16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

AN INTEGRATED METHOD FOR BLIND SEPARATION AND DEREVERBERATION OF CONVOLUTIVE AUDIO MIXTURES

Takuya Yoshioka, Tomohiro Nakatani, and Masato Miyoshi

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
email: takuya@cslab.kecl.ntt.co.jp

ABSTRACT

This paper proposes an integrated method for the blind separation and dereverberation of convolutive audio mixtures. The proposed method is based on multichannel blind deconvolution in the frequency domain. Significant points to be emphasized are as follows: (1) The objective function for optimizing the deconvolution system is derived based on a time-varying all-pole model of source signals, which was proven to be effective for single source dereverberation. This provides the proposed method with the capacity for both separation and dereverberation. (2) An efficient optimization algorithm is developed. This algorithm is realized by decomposing the deconvolution system into an instantaneous separation part and a multichannel auto-regressive part. Illustrative experimental results with an RT_{60} of 0.6 seconds are reported, where the proposed method showed superiority over a conventional frequency-domain blind separation method.

1. INTRODUCTION

When recording sound with microphones in a room, acoustic signals emitted from sound sources are often reverberated and coupled with each other. The acoustic signals observed at the microphones are thus linear convolutive mixtures of the source signals. Recovering the individual source signals from the observed signals, namely the blind separation and dereverberation of the convolutive mixtures, will be useful for many audio applications.

There are two common approaches to this task. One is blind source separation (BSS) in the frequency domain. It is well known that the frequency-domain BSS approach fails in separation when reverberation time is long [1]. The other is multichannel blind deconvolution (MBD), which may be applied either in the time domain [2, 3] or in the frequency domain [4]. Ideally, the MBD approach is able to separate and dereverberate the convolutive mixtures even under such highly reverberant conditions. In reality, however, such conditions oblige us to use high order deconvolution systems. Since most conventional MBD methods are based on the gradient descent method, the order increase results in poor convergence performance and large computational cost. In addition, the conventional MBD methods cannot achieve dereverberation since the objective functions employed by these methods provide insufficient information about the temporal structures of source signals.

Recently, on the other hand, we investigated the single source dereverberation. In our previous reports, we derived objective functions for the dereverberation based on a time-varying all-pole (TVAP) model of a source signal (see, for example, [5]). Optimizing dereverberation systems so that one of the objective functions is maximized led to high dereverberation performance.

In this paper, we embed the TVAP source model in the MBD approach. This enables us to achieve both the blind separation and dereverberation of convolutive mixtures even under highly reverberant conditions. The main results are as follows.

1. We develop a method for blind separation and dereverberation in the framework of the maximum likelihood (ML) estimation. The statistical model of observed signals is defined based on the TVAP source model. The parameters of this observation model consist of all-pole source parameters and matrices of a deconvolution system. These parameters are optimized to maximize the likelihood function.
2. We propose to decompose the deconvolution system into an instantaneous separation part and a multichannel auto-regressive (AR) part. This decomposition allows us to obtain an efficient optimization algorithm that comprises three analytically calculated optimization steps.

The proposed method is a form of integration of blind separation and dereverberation in the sense that the instantaneous separation system and the multichannel AR system are optimized by using a common objective function (i.e. the likelihood function). Indeed, if the AR part is forced to be an identity system, the proposed method reduces to a new frequency-domain BSS method based on second-order statistics. On the other hand, if the instantaneous separation part is fixed at an identity system, the proposed method reduces to a frequency-domain dereverberation method similar to [6].

2. BLIND SEPARATION AND DEREVERBERATION

Suppose there are M sound sources and M microphones. Let $s^{(m_1)}(n)$ and $y^{(m_2)}(n)$ denote the m_1-th source signal and the m_2-th observed signal, respectively. We summarize the source signals and the observed signals in vectors, respectively, as $\boldsymbol{s}(n) = [s^{(1)}(n),\cdots,s^{(M)}(n)]^T$ and $\boldsymbol{y}(n) = [y^{(1)}(n),\cdots,y^{(M)}(n)]^T$, where superscript T stands for non-conjugate transposition. The observed signal vector, $\boldsymbol{y}(n)$, is generated by

$$\boldsymbol{y}(n) = \sum_k B(k)^H \boldsymbol{s}(n-k), \tag{1}$$

where superscript H stands for conjugate transposition and the (m_1,m_2)-th entry of $B(k)^H$ is the k-th coefficient of the room transfer function between the m_1-th source and the m_2-th microphone.

The proposed method postulates time-frequency domain signal representation. Short-time Fourier transform (STFT) is used for the time-frequency analysis. Unlike conventional frequency domain BSS methods such as [7], the proposed method uses a short time frame of about 30 milliseconds in order to utilize the TVAP source model effectively.

Let $S_{t,l}^{(m_1)}$ and $Y_{t,l}^{(m_2)}$ denote the spectral component of $s^{(m_1)}(n)$ and that of $y^{(m_2)}(n)$, respectively, at the t-th frame and the l-th frequency band. We summarize the source spectral components and the observed spectral components in vectors, respectively, as $\boldsymbol{s}_{t,l} = [S_{t,l}^{(1)},\cdots,S_{t,l}^{(M)}]^T$ and $\boldsymbol{y}_{t,l} = [Y_{t,l}^{(1)},\cdots,Y_{t,l}^{(M)}]^T$. We assume that the sequence of the source spectral component vectors in the l-th frequency band can be recovered with causal finite impulse response (FIR) filters of order K_l as

$$\boldsymbol{s}_{t,l} = \sum_{k=0}^{K_l} W_{k,l}^H \boldsymbol{y}_{t-k,l}. \tag{2}$$

We admit that (2) is not an exact inverse of mixing model (1) due mainly to its causality. However, by restricting the deconvolution

system to be causal, it is possible to develop an effective optimization algorithm. Therefore, this inverse model would serve as a good starting point for integrating blind separation and dereverberation.

Now suppose that we observe the spectral sequences at frames $t = 0, \cdots, T-1$. We collectively represent the observed spectra and the corresponding unknown source spectra as $\mathscr{Y} = \{\boldsymbol{y}_{t,l}\}_{0\le t\le T-1, 0\le l\le L-1}$ and $\mathscr{S} = \{\boldsymbol{s}_{t,l}\}_{0\le t\le T-1, 0\le l\le L-1}$, respectively, where T is the number of frames and L is the number of frequency bands. The task of blind separation and dereverberation is defined as setting up the deconvolution system parameters, $\mathscr{W} = \{\{W_{k,l}\}_{0\le k\le K_l}\}_{0\le l\le L-1}$, so that the unknown source spectra, $\mathscr{S}$, are recovered.

3. STATISTICAL FORMULATION

In this section, we formulate the blind separation and dereverberation task as a maximum likelihood (ML) estimation task to set up an objective function for optimizing a deconvolution system.

3.1 Source model

Fourmulation as an ML estimation task requires a source signal model. Here, we use the TVAP source model, which may allow us to achieve both separation and dereverberation unlike conventional source models such as the non-Gaussian independent and identically distributed (i.i.d.) model. The TVAP source model assumes the following conditions. These assumptions have been widely accepted in the literature, especially for speech signals.

1. The power spectral density (PSD) of each source signal is an all-pole form. Let ${}_s\lambda_t^{(m)}(\omega)$ denote the PSD of the m-th source signal at the t-th frame and angular frequency ω. Then, ${}_s\lambda_t^{(m)}(\omega)$ is represented as

$$ {}_s\lambda_t^{(m)}(\omega) = \frac{{}_sv_t^{(m)}}{|A_t^{(m)}(e^{j\omega})|^2} \tag{3} $$

$$ A_t^{(m)}(z) = 1 - a_{t,1}^{(m)} z^{-1} - \cdots, -a_{t,P}^{(m)} z^{-P}, \tag{4} $$

   where $\{a_{t,1}^{(m)}, \cdots, a_{t,P}^{(m)}\}$ and ${}_sv_t^{(m)}$ are called linear predictor coefficients (LPCs) and a prediction residual, respectively.

2. Each source spectral component $S_{t,l}^{(m)}$ follows a complex Gaussian process with mean 0 and variance ${}_s\lambda_t^{(m)}(2\pi l/L)$. Therefore, we have

$$ p(S_{t,l}^{(m)}; a_{t,1}^{(m)}, \cdots, a_{t,P}^{(m)}, {}_sv_t^{(m)}) = \mathscr{N}_{\mathbb{C}}\{S_{t,l}^{(m)}; 0, {}_s\lambda_t^{(m)}(2\pi l/L)\}, \tag{5} $$

   where $N_{\mathbb{C}}\{\boldsymbol{x}; \boldsymbol{\mu}, \Sigma\}$ is the PDF of a complex Gaussian random variable $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ. If the dimension of $\boldsymbol{x}$ is D, the PDF is defined as [8]

$$ \mathscr{N}_{\mathbb{C}}\{\boldsymbol{x}; \boldsymbol{\mu}, \Sigma\} = \frac{1}{\pi^D \det\Sigma} \exp\{-\boldsymbol{x}^H \Sigma^{-1} \boldsymbol{x}\}. \tag{6} $$

3. For any m, $S_{t_1,l_1}^{(m)}$ and $S_{t_2,l_2}^{(m)}$ are statistically independent unless $(t_1, l_1) = (t_2, l_2)$.

In addition, we assume the following condition as regards the relationship between spectral components of different sources.

4. If $m_1 \neq m_2$, $S_{t_1,l_1}^{(m_1)}$ and $S_{t_2,l_2}^{(m_2)}$ are statistically independent for any (t_1, l_1, t_2, l_2).

The most important property of the TVAP source model is that the PSD varies every short time frame. It was shown in [6] that the power spectral time-variance plays an important role in accomplishing dereverberation. Note that the TVAP source model is different from the time-invariant auto-regressive source model of [9] and the nonstationary source signal model with a relatively long time frame used in [10].

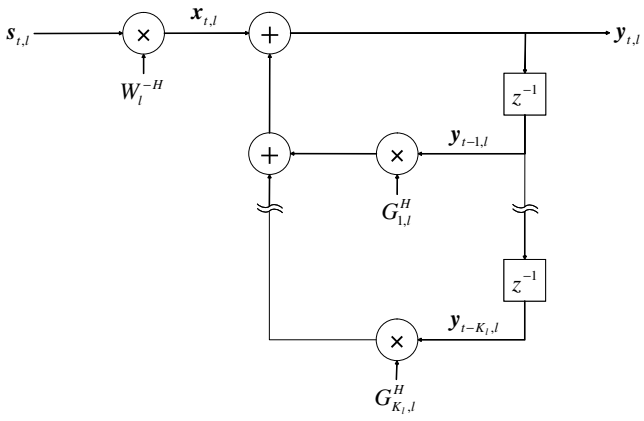

Figure 1: Schematic diagram of forward model.

3.2 Equivalent Forward Model

In order to derive the statistical observation model, we transform inverse model (2) into a forward model as follows. (2) is rewritten as

$$ \boldsymbol{s}_{t,l} = W_{0,l}^H \Big(\boldsymbol{y}_{t,l} + \sum_{k=1}^{K_l} W_{0,l}^{-H} W_{k,l}^H \boldsymbol{y}_{t-k,l} \Big). \tag{7} $$

Rewriting (7) by using $W_l = W_{0,l}$ and $G_{k,l} = W_{k,l} W_{0,l}^{-1}$ yields

$$ \boldsymbol{y}_{t,l} = \sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{y}_{t-k,l} + \boldsymbol{x}_{t,l} \tag{8} $$

$$ \boldsymbol{x}_{t,l} = W_l^{-H} \boldsymbol{s}_{t,l}. \tag{9} $$

Now, let us represent the m-th entry of $\boldsymbol{x}_{t,l}$ by $X_{t,l}^{(m)}$. (9) means that $S_{t,l}^{(1)}, \cdots, S_{t,l}^{(M)}$ are mixed together into $X_{t,l}^{(1)}, \cdots, X_{t,l}^{(M)}$ with frequency-band dependent mixing matrix W_l^{-H}. Then, by (8), these mixtures are further convolutively mixed via the multichannel AR system $G_l(z) = (I - \sum_{k=1}^{K_l} G_{k,l}^H z^{-k})^{-1}$. In this sense, the set of (8) and (9) defines the forward model that generates the observed spectral component vector $\boldsymbol{y}_{t,l}$ from the source spectral component vector $\boldsymbol{s}_{t,l}$. This forward model is depicted in Figure 1. We hereafter call W_l and $G_{k,l}$ a separation matrix and a regression matrix, respectively. Note that this model is an extension of the MIMO-AR model [11] to the case of complex numbers.

3.3 Maximum Likelihood Estimation

Based on the above setup, the probability density function (PDF) of an observed spectral component vector conditioned on its past sequence is given by

$$ p(\boldsymbol{y}_{t,l} | \boldsymbol{y}_{t-1,l}, \cdots, \boldsymbol{y}_{t-K_l,l}; \Theta) = \mathscr{N}_{\mathbb{C}}\Big\{ \boldsymbol{y}_{t,l}; \sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{y}_{t-k,l}, \Big(W_l \, {}_s\Lambda_t\Big(\frac{2\pi l}{L}\Big)^{-1} W_l^H \Big)^{-1} \Big\}, \tag{10} $$

where

$$ {}_s\Lambda_t(\omega) = \mathrm{diag}\{ {}_s\lambda_t^{(1)}(\omega), \cdots {}_s\lambda_t^{(M)}(\omega)\} \tag{11} $$

and Θ is the set of all parameters. Specifically, Θ is defined as

$$ \Theta = \{ {}_s\Theta, {}_w\Theta, {}_g\Theta\}, \tag{12} $$

where

$$ {}_s\Theta = \{a_{t,1}^{(m)}, \cdots, a_{t,P}^{(m)}, {}_sv_t^{(m)}\}_{1\le m\le M, 0\le t\le T-1} \quad (13) $$

$$ {}_w\Theta = \{W_l\}_{0\le l\le L-1} \quad (14) $$

$$ {}_g\Theta = \{\{G_{k,l}\}_{1\le k\le K_l}\}_{0\le l\le L-1}. \quad (15) $$

${}_s\Theta$, ${}_w\Theta$, and ${}_g\Theta$ are sets of source parameters, separation matrices, and regression matrices, respectively.

As described in Section 2, the goal of the blind separation and dereverberation task is to estimate separation matrices ${}_w\Theta$ and regression matrices ${}_g\Theta$, (recall that ${}_w\Theta$ and ${}_g\Theta$ collectively provide the same information as deconvolution-system parameters $\mathscr{W}$). On the other hand, the PDF of $\boldsymbol{y}_{t,l}$ depends on ${}_s\Theta$ as well as ${}_w\Theta$ and ${}_g\Theta$ as shown in (10). Therefore, we estimate all the parameters in Θ ($=\{{}_s\Theta, {}_w\Theta, {}_g\Theta\}$) by using the ML estimation method.

The ML estimation method maximizes the log likelihood function $\mathscr{L}(\Theta;\mathscr{Y}) = \log p(\mathscr{Y};\Theta)$ with respect to parameter set Θ. By using (10), the log likelihood function is written as

$$ \begin{aligned} \mathscr{L}(\Theta;\mathscr{Y}) = \sum_{l=0}^{L-1}\sum_{t=0}^{T-1}\Big\{ & \log\det\Big(W_l\,{}_s\Lambda_t\Big(\frac{2\pi l}{L}\Big)^{-1}W_l^H\Big) \\ & - \Big(\boldsymbol{y}_{t,l} - \sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{y}_{t-k,l}\Big)^H W_l\,{}_s\Lambda_t\Big(\frac{2\pi l}{L}\Big)^{-1}W_l^H \\ & \times\Big(\boldsymbol{y}_{t,l} - \sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{y}_{t-k,l}\Big)\Big\}. \end{aligned} \quad (16) $$

The proposed algorithm for maximizing (16) is described in Section 4.

4. PROPOSED ALGORITHM

We maximize log likelihood function (16) by using the coordinate ascent method. This method iteratively updates the estimate of Θ. Let ${}_s\hat{\Theta}^{(i)}$, ${}_g\hat{\Theta}^{(i)}$, and ${}_w\hat{\Theta}^{(i)}$ denote the estimates of ${}_s\Theta$, ${}_g\Theta$, and ${}_w\Theta$, respectively, after the i-th iteration. Each iteration comprises the following three maximization steps:

$$ {}_s\hat{\Theta}^{(i+1)} = \operatorname*{argmax}_{{}_s\Theta} \mathscr{L}(\Theta;\mathscr{Y})\big|_{{}_g\Theta={}_g\hat{\Theta}^{(i)}, {}_w\Theta={}_w\hat{\Theta}^{(i)}} \quad (17) $$

$$ {}_g\hat{\Theta}^{(i+1)} = \operatorname*{argmax}_{{}_g\Theta} \mathscr{L}(\Theta;\mathscr{Y})\big|_{{}_s\Theta={}_s\hat{\Theta}^{(i+1)}, {}_w\Theta={}_w\hat{\Theta}^{(i)}} \quad (18) $$

$$ {}_w\hat{\Theta}^{(i+1)} = \operatorname*{argmax}_{{}_w\Theta} \mathscr{L}(\Theta;\mathscr{Y})\big|_{{}_s\Theta={}_s\hat{\Theta}^{(i+1)}, {}_g\Theta={}_g\hat{\Theta}^{(i+1)}}. \quad (19) $$

Below, we derive algorithms for (17), (18), and (19).

4.1 Update of source parameters

Source parameters ${}_s\Theta$ are updated as follows. Since we now have the tentative estimates of regression matrices, ${}_w\Theta^{(i)} = \{\hat{W}_l^{(i)}\}_{0\le l\le L-1}$, and those of separation matrices, ${}_g\Theta^{(i)} = \{\{\hat{G}_{k,l}^{(i)}\}_{1\le k\le K_l}\}_{0\le l\le L-1}$, we can estimate source spectra $\mathscr{S} = \{S_{t,l}^{(m)}\}_{1\le m\le M, 0\le t\le T-1, 0\le l\le L-1}$ by substituting $\hat{W}_l^{(i)}$ and $\hat{G}_{k,l}^{(i)}$ into (8) and (9). We denote the estimate of $S_{t,l}^{(m)}$ by $\hat{S}_{t,l}^{(m)}$. By applying linear predictive analysis to $\{\hat{S}_{t,l}^{(m)}\}_{0\le l\le L-1}$ for each source m and frame t, we obtain the updated source parameters, ${}_s\hat{\Theta}^{(i+1)}$.

4.2 Update of regression matrices

It may be found that all regression matrices for the l-th frequency band, $G_{1,l}, \cdots, G_{K_l,l}$, that maximize log likelihood function (16) depend on each other. In order to treat them jointly, let us put all the entries of the regression matrices related to the l-th frequency band into a vector as

$$ \boldsymbol{g}_l = [\boldsymbol{g}_{1,l}, \cdots, \boldsymbol{g}_{K_l,l}]_{1\times M^2K_l} \quad (20) $$

$$ \boldsymbol{g}_{k,l} = [\boldsymbol{g}_{k,l}^{(1)T}, \cdots, \boldsymbol{g}_{k,l}^{(M)T}]_{1\times M^2}, \quad (21) $$

where $\boldsymbol{g}_{k,l}^{(m)}$ denotes the m-th column of $G_{k,l}$. By using $\boldsymbol{g}_l$, the term $\sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{y}_{t-k,l}$ in (16) is rewritten as

$$ \sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{y}_{t-k,l} = \boldsymbol{Y}_{t-1,l}\boldsymbol{g}_l^H, \quad (22) $$

where

$$ \boldsymbol{Y}_{t-1,l} = [\tilde{\boldsymbol{y}}_{t-1,l}, \cdots, \tilde{\boldsymbol{y}}_{t-K_l,l}]_{M\times M^2K_l} \quad (23) $$

$$ \tilde{\boldsymbol{y}}_{t-k,l} = \begin{bmatrix} \boldsymbol{y}_{t-k,l}^T & & O \\ & \ddots & \\ O & & \boldsymbol{y}_{t-k,l}^T \end{bmatrix}_{M\times M^2}. \quad (24) $$

By substituting (22) into (16) and then organizing the resultant equation with respect to $\boldsymbol{g}_l$, we obtain

$$ \begin{aligned} \mathscr{L}(\Theta;\mathscr{Y}) = & -\sum_{t=0}^{T-1}(\boldsymbol{y}_{t,l} - \boldsymbol{Y}_{t-1,l}\boldsymbol{g}_l^H)^H W_l\,{}_s\Lambda_t\Big(\frac{2\pi l}{L}\Big)^{-1} \\ & \times W_l^H(\boldsymbol{y}_{t,l} - \boldsymbol{Y}_{t-1,l}\boldsymbol{g}_l^H) + \text{constant}. \end{aligned} \quad (25) $$

$\hat{\boldsymbol{g}}_l^{(i+1)}$ that maximizes (25) under conditions ${}_s\Theta = {}_s\hat{\Theta}^{(i+1)}$ and ${}_w\Theta = {}_w\hat{\Theta}^{(i)}$ is easily obtained as

$$ \begin{aligned} (\hat{\boldsymbol{g}}_l^{(i+1)})^H = & \Big(\sum_{t=0}^{T-1}\boldsymbol{Y}_{t-1,l}^H\hat{W}_l^{(i)}\,{}_s\hat{\Lambda}_t^{(i+1)}\Big(\frac{2\pi l}{L}\Big)^{-1}(\hat{W}_l^{(i)})^H\boldsymbol{Y}_{t-1,l}\Big)^{-1} \\ & \times\Big(\sum_{t=0}^{T-1}\boldsymbol{Y}_{t-1,l}^H\hat{W}_l^{(i)}\,{}_s\hat{\Lambda}_t^{(i+1)}\Big(\frac{2\pi l}{L}\Big)^{-1}(\hat{W}_l^{(i)})^H\boldsymbol{y}_{t,l}\Big), \end{aligned} \quad (26) $$

where

$$ {}_s\hat{\Lambda}_t^{(i+1)}(\omega) = {}_s\Lambda_t(\omega)\big|_{{}_s\Theta={}_s\hat{\Theta}^{(i+1)}} \quad (27) $$

Vectors $\hat{\boldsymbol{g}}_0^{(i+1)}, \cdots, \hat{\boldsymbol{g}}_{L-1}^{(i+1)}$ constitute the updated estimate of ${}_g\Theta$, namely ${}_g\hat{\Theta}^{(i+1)}$.

4.3 Update of separation matrices

A necessary condition for W_l to maximize log likelihood function (16) is

$$ \begin{aligned} \frac{1}{2}\frac{\partial\mathscr{L}(\Theta;\mathscr{Y})}{\partial W_l} = & W_l^{-H} - \frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{x}_{t,l}\boldsymbol{x}_{t,l}^H W_l\,{}_s\Lambda_t\Big(\frac{2\pi l}{L}\Big)^{-1} \\ = & O. \end{aligned} \quad (28) $$

(28) is rewritten as

$$ \frac{1}{T}\sum_{t=0}^{T-1}\Big\{W_l^H\boldsymbol{x}_{t,l}\boldsymbol{x}_{t,l}^H W_l\,{}_s\Lambda_t\Big(\frac{2\pi l}{L}\Big)^{-1}\Big\} = I_M, \quad (29) $$

where I_M stands for the M-dimensional identity matrix. For any number of microphones M, (29) may be solved by using, for example, the quasi-Newton method. Instead, we here describe an alternative efficient algorithm for solving (29) for the stereo microphones, namely where $M = 2$.

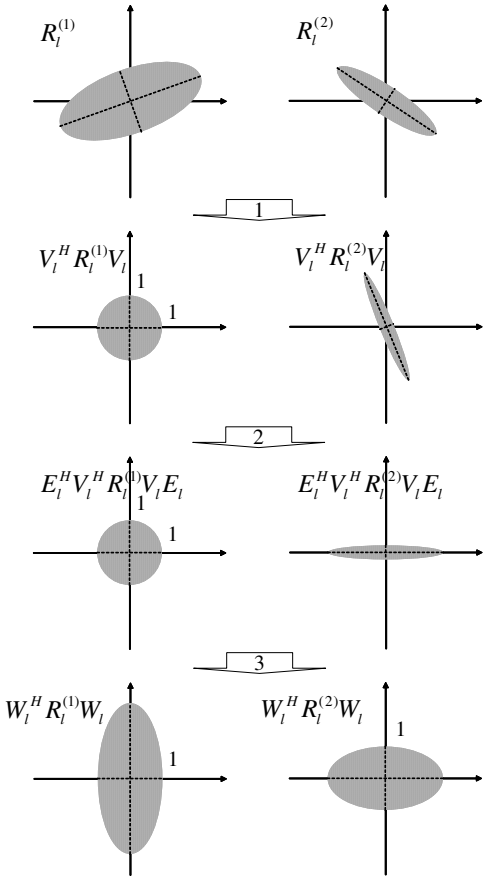

Figure 2: Diagram of algorithm for solving (32) and (33).

In order to simplify the notation, we define $\alpha_{t,l}^{(m)}$ as

$$\alpha_{t,l}^{(m)} = {}_s\lambda_t^{(m)}(2\pi l/L)^{-1} \tag{30}$$

Substituting (30) into (29) yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(W_l^H \boldsymbol{x}_{t,l}\boldsymbol{x}_{t,l}^H W_l \mathrm{diag}\{\alpha_{t,l}^{(1)}, \alpha_{t,l}^{(2)}\}\right) = I_2. \tag{31}$$

Condition (31) can be split into the following two conditions:

$$W_l^H R_l^{(1)} W_l = \begin{bmatrix} 1 & \star \\ 0 & \star \end{bmatrix} \tag{32}$$

$$W_l^H R_l^{(2)} W_l = \begin{bmatrix} \star & 0 \\ \star & 1 \end{bmatrix}, \tag{33}$$

where $\star$ stands for any number, and

$$R_l^{(1)} = \frac{1}{T}\sum_{t=0}^{T-1} \alpha_{t,l}^{(1)} \boldsymbol{x}_{t,l}\boldsymbol{x}_{t,l}^H \tag{34}$$

$$R_l^{(2)} = \frac{1}{T}\sum_{t=0}^{T-1} \alpha_{t,l}^{(2)} \boldsymbol{x}_{t,l}\boldsymbol{x}_{t,l}^H. \tag{35}$$

The set of equations (32) and (33) can be solved according to the following algorithm, which is also illustrated in Figure 2.

1. Calculate the inverse square root of $R_l^{(1)}$, which satisfies

$$V_l^H R_l^{(1)} V_l = I_2. \tag{36}$$

2. Find unitary matrix E_l such that

$$E_l^H V_l^H R_l^{(2)} V_l E_l = \mathrm{diag}\{d_l^{(1)}, d_l^{(2)}\}. \tag{37}$$

   Since $V_l^H R_l^{(2)} V_l$ is an Hermitian matrix, E_l is obtained via eigen-decomposition.
3. Then, we obtain W_l as

$$W_l = V_l E_l \begin{bmatrix} 1 & 0 \\ 0 & d_l^{(2)^{-\frac{1}{2}}} \end{bmatrix}. \tag{38}$$

The estimate of unmixing matrix W_l after the $(i+1)$-th iteration, $\hat{W}_l^{(i+1)}$, is obtained by letting ${}_s\Theta = {}_s\hat{\Theta}^{(i+1)}$ and ${}_g\Theta = {}_g\hat{\Theta}^{(i+1)}$ in (34) and (35).

After $\hat{W}_l^{(i+1)}$ is obtained for all l, the scales and permutations of the unmixing matrices are aligned. The scale alignment is performed based on the minimal distortion principle [12]. The permutation alignment is done by using the method described in [7].

Thus, the derivation of the proposed algorithm is completed. It is noteworthy that if we force regression order K_l to be 0 for any l, the proposed method reduces to a new frequency-domain BSS method based on second-order statistics (SOS).

5. EXPERIMENTAL RESULTS

We show three experimental results that demonstrate the performance of the proposed method. All experiments dealt with a two-source two-microphone case.

5.1 Anechoic mixing

In the first experiment, observed signals were instantanous mixtures of source signals. The objective is to validate the algorithm for optimizing unmixing matrices shown in Section 4.3.

In this experiment, the proposed method was tested in 100 trials. Each trial used a mixing matrix that was independently generated from a zero-mean unit-variance Gaussian distribution. The two source signals were male and female speech signals taken from the JNAS database. The sampling rate was 8 kHz and the signal length was 5 seconds. The parameters were set as follows: the STFT frame size was 256 points, the STFT frame shift was 128 points, the window function was a Hanning window, the regression order, K_l, was 0 for any l, the number of poles, P, was 12, and there were 3 iterations. Recall that setting K_l at 0 reduces the proposed method to a frequency-domain BSS method.

Improvement of signal to interference and noise ratio (SINR) [13] averaged over the 100 trials was 14.7 dB. This result indicates the validity of the proposed unmixing-matrix optimization algorithm.

5.2 Convolutive mixing of speech and noise

In the second and third experiments, the observed signals were synthesized by convolutively mixing two source signals by using room impulse responses measured in a room with an RT_{60} of 0.6 seconds. The objective of these experiments is to evaluate the overall performance of the proposed method.

The second experiment used a male utterance and a pink noise as source sounds. The parameters were common to those of the first experiment except that the regression order, K_l, was set at 40 here.

Figure 3 shows the waveforms of the clean speech, the observed speech, and the estimated speech. The speech waveform estimated with the conventional frequency-domain BSS method reported in [7] is also plotted. The conventional frequency-domain BSS method used a 1024-point Hanning window with 256-point overlap. We see that the proposed method canceled out a larger amount of the interfering noise than the conventional method. This is because the proposed method can reduce the reverberant components of speech

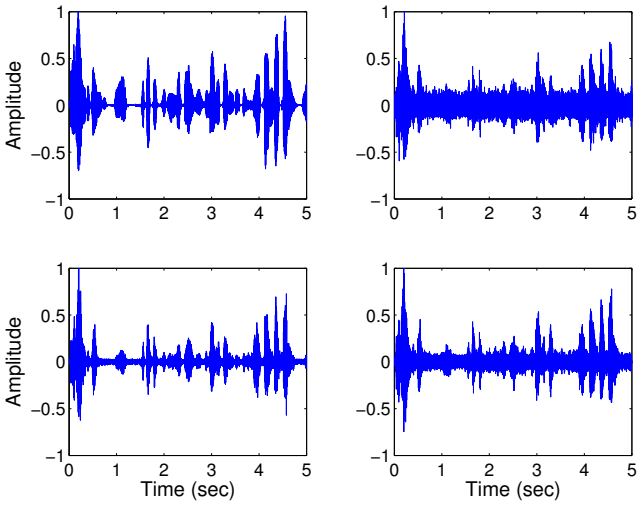

Figure 3: (Upper left) Clean speech waveform. (Upper right) Observed speech waveform. (Lower left) Speech waveform estimated with proposed method. (Lower right) Speech waveform estimated with conventional method.

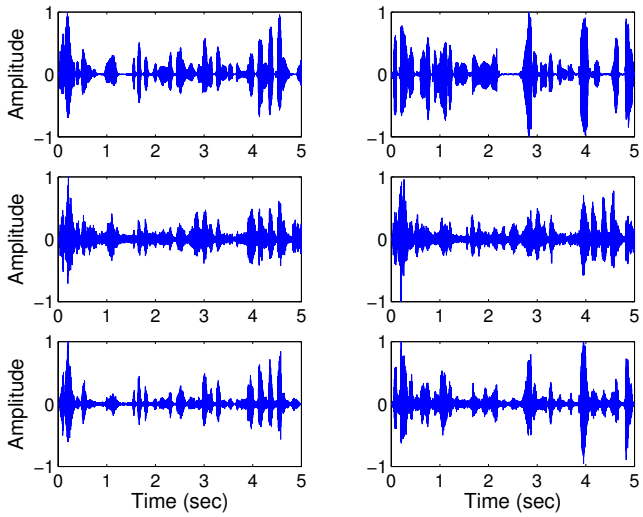

Figure 4: (Top) Clean waveforms of male speech (left) and female speech (right). (Middle) Observed speech waveforms. (Bottom) Estimated waveforms of male speech (left) and female speech (right).

and noise, which are never canceled out without using a deconvolution system. Indeed, the speech estimated with the proposed method sounded less reverberant.

5.3 Convolutive mixing of two speeches

In the third experiment, the source signals were the male and female speech signals that were used in the first experiemnt. The other conditions were common to those of the second experiment.

Figure 4 shows the waveforms of the source signals, the observed signals, and the estimates of the source signals. We find that the male speech was estimated with a high degree of accuracy. On the other hand, the estimated female speech signal still contained an male speech interfering component, though the magnitude of the male speech was reduced to some extent. This may be attributed to limiting the deconvolution system to a causal system. A thorough evaluation and further improvement of the proposed method will be included in future work.

6. CONCLUSION

This paper described a method for the blind separation and dereverberation of convolutive audio mixtures. The proposed method was derived in the framework of the ML estimation method. The TVAP source model was introduced to derive the likelihood function. The optimization algorithm consisted of three analytically calculated optimization steps. The proposed method yielded promising experimental results.

REFERENCES

[1] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 2, pp. 109–116, 2003.

[2] K. Kokkinakis and A. K. Nandi, "Multichannel blind deconvolution for source separation in convolutive mixtures of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 200–212, 2006.

[3] S. Douglas, H. Sawada, and S. Makino, "Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 1, pp. 92–104, 2005.

[4] S. Araki, S. Makino, R. Aichner, T. Nishikawa, and H. Saruwatari, "Subband-based blind separation for convolutive mixtures of speech," *IEICE Trans. Fund.*, vol. E88-A, no. 12, 2005.

[5] T. Yoshioka, T. Hikichi, and M. Miyoshi, "Dereverberation by using time-variant nature of speech production system," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007, article ID 65698, 15 pages, doi:10.1155/2007/65698.

[6] T. Nakatani, B. H. Juang, T. Yoshioka, K. Kinoshita, and M. Miyoshi, "Importance of energy and spectral features in Gaussian source model for speech dereverberation," in *Proc. IEEE Worksh. Appl. Signal Process. Audio, Acoust.*, 2007, pp. 299–302.

[7] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Int'l Symp. Circ., Syst.*, 2007, pp. 3247–3250.

[8] A. van den Bos, "The multivariate complex normal distribution–A generalization," *IEEE Trans. Inform. Theory*, vol. 41, no. 2, pp. 537–539, 1995.

[9] S. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Comm.*, vol. 39, no. 1, pp. 65–78, 2003.

[10] H. Buchner, R. Aichner, and W. Kellerman, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, 2005.

[11] T. Routtenberg and J. Tabrikian, "MIMO-AR system identification and blind source separation using GMM," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2007, pp. 761–764.

[12] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Int'l Conf. Ind. Comp. Anal., Blind Signal Sep.*, 2001, pp. 722–727.

[13] S. Douglas, M. Gupta, H. Sawada, and S. Makino, "Spatio-temporal FastICA algorithms for the blind separation of convolutive mixtures," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 15, no. 5, pp. 1511–1520, 2007.