SEPARATION MATRIX OPTIMIZATION USING ASSOCIATIVE MEMORY MODEL FOR BLIND SOURCE SEPARATION

Motoi Omachi[*], *Tetsuji Ogawa*[*], *Tetsunori Kobayashi*[*], *Masaru Fujieda*[†], *Kazuhiro Katagiri*[†]

[*] Department of the computer science, Waseda University, Japan
[†] Oki Electric Industry Co., Ltd., Japan.

ABSTRACT

A source signal is estimated using an associative memory model (AMM) and used for separation matrix optimization in linear blind source separation (BSS) to yield high quality and less distorted speech. Linear-filtering-based BSS, such as independent vector analysis (IVA), has been shown to be effective in sound source separation while avoiding non-linear signal distortion. This technique, however, requires several assumptions of sound sources being independent and generated from non-Gaussian distribution. We propose a method for estimating a linear separation matrix without any assumptions about the sources by repeating the following two steps: estimating non-distorted reference signals by using an AMM and optimizing the separation matrix to minimize an error between the estimated signal and reference signal. Experimental comparisons carried out in simultaneous speech separation suggest that the proposed method can reduce the residual distortion caused by IVA.

Index Terms— convolutional neural network, denoising autoencoder, associative memory model, linear filtering, blind source separation

1. INTRODUCTION

Time-frequency (TF) masking [1–3] and linear filtering [4–6] are frequently used on multichannel blind source separation (BSS). TF masking estimates a target source by using a non-linear mask that passes only the TF components where the target source dominantly exists. TF masking has been shown to be effective in source separation but yields unexpected harmful non-linear distortions, which degrade the sound quality of separated signals. Linear filtering, on the other hand, estimates sound sources by using a linear separation matrix that compensates for the effect of the mixing process of multiple sound sources. This method can be advantageous over TF masking in terms of the quality of the separated sounds because non-linear distortion, inevitable in TF masking, can be avoided in linear filtering-based methods.

In general, the separation matrix for linear BSS is estimated under the assumption that sound sources are statistically independent and generated from a non-Gaussian distribution. Independent component analysis (ICA) [4] and independent vector analysis (IVA) [5] are well-known strategies of this approach. These strategies, however, can degrade the performance of the separation when the sound sources are not independent. In addition, they do not take account of a property of the sound sources explicitly. To handle the problem, many attempts to use a prior distribution suitable for expressing the property have been made [7, 8]. More consideration, however, is needed to overcome the assumption of the sound sources being independent.

We propose a separation matrix optimization method using a neural network-based associative memory model (AMM) to achieve high-performance BSS without any assumptions of sound sources. The proposed method optimizes the separation matrix to minimize the error between the distorted signal and corresponding reference signal, which is estimated as an ideal (non-distorted) signal by using the AMM. An alternating optimization is adopted: AMM-based reference signal estimation using the current separation matrix and separation matrix optimization using the reference signal estimated in the previous step. Note that the proposed method does not require any assumptions of sound sources that are necessary in ICA and IVA. In addition, less distortion in the separated sounds can be expected because the proposed method is based on linear filtering.

We used a denoising auto-encoder (DAE) [9] as an AMM. A DAE can estimate a non-distorted signal from an input with distortion and has been shown to be effective in dereverberation [10] and denoising [11]. The separated signal generally contains distortion attributed to a residual of the interference source and over-subtraction. A DAE-based AMM therefore may be useful for reducing such distortion and yielding the reference signal in the proposed method.

The spectra estimated using a DAE are known to be over-smoothed. The DAE output therefore is not directly used as an estimate of the target source but has been used in filter estimation, e.g., design of the Wiener filter [12] and TF mask [13]. In contrast, we attempt to use the DAE output to estimate the linear separation matrix for BSS.

The rest of the present paper is organized as follows. In Section 2, we review linear filtering-based BSS. In Section 3, we describe an algorithm of the proposed method. In Section 4, we discuss experimental comparisons in sound source

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

separation to verify the effectiveness of the proposed method. Finally, we present our conclusion in Section 5.

2. LINEAR FILTERING-BASED BSS

This section briefly describes the method of estimating the separation matrix for linear filtering-based BSS. The aim of the BSS is to estimate an inverse filter to eliminate the effect of the mixing process of multiple sound sources when this process is unknown. Assume that N_s sources are estimated from N_m observations under the condition of $N_s \leq N_m$. The mixing process in the frequency domain is written as $\mathbf{Z}(\omega,\tau) = \mathbf{H}(\omega)\mathbf{S}(\omega,\tau)$, where ω and τ denote the discrete frequency and frame index, respectively; $\mathbf{S}(\omega,\tau) \in \mathbb{C}^{N_s\times 1}$ denotes source signals, and $\mathbf{Z}(\omega,\tau) \in \mathbb{C}^{N_m\times 1}$ and $\mathbf{H}(\omega) \in \mathbb{C}^{N_m\times N_s}$, observed signals and a mixing matrix (i.e., transfer function). By applying the separation matrix $\mathbf{W}(\omega) \in \mathbb{C}^{N_s\times N_m}$ to $\mathbf{Z}(\omega,\tau)$, the outputs $\mathbf{Y}(\omega,\tau) \in \mathbb{C}^{N_s\times 1}$, which are the estimates of source spectra, are obtained as $\mathbf{Y}(\omega,\tau) = \mathbf{W}(\omega)\mathbf{Z}(\omega,\tau) = \mathbf{W}(\omega)\mathbf{H}(\omega)\mathbf{S}(\omega,\tau)$.

In BSS, $\mathbf{W}(\omega)$ is estimated such that $\mathbf{Y}(\omega,\tau)$ could be statistically independent. Particularly, IVA can yield an advantage over ICA in avoidance of the permutation problem. IVA estimates $\mathbf{W}(\omega)$ by maximizing the independence between the vectors consisting of spectral components of the estimated source. As a multivariate distribution, assumed as a prior of the sound source, takes into account the relation among frequencies, IVA is free from the permutation problem. Separation results, however, can be degraded when the sound sources are not independent. In addition, the signal estimated on the basis of source independency is not always consistent with the signal of the original source.

3. SEPARATION MATRIX ESTIMATION USING AMM

We attempt to associate a distorted speech with the corresponding non-distorted speech using a neural network-based AMM and apply the estimate of the non-distorted signal to optimization of the separation matrix. In the present study, a DAE was used on the AMM and the estimate of the non-distorted signal was referred to as a reference signal. The separation matrix $\mathbf{W}(\omega)$ was estimated by minimizing the error between the separation output $\mathbf{Y}(\omega,\tau)$, which generally contains a residual distortion, and the reference signal $\hat{\mathbf{S}}(\omega,\tau)$. Figure 1 shows the proposed method. The method adopts an alternating optimization approach that repeats the following two steps: estimating the reference $\hat{\mathbf{S}}(\omega,\tau)$ using $\mathbf{Y}(\omega,\tau)$ obtained through the current $\mathbf{W}(\omega)$; and optimizing $\mathbf{W}(\omega)$ on the basis of the error $J(\omega)$ between $\hat{\mathbf{S}}(\omega,\tau)$ estimated in the previous step and $\mathbf{Y}(\omega,\tau)$. These two steps are repeated until $J(\omega)$ is converged. The rest of this section describes the details regarding reference spectrum estimation and separation matrix optimization.

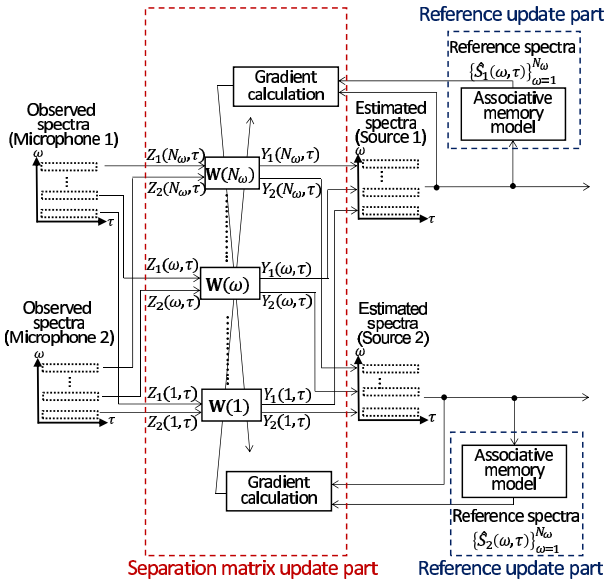


Fig. 1. *Schematic diagram of proposed method when two sources are estimated from two observations. System developed consists of reference spectrum estimation and separation matrix optimization.*

3.1. Reference signal estimation using AMM

A convolutional neural network (CNN) [14] is applied to a DAE-based AMM that associates a signal containing a distortion with a non-distorted signal. Figure 2 illustrates a DAE using a CNN. A CNN has a suitable structure to extract local features from a TF pattern of a speech spectrum. In addition, the distortion attributed to separation processing locally exists on the TF pattern. We therefore expect that a DAE based on a CNN can make it possible to eliminate such local distortions.

Partial TF patterns of 513 bins×10 frames (8000 Hz×160 ms) are extracted from a logarithmic power spectrum of distorted speech at intervals of five frames. These patterns are normalized such that the mean and standard deviation would take zero and one, respectively, then used for the inputs of the CNN. Each unit on the convolutional layer is obtained by applying 94 convolutional filters of 30 bins×5 frames (468.8 Hz×80 ms) over the input TF pattern, where the convolutional window is shifted by 15 bins (234.4 Hz) and 2 frames (32 ms). Principle component analysis (PCA) is carried out using all the partial TF patterns of 30 bins×5 frames of training data. Then 94 eigenvectors of 30 bins×5 frames are used as the convolutional filters, where the top 94 eigenvalues yield the cumulative contribution ratio of 95%. The convolutional layer detects the same local pattern in the different position of the TF pattern. The bottleneck layer extracts higher-order features expressing the relation among different positions of local patterns detected on the convolutional layer. The dimensionality of the bottleneck features is 2,253. To determine this dimensionality, PCA is applied to all the 9,306-dimensional convolutional layer outputs of the training data and the resulting dimensionality is determined such that

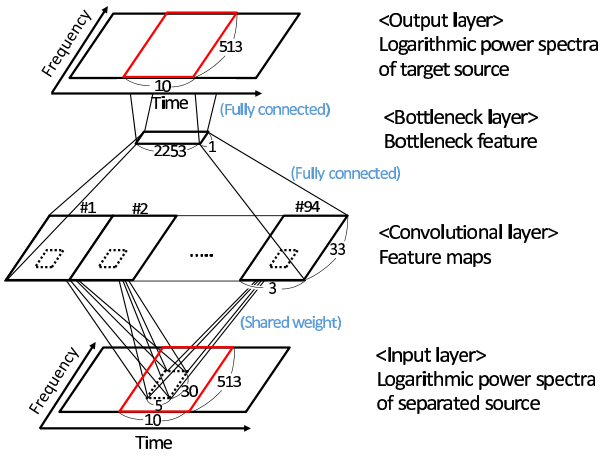


Fig. 2. *Denoising auto-encoder-based associative memory model. Logarithmic power spectra of target sources are estimated from those of separated signal containing distortion.*

the cumulative contribution ratio can be more than 95%. In the output layer, the spectral estimates of 513 bins×10 frames are obtained.

The weights and biases in the CNN are determined by pre-training based on the auto-encoder followed by fine-tuning. Early stopping is conducted with development data to avoid over-fitting. The CNN is trained with both of two types of data pairs as follows:

- **clean-clean:** the non-distorted speech is used for both input and supervisory signals.
- **separated-clean:** the separated speech obtained by applying IVA to the simultaneous speech is used for the input signal and those of the corresponding non-distorted signal are used for the supervisory signal.

3.2. Separation matrix optimization

We estimate a linear projection matrix $\mathbf{M}(\omega) \in \mathbb{C}^{N_s \times N_s}$ for eliminating the distortion included in $\mathbf{Y}(\omega,\tau)$, yielding $\bar{\mathbf{Y}}(\omega,\tau) = \mathbf{M}(\omega)\mathbf{Y}(\omega,\tau)$. It is ideal that $\bar{\mathbf{Y}}(\omega,\tau)$ be consistent to the target source. The reference signal $\hat{\mathbf{S}}(\omega,\tau)$ therefore is used as $\bar{\mathbf{Y}}(\omega,\tau)$ for estimating $\mathbf{M}(\omega)$. In addition, the separation matrix estimated through IVA, $\mathbf{W}^{(\mathrm{IVA})}(\omega)$, is used as the initial value for AMM-based optimization. In this case, $\bar{\mathbf{Y}}(\omega,\tau)$ is written as

$$\begin{aligned}\bar{\mathbf{Y}}(\omega,\tau) &= \mathbf{M}(\omega)\mathbf{Y}(\omega,\tau) = \mathbf{M}(\omega)\mathbf{W}^{(\mathrm{IVA})}(\omega)\mathbf{Z}(\omega,\tau) \\ &= \bar{\mathbf{W}}(\omega)\mathbf{Z}(\omega,\tau), \end{aligned} \quad (1)$$

where $\bar{\mathbf{W}}(\omega) = \mathbf{M}(\omega)\mathbf{W}^{(\mathrm{IVA})}(\omega)$. Since $\mathbf{M}(\omega)\mathbf{W}^{(\mathrm{IVA})}(\omega)$ is a linear transformation, Eq. (1) represents linear BSS, i.e., the linear separation matrix $\bar{\mathbf{W}}(\omega)$ is applied to the observed signals $\mathbf{Z}(\omega,\tau)$. The updated $\bar{\mathbf{W}}(\omega)$ is obtained by estimating $\mathbf{M}(\omega)$ followed by calculating $\mathbf{M}(\omega)\mathbf{W}^{(\mathrm{IVA})}(\omega)$. Here, $\mathbf{M}(\omega)$ is estimated to minimize the error between $\mathbf{Y}(\omega,\tau)$ and $\hat{\mathbf{S}}(\omega,\tau)$. The error function is described as

$$J(\omega) = \sum_{\tau=1}^{N_\tau} \left\| \log|\hat{\mathbf{S}}(\omega,\tau)|^2 - \log|\mathbf{M}(\omega)\mathbf{Y}(\omega,\tau)|^2 \right\|^2, \quad (2)$$

where N_τ denotes the number of frames. Then $\mathbf{M}(\omega)$ is optimized using a gradient descent method as

$$\mathbf{M}^{(j+1)}(\omega) \leftarrow \mathbf{M}^{(j)}(\omega) - \mu \frac{\mathbf{G}(\omega)}{\|\mathbf{G}(\omega)\|}, \quad (3)$$

where $\mathbf{G}(\omega) = \partial J(\omega)/\partial \mathbf{M}^*(\omega)$, and μ, j and * denote the learning rate, index of the update and complex conjugate operator, respectively. The optimization algorithm used is summarized in Algorithm 1.

Algorithm 1 Algorithm for separation matrix optimization.

Require: Observed signal $\mathbf{Z}(\omega,\tau)$
Require: Initial separation matrix $\mathbf{W}^{(\mathrm{IVA})}(\omega)$
Require: #epochs for reference signal update N_R, #epochs for filter update N_M, learning rate μ

$\mathbf{M}^{(0)}(\omega) = \mathbf{I}$ ($\mathbf{I}$: identity matrix).
$\mathbf{Y}^{(0)}(\omega,\tau) = \mathbf{W}^{(\mathrm{IVA})}(\omega)\mathbf{Z}(\omega,\tau)$.
Estimate $\hat{\mathbf{Y}}^{(0)}(\omega,\tau)$ from $\mathbf{Y}^{(0)}(\omega,\tau)$ using AMM.
for $i = 0 : N_R$-1
  for $j = 0 : N_M$-1
    Calculate gradient $\mathbf{G}^{(j)}(\omega)$ using $\hat{\mathbf{Y}}^{(i)}(\omega,\tau)$ and $\mathbf{Y}^{(i)}(\omega,\tau)$.
    $\mathbf{M}^{(j+1)}(\omega) = \mathbf{M}^{(j)}(\omega) - \mu\mathbf{G}^{(j)}(\omega)/||\mathbf{G}^{(j)}(\omega)||$.
  end for
  $\bar{\mathbf{M}}(\omega) = \mathbf{M}^{(N_M)}(\omega)$.
  $\mathbf{Y}^{(i+1)}(\omega,\tau) = \bar{\mathbf{M}}(\omega)\mathbf{W}^{(\mathrm{IVA})}\mathbf{Z}(\omega,\tau)$.
  Estimate $\hat{\mathbf{Y}}^{(i+1)}(\omega,\tau)$ from $\mathbf{Y}^{(i+1)}(\omega,\tau)$ using AMM.
  $\mathbf{M}^{(0)}(\omega) = \bar{\mathbf{M}}(\omega)$.
end for

Output: $\bar{\mathbf{W}}(\omega) = \bar{\mathbf{M}}(\omega)\mathbf{W}^{(\mathrm{IVA})}(\omega)$.

4. SOURCE SEPARATION EXPERIMENT

Experimental comparisons were conducted in simultaneous speech separation to verify the effectiveness of the proposed method.

4.1. Evaluation items

The BSS methods we evaluated are as follows:

- **IVA**: auxiliary-function-based IVA [6]
- **IVA-AMM (Proposed)**: separation matrix optimization using an AMM in which the matrix is initialized by **IVA**

The signal-to-interference ratio (SIR) and signal-to-distortion ratio (SDR) are used for speech quality measures. The SIR represents the ratio of the target source components to the components which is not included in the target source out of the separated speech components. The SDR

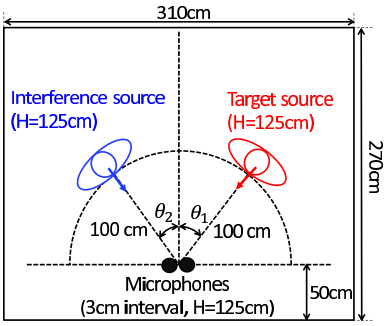


Fig. 3. *Experimental environment.*

represents the ratio of the target source components to the distortion caused in separation processing. The SIR and SDR are calculated as follows:

$$\text{SIR [dB]} = \frac{1}{N_s N_\omega}\sum_{i=1}^{N_s}\sum_{\omega=1}^{N_\omega} 10\log_{10}\frac{\sum_{\tau=1}^{N_\tau}|S_i(\omega,\tau)|^2}{\sum_{j=1}^{N_s}\sum_{\tau=1}^{N_\tau}(1-\delta_{ij})|Y_{ij}(\omega,\tau)|^2}$$

$$\text{SDR [dB]} = \frac{1}{N_s N_\omega}\sum_{i=1}^{N_s}\sum_{\omega=1}^{N_\omega} 10\log_{10}\frac{\sum_{\tau=1}^{N_\tau}|S_i(\omega,\tau)|^2}{\sum_{\tau=1}^{N_\tau}(|S_i(\omega,\tau)|-|Y_{ii}(\omega,\tau)|)^2},$$

where $S_i(\omega,\tau)$ denotes the i-th source component; $Y_{ij}(\omega,\tau)$ denotes the estimate of the i-th source when only the j-th source exists; N_ω denotes the number of frequency bins; and δ_{ij} returns one when $i=j$ and zero when $i\neq j$.

The reference signal estimation and separation matrix optimization were repeated up to 30 and 5000 times, respectively. The learning rate at the optimization step in (3) and that in training the CNN were initialized by 0.0001 and 0.01, respectively. The new-bob learning rate strategy was applied to both stages. In training the CNN, the size of the mini-batch was 100.

4.2. Speech material

Speech materials for evaluation and training were generated by convoluting a dry source with an impulse response. Figure 3 illustrates the acoustic environment used. Table 1 lists the directions of the target and interference sources, θ_1 and θ_2, used to yield the training, development, and test sets. The signals were assumed to be observed at a microphone array with two microphones spaced 3 cm apart. The impulse response was recorded in the environment shown in Fig. 3 to yield the test set and obtained by delay-based approximation to yield the training and development sets. In this experiment, the effect of the reflection and reverberation was removed from the impulse response.

4.2.1. Evaluation

Thirty utterances spoken by ten females (three utterances for each) were randomly selected from the Japanese newspaper article sentence database for target and interference speech, convoluted with the impulse response, then overlapped, yielding simultaneous speech for evaluation.

Table 1. Direction of target and interference sources.

data set	**source direction of** (θ_1,θ_2)
training	(-15,15), (-45,45), (-75,75), (-90,90)
development	(-60,60)
testing	(-30,30), (-30,0), (0,-30), (0,30), (30,0), (30,-30)

Table 2. Averaged SIR and averaged SDR.

Source angle	**SIR [dB]**		**SDR [dB]**	
(θ_1,θ_2) **[deg.]**	**IVA**	**IVA-AMM**	**IVA**	**IVA-AMM**
(-30,30)	32.8	32.8	10.1	11.7
(-30,0)	30.6	30.7	9.4	9.8
(0,-30)	30.2	30.0	9.2	10.8
(0,30)	28.0	27.5	8.4	8.6
(30,0)	26.7	26.4	8.0	8.7
(30,-30)	29.3	29.2	10.0	11.4

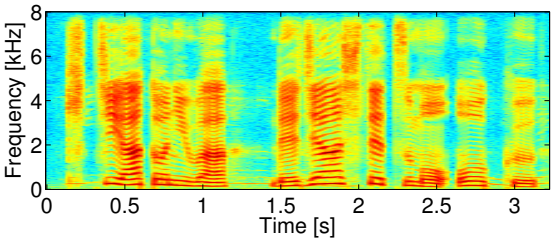
4.2.2. CNN training

A CNN was trained on both the **clean-clean** and **separated-clean** data pairs. Non-distorted speech utterances spoken by four females in a phoneme-balanced sentence database were used on the input and supervisory signals for training the CNN. Those utterances were also used for the dry sources to yield the simultaneous speech signals. Note that the speakers in the training and development sets are different from those used in the test set. **Clean-clean** consists of 1,800 utterances spoken by four females (450 utterances for each). The separated speech for **separated-clean** was obtained by applying IVA to the utterances simultaneously spoken by two speakers out of the four speakers. Four female speakers yield 12 pairs of speakers. In this case, nine pairs of speakers and one pair of speakers were used on the training and development sets, respectively. For the training set, 3,600 utterances of separated speech (9 pairs $\times$ 2 speakers $\times$ 50 utterances for each speaker $\times$ 4 conditions) were used for the input signals and the corresponding 3,600 non-distorted utterances were used for the supervisory signals. For the development set, 104 utterances of separated speech (1 pair $\times$ 2 speakers $\times$ 52 utterances for each speaker $\times$ 1 condition) were used for the input signals.

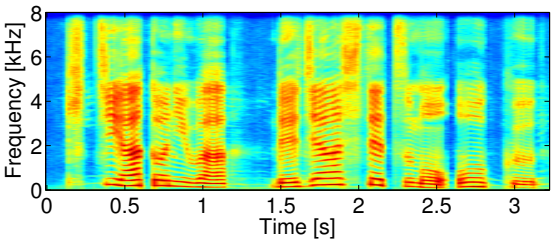
4.3. Experimental results

Table 2 lists the SIR and SDR averaged over 30 utterances. The SIRs of **IVA-AMM** were similar to those of **IVA**. These results indicate that in the present experiment, **IVA** was good enough to eliminate the interference components; therefore, **IVA-AMM** could not demonstrate its effectiveness. The **IVA-AMM**, on the other hand, out performed **IVA** in terms of the SDR, irrespective of source locations. This suggests that **IVA-AMM** can reduce the residual distortion attributed to **IVA** and estimate the sound sources with high accuracy.

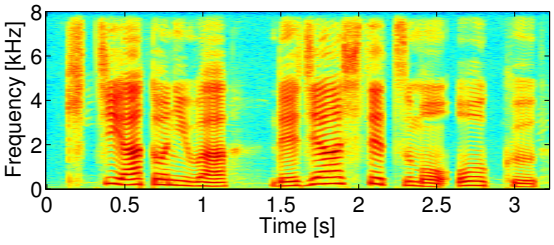
Figure 4 shows the specta of the (a) separated signal from



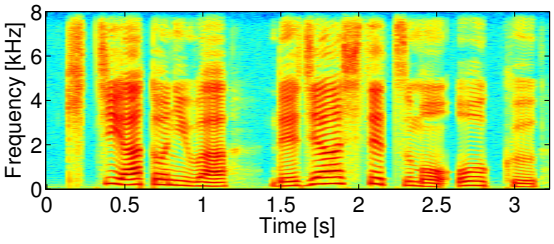
(a) Speech spectra estimated with IVA.



(b) Reference spectra estimated with CNN.



(c) Speech spectra estimated with proposed method.



(d) Original speech spectra.

Fig. 4. *Example of logarithmic power spectra.*

IVA, (b) reference signal, (c) estimated signal from **IVA-AMM**, and (d) target signal (ground truth). We can see that **IVA** processing yielded unexpected harmful spectral distortion over a range from 1.2 to 1.5 s and from 0 to 2 kHz (a); such distortion does not appear in the reference spectrum estimated using the AMM (b); and the spectrum obtained using **IVA-AMM** (c) was close to the target source spectrum (d).

5. CONCLUSION

We achieved high-quality BSS by incorporating AMM-based reference signal estimation in separation matrix optimization. The proposed method optimizes the separation matrix such that the estimates could be brought close to the ideal non-distorted signal estimated using the AMM and can accurately estimate original sources without any assumptions of the sources necessary in conventional ICA and IVA-based BSS. Experimental comparisons carried out in simultaneous speech separation demonstrated that the proposed method reduced the distortion caused by IVA.

REFERENCES

[1] H. Sawada *et al.*, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, March 2011.

[2] S. Araki *et al.*, "Simultaneous clustering of mixing and spectral model parameters for blind sparse source separation," in *Proc. ICASSP2010*, March 2010, pp. 5–8.

[3] A. Alinaghi *et al.*, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *Proc. ICASSP2011*, May 2011, pp. 209–212.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, Nov. 1998.

[5] T. Kim *et al.*, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, Jan. 2007.

[6] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA2011*, Oct. 2011, pp. 189–192.

[7] I. Lee *et al.*, "Adaptive independent vector analysis for the separation of convoluted mixtures using em algorithm," in *ICASSP2008*, March 2008, pp. 145–148.

[8] Y. Liang *et al.*, "Independent vector analysis with a generalized multivariate gaussian source prior for frequency domain blind source separation," *Signal Processing*, vol. 105, pp. 175–184, 2014.

[9] P. Vincent *et al.*, "Extracting and composing robust features with denoising autoencoders," in *Proc. ICML2008*, March 2008, pp. 1096–1103.

[10] T. Ishii *et al.*, "Reverberant speech recognition based on denoising autoencoder.," in *Proc. INTERSPEECH2013*, Aug. 2013, pp. 3512–3516.

[11] Y. Xu *et al.*, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Proc. Let.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[12] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Commun.*, vol. 60, pp. 13 – 29, 2014.

[13] P.-S. Huang *et al.*, "Deep learning for monaural speech separation," in *Proc. ICASSP2014*, May 2014, pp. 1562–1566.

[14] L. Deng *et al.*, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *Proc. ICASSP2013*, May 2013, pp. 6669–6673.