incorporating an echoic mixing model into MFHMM. Parameter inference of the present model allows us to simultaneously solve source separation, source activity detection and dereverberation 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 under a reverberant condition, Sec. 3 presents the generative model of a source signal based on an HMM, Sec. 4 dervies a parameter estimation algorithm for the present model and Sec. 5 presents some results of a source separation experiment.

2. ECHOIC MIXING MODEL

First we consider a situation where I source signals are recorded by M microphones. Here, let $y_m(t) \in \mathbb{R}$ be the observed signal at the m-th microphone, and $s_i(t) \in \mathbb{R}$ be the signal of the i-th source. The observed signal can be written in the time domain:

$$\boldsymbol{y}(t) = \sum_{i=1}^{I} \int_{-\infty}^{\infty} \boldsymbol{a}_i(\tau) s_i(t-\tau) \mathrm{d}\tau, \tag{1}$$

where $\boldsymbol{y}(t) = (y_1(t), \ldots, y_M(t))^{\mathsf{T}} \in \mathbb{R}^M$ and $\boldsymbol{a}_i(t) = (a_{i,1}(t), \ldots, a_{i,M}(t))^{\mathsf{T}} \in \mathbb{R}^M$. $a_{i,m}(t)$ denotes the impulse response between source i and microphone m. 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 time-frequency domain:

$$\boldsymbol{y}(\omega_k, t_l) = \sum_{i=1}^{I} \boldsymbol{a}_i(\omega_k) s_i(\omega_k, t_l), \tag{2}$$

where $\boldsymbol{y}(\omega_k, t_l) = (y_1(\omega_k, t_l), \ldots, y_M(\omega_k, t_l))^{\mathsf{T}} \in \mathbb{C}^M$ and $\boldsymbol{a}_i(\omega_k) = (a_{i,1}(\omega_k), \ldots, a_{i,M}(\omega_k))^{\mathsf{T}} \in \mathbb{C}^M$. 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. $\boldsymbol{a}_i(\omega_k)$ denotes the frequency array response for source i at frequency ω_k. However, in a reverberant condition, the length of the impulse responses are relatively long and so an instantaneous mixture approximation is not always true. Therefore we approximately express the observed signals as a form of a convolution of the frequency array response and the source signal in the time-frequency domain:

$$\boldsymbol{y}(\omega_k, t_l) \approx \sum_{i=1}^{I} \sum_{n=0}^{N} \boldsymbol{a}_i(\omega_k, t_n) s_i(\omega_k, t_l - t_n). \tag{3}$$

$0 \leq n \leq N$ is the time index of the frequency array response in the time-frequency domain. Note that $\boldsymbol{a}_i(\omega_k, t_1 : t_N)$ denote the frequency array responses which correspond to the impulse responses out of the frames of the STFT. This approximation is useful for dereverberation [7] and the validity of the approximation is experimentally shown. For convenience of notation, we hereafter use subscripts k, l and n to indicate ω_k, t_l and t_n respectively.

3. MULTICHANNEL FACTORIAL HMM

3.1. Generative process of observed signals

Here we describe the generative process of an observed signal based on Eq. (3). 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 $\sigma^2_{i,k,l}$,

$$s_{i,k,l} | \sigma_{i,k,l} \sim \mathcal{N}_{\mathbb{C}}(s_{i,k,l}; 0, \sigma^2_{i,k,l}), \tag{4}$$

where $\sigma^2_{i,k,l}$ denotes the power spectral density of i-th source at frequency k and time l and $\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \propto \exp(-(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{H}} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}))$. From Eq. (3) and Eq. (4), $\boldsymbol{y}_{k,l}$ is also normally distributed such that

$$\begin{aligned} \boldsymbol{y}_{k,l} | \boldsymbol{a}_{1:I,k,0:N}, \sigma_{i,k,l-N:l} \\ \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; 0, \sum_{i,n} \boldsymbol{C}_{i,k,n} \sigma^2_{i,k,l-n}), \end{aligned} \tag{5}$$

conditioned on $\boldsymbol{a}_{1:I,k,0:N}$ and $\sigma_{i,k,l-N:l}$ where $\boldsymbol{C}_{i,k,n} = \boldsymbol{a}_{i,k,n} \boldsymbol{a}^{\mathsf{H}}_{i,k,n}$. $\boldsymbol{C}_{i,k,n}$ represents n-th spatial correlation matrix of the i-th source at frequency k.

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

In the regular NMF model (see [8]), 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

$$\sigma^2_{i,k,l} = w_{i,k} h_{i,l}, \tag{6}$$

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

$$\boldsymbol{s}_{i,k,l} | w_{i,k}, h_{i,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_{i,k,l}; 0, w_{i,k} h_{i,l}), \tag{7}$$

conditioned on $w_{i,k}$ and $h_{i,l}$.

3.3. 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 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 model the sequence of the power spectra and the total powers of each source using an HMM.

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:

$$z_{i,l} | z_{i,l-1} \sim \mathrm{Categorical}(z_{i,l}; \boldsymbol{\rho}_{z_{i,l-1}}), \tag{8}$$

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 $z_{i,l}$,

$$h_{i,l}|z_{i,l} \sim \mathrm{Gamma}(h_{i,l};\alpha_{z_{i,l}}\beta_{z_{i,l}}), \tag{9}$$

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. We expect that this setting allows us to solve source separation and source 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

$$\begin{aligned} &s_{i,k,l}|w_{i,k,1:D},h_{i,l},z_{i,l} \\ &\quad\sim \mathcal{N}_{\mathbb{C}}(s_{i,k,l};0,w_{i,k,z_{i,l}}h_{i,l}). \end{aligned} \tag{10}$$

Since the generative model of $\boldsymbol{y}_{k,l}$ contains multiple HMMs associated with the underlying sources, the overall model can be viewed as a Factorial HMM. Our overall generative model is given by Eqs. (8), Eqs. (9) and

$$\begin{aligned} &\boldsymbol{y}_{k,l}|\boldsymbol{a}_{1:I,k,0:N},w_{1:I,k,1:D},h_{1:I,l-N:l},z_{1:I,l-N:l} \\ &\quad\sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l};0,\sum_{i,n}\boldsymbol{C}_{i,k,n}w_{i,k,z_{i,l-n}}h_{i,l-n}), \end{aligned} \tag{11}$$

conditioned on $\boldsymbol{a}_{1:I,k,0:N}$, $w_{1:I,k,1:D}$, $h_{1:I,l-N:l}$ and $z_{1:I,l-N:l}$.

3.4. Related work

The present model is equivalent to the model proposed in [6] when the number N is set at 0, which leads to an anechoic mixing model.

4. ALGORITHM FOR PARAMETER ESTIMATION

4.1. 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 $\boldsymbol{W} = w_{1:I,1:K,1:D}$, $\boldsymbol{H} = h_{1:I,1:L}$, $\boldsymbol{C} = \boldsymbol{C}_{1:I,1:K,0:N}$ 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

$$p(\Theta|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y},\Theta)}{p(\boldsymbol{Y})}, \tag{12}$$

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. 3, we can write the joint distribution $p(\boldsymbol{Y},\Theta)$ as

$$p(\boldsymbol{Y},\Theta) \propto p(\boldsymbol{Y}|\Theta)p(\boldsymbol{H}|\boldsymbol{Z})p(\boldsymbol{Z}). \tag{13}$$

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

$$\hat{\Theta} = \operatorname*{argmax}_{\Theta} \log p(\Theta|\boldsymbol{Y}). \tag{14}$$

By using Eqs. (12), (13) and (14), the current optimization problem can be rewritten as

$$\begin{aligned} \hat{\Theta} = \operatorname*{argmax}_{\Theta}\big(&\log p(\boldsymbol{Y}|\Theta) \\ &+ \log p(\boldsymbol{H}|\boldsymbol{Z}) + \log p(\boldsymbol{Z})\big). \end{aligned} \tag{15}$$

According to the generative model defined in Sec. 3, $\log p(\boldsymbol{Y}|\Theta)$ is written as

$$\begin{aligned} &\log p(\boldsymbol{Y}|\Theta) \\ &= -\frac{1}{2}\sum_{k,l}(M\log 2\pi + \log|\hat{\boldsymbol{X}}_{k,l}| + \boldsymbol{y}_{k,l}{}^{\mathsf{H}}\hat{\boldsymbol{X}}_{k,l}^{-1}\boldsymbol{y}_{k,l}), \end{aligned} \tag{16}$$

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

4.2. 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-decreasing under the updates $\Theta \leftarrow \mathrm{argmax}_{\Theta} L^{+}(\Theta,\Lambda)$ and $\Lambda \leftarrow \mathrm{argmax}_{\Lambda} L^{+}(\Theta,\Lambda)$. 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.

$$\begin{aligned} &L(\Theta) \\ \geq &L^{+}(\Theta,\Lambda) \\ = &-\frac{1}{2}\sum_{k,l}\Bigg\{\sum_{i,n}\Bigg(\frac{\mathrm{tr}(\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,n}\boldsymbol{C}_{i,k,n}^{-1}\boldsymbol{R}_{i,k,l,n})}{w_{i,k,z_{i,l-n}}h_{i,l-n}} \\ &+\mathrm{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{C}_{i,k,n})w_{i,k,z_{i,l-n}}h_{i,l-n}\Bigg) + \log|\boldsymbol{U}_{k,l}| - M\Bigg\} \\ &+\sum_{i,l}\big\{(\alpha_{z_{i,l}}-1)\log h_{i,l} - h_{i,l}/\beta_{z_{i,l}} - \alpha_{z_{i,l}}\log\beta_{z_{i,l}}\big\} \\ &+\log p(\boldsymbol{Z}), \end{aligned} \tag{17}$$

where $\boldsymbol{R}_{i,k,l,n}$ and $\boldsymbol{U}_{k,l}$ are auxiliary variables that satisfy Hermitian positive definiteness and $\sum_{i,n}\boldsymbol{R}_{i,k,l,n} = \boldsymbol{I}$. We

denote the set of the auxiliary variables as Λ. $\mathrm{tr}(\cdot)$ is the trace of a matrix. The equality $L(\Theta) = L^+(\Theta, \Lambda)$ is satisfied when

$$\boldsymbol{R}_{i,k,l,n} = \boldsymbol{C}_{i,k,n} w_{i,k,z_{i,l-n}} h_{i,l-n} \hat{\boldsymbol{X}}_{k,l}^{-1}, \tag{18}$$

$$\boldsymbol{U}_{k,l} = \hat{\boldsymbol{X}}_{k,l}. \tag{19}$$

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. (18) and (19). 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

$$\frac{\partial L^+}{\partial w_{i,k,z_{i,l}}} = \frac{1}{2} \sum_{l,n} \left(\frac{\mathrm{tr}(\boldsymbol{y}_{k,l+n} \boldsymbol{y}_{k,l+n}{}^{\mathsf{H}} \boldsymbol{R}_{i,k,l+n,n} \boldsymbol{C}_{i,k,n}^{-1} \boldsymbol{R}_{i,k,l+n,n})}{w_{i,k,z_{i,l}}^2 h_{i,l}} - \mathrm{tr}(\boldsymbol{U}_{k,l+n}^{-1} \boldsymbol{C}_{i,k,n}) h_{i,l} \right), \tag{20}$$

$$\frac{\partial L^+}{\partial h_{i,l}} = \frac{1}{2} \sum_{k,n} \left(\frac{\mathrm{tr}(\boldsymbol{y}_{k,l+n} \boldsymbol{y}_{k,l+n}{}^{\mathsf{H}} \boldsymbol{R}_{i,k,l+n,n} \boldsymbol{C}_{i,k,n}^{-1} \boldsymbol{R}_{i,k,l+n,n})}{w_{i,k,z_{i,l}} h_{i,l}^2} - \mathrm{tr}(\boldsymbol{U}_{k,l+n}^{-1} \boldsymbol{C}_{i,k,n}) w_{i,k,z_{i,l}} \right) + (\alpha_{z_{i,l}} - 1)/h_{i,l} - 1/\beta_{z_{i,l}}, \tag{21}$$

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

$$w_{i,k,z_{i,l}} \leftarrow \sqrt{\frac{\sum_{l,n} \frac{\mathrm{tr}(\boldsymbol{y}_{k,l+n} \boldsymbol{y}_{k,l+n}{}^{\mathsf{H}} \boldsymbol{R}_{i,k,l+n,n} \boldsymbol{C}_{i,k,n}^{-1} \boldsymbol{R}_{i,k,l+n,n})}{h_{i,l}}}{\sum_{l,n} \mathrm{tr}(\boldsymbol{U}_{k,l+n}^{-1} \boldsymbol{C}_{i,k,n}) h_{i,l}}}, \tag{22}$$

$$h_{i,l} \leftarrow \frac{(\alpha_{z_{i,l}} - 1) + \sqrt{(\alpha_{z_{i,l}} - 1)^2 + \mu_{i,l} \nu_{i,l}}}{\nu_{i,l}}, \tag{23}$$

where

$$\mu_{i,l} = \sum_{k,n} \frac{\mathrm{tr}(\boldsymbol{y}_{k,l+n} \boldsymbol{y}_{k,l+n}{}^{\mathsf{H}} \boldsymbol{R}_{i,k,l+n,n} \boldsymbol{C}_{i,k,n}^{-1} \boldsymbol{R}_{i,k,l+n,n})}{w_{i,k,z_{i,l}}}, \tag{24}$$

$\nu_{i,l}$

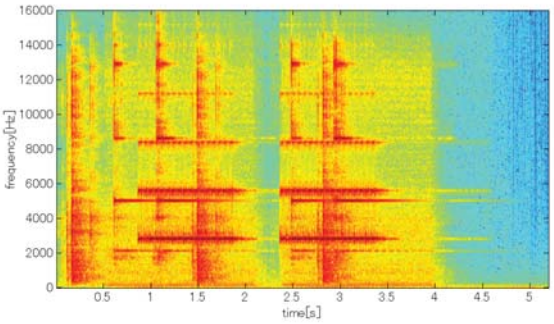

Fig. 1. Spectrogram of a mixture signal.

$$= \sum_{k,n} \mathrm{tr}(\boldsymbol{U}_{k,l+n}^{-1} \boldsymbol{C}_{i,k,n}) w_{i,k,z_{i,l}} + 2/\beta_{z_{i,l}}. \tag{25}$$

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

$$\frac{\partial L^+}{\partial \boldsymbol{C}_{i,k,n}} = \sum_{l} \left(\frac{\boldsymbol{C}_{i,k,n}^{-1} \boldsymbol{R}_{i,k,l,n} \boldsymbol{y}_{k,l} \boldsymbol{y}_{k,l}{}^{\mathsf{H}} \boldsymbol{R}_{i,k,l,n} \boldsymbol{C}_{i,k,n}^{-1}}{w_{i,k,z_{i,l-n}} h_{i,l-n}} - \boldsymbol{U}_{k,l}^{-1} w_{i,k,z_{i,l-n}} h_{i,l-n} \right). \tag{26}$$

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

$$\boldsymbol{C}_{i,k,n} \boldsymbol{A}_{i,k,n} \boldsymbol{C}_{i,k,n} = \boldsymbol{B}_{i,k,n}, \tag{27}$$

where

$$\boldsymbol{A}_{i,k,n} = \sum_{l} w_{i,k,z_{i,l-n}} h_{i,l-n} \hat{\boldsymbol{X}}_{k,l}^{-1},$$

$$\boldsymbol{B}_{i,k,n} = \boldsymbol{C}_{i,k,n} (\sum_{l} w_{i,k,z_{i,l-n}} h_{i,l-n} \hat{\boldsymbol{X}}_{k,l}^{-1} \boldsymbol{y}_{k,l} \boldsymbol{y}_{k,l}{}^{\mathsf{H}} \hat{\boldsymbol{X}}_{k,l}^{-1}) \boldsymbol{C}_{i,k,n}. \tag{28}$$

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

$$\begin{bmatrix} 0 & -\boldsymbol{A}_{i,k,n} \\ -\boldsymbol{B}_{i,k,n} & 0 \end{bmatrix}, \tag{29}$$

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

$$\boldsymbol{e}_{m,i,k,n} = \begin{bmatrix} \boldsymbol{f}_{m,i,k,n} \\ \boldsymbol{g}_{m,i,k,n} \end{bmatrix}, \tag{30}$$

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

$$\boldsymbol{C}_{i,k,n} \leftarrow \boldsymbol{G}_{i,k,n} \boldsymbol{F}_{i,k,n}^{-1}, \tag{31}$$

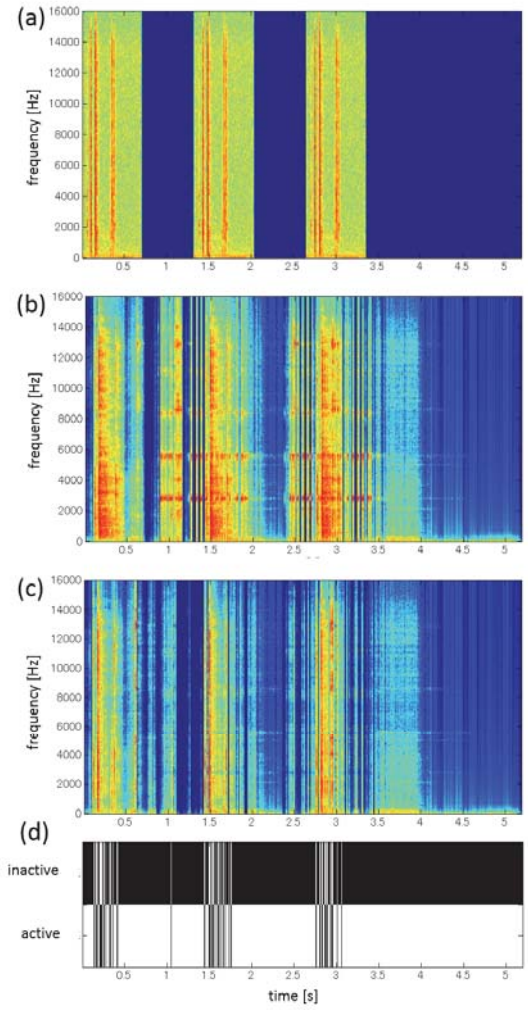

Fig. 2. (a)A spectrogram of the source signal of the stapler recorded in an anechoic condition, (b)that of the separated signal of the stapler obtained by the conventional method, (c)that of the separated and dereverbed signal of the stapler obtained by the proposed method, and (d)the acoustic event detection result obtained by the proposed method. Black indicates the state is assigned at the time.

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

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

5. EXPERIMENTAL EVALUATION

We evaluated the performance of the present method in terms of the abilities of source separation, source activity detection and dereverberation in a supervised case. We used a mixed stereo signal as the experimental data, each of which we obtained by mixing the non-speech signals (sounds of a cell phone, a bell and a stapler) from the RWCP database[9] and was convolved with the measured room impulse response from the RWCP database [9] (in which the distance between the microphones was 2.83 cm and the reverberation time was 600 ms). Thus, the three signals were artificially located 60°, 90° and 130° from the microphones respectively. Fig. 1 shows a spectrogram of the observed signal in the reverberant condition. 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. We set the number of states of HMMs D as 2. We expected that $d = 1$ was an inactive state and $d = 2$ was an active state, by setting α_1 and β_1 as 1 and 10^{-2} respectively, and α_2 and β_2 as 1 and 10^{20} respectively. The diagonal elements of $\boldsymbol{C}_{i,k,0}$ were initially all set to $1/\sqrt{M}$, and the off-diagonal elements were initially set to zero. For $n = 1, \ldots, N$, the diagonal elements of $\boldsymbol{C}_{i,k,n}$ were set to $10^{-3}/\sqrt{M}$, and the off-diagonal elements were also set to zero initially. We first learned $\boldsymbol{W}$, $\boldsymbol{H}$ and $\boldsymbol{\rho}$ from the three clean source signal recorded in the anechoic condition with the proposed method starting from random initial matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ (therefore this experiment was a supervised source separation). The parameter estimation algorithm was run for 150 iterations. In order to avoid an undesirable local optima, we iterated the proposed algorithm 100 times with setting N as 0, then gradually increased N up to 20 according to the iteration. We chose the method proposed in [6] as a comparison. This method equals to be our proposed method when we set $N = 0$. The separated signal $\hat{\boldsymbol{y}}_{i,k,l}$ was obtained by Wiener filtering

$$\hat{\boldsymbol{y}}_{i,k,l} = w_{i,k,z_{i,l}} h_{i,l} \boldsymbol{C}_{i,k,0} \hat{\boldsymbol{X}}_{k,l}^{-1} \boldsymbol{y}_{k,l}. \tag{32}$$

As evaluation measures, we used the signal-to-interference ratio (SIR) [10]. The SIR is expressed in decibels (dB), and a higher SIR indicates superior quality. The input SIR of the sounds of the bell, the cell phone and the stapler were -16.61, 16.58 and -39.17 [dB], respectively.

Table 1. The SIRs of the three sources obtained by the conventional and proposed methods.

SIR [dB]	bell	phone	stapler
Proposed	43.35	32.05	7.19
Conventional	32.37	30.15	-8.77

Table 1 shows the SIRs obtained by the conventional and proposed methods. The average of the SIRs obtained with the proposed method was 9.61 [dB] more than that obtained with the conventional approach. These results show the effectiveness of the proposed method for source separation in a reverberant condition. Fig. 2 shows (a)a spectrogram of the source signal of the stapler recorded in an anechoic condition, (b)that of the separated signal of the stapler obtained by the conventional method, (c)that of the separated and dereverbed signal of the stapler obtained by the proposed method,

and (d)the acoustic event detection result obtained by the proposed method respectively. Black indicates the state is assigned at the time. We can see that the reverberant components were relatively removed by the proposed method and the acoustic event was roughly detected.

6. CONCLUSION

This paper proposed a unified approach for jointly solving underdetermined source separation, audio event detection and dereverberation of convolutive mixtures based on MFHMM with an echoic mixing model. Through an experiment of supervised source separation under reverberant conditions, we confirmed that the proposed method provided a 9.61 dB improvement compared with the conventional method in terms of the signal-to-interference ratio.

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," in *Proc. 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2003)*, pp. 177–180, 2003.

[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," in *Proc. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, pp. 261–264, 2012.

[6] T. Higuchi, H. Takeda, T. Nakamura and H. Kameoka, "A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden Markov models," in *Proc. 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*, to appear.

[7] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B.-H. Juang, "Blind speech dereverberation with multichannel linear prediction based on short time Fourier transform representation," in *Proc. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pp. 85–88, 2008.

[8] 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.

[9] 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.

[10] 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.