Encoder Re-training with Mixture Signals on FastMVAE Method

Shuhei Yamaji*, Taishi Nakashima*, Nobutaka Ono*, Li Li† and Hirokazu Kameoka†
* Tokyo Metropolitan University, Tokyo, Japan
† NTT Communication Science Laboratories, Kanagawa, Japan.

***Abstract*—In this paper, we propose a new network training to improve the source separation performance of the fast multichannel variational autoencoder (FastMVAE) method. The FastMVAE method is very effective for supervised source separation. It also significantly reduces the processing time by replacing the backpropagation steps in the MVAE method with a single forward propagation of the encoder for estimating latent variables. In previous studies, the encoder is trained together with the decoder using clean speech. In contrast, in this study, we re-train only the encoder by using the mixed signals with the decoder fixed. More specifically, using the imperfectly separated signals obtained in the process of the source separation algorithm, we train the encoder to find the optimal latent variables that minimize the objective function for source separation. Experimental results show that the proposed method reduces the objective function at almost every iteration and achieves higher separation performance than the conventional method.**

I. Introduction

Multichannel blind source separation is a promising technique to extract each source signal from a mixture. It has been remarkably developed in the last two decades [1], and frequency-domain independent component analysis (ICA) [2], independent vecotr analysis (IVA) [3], [4], auxiliary function based IVA (AuxIVA) [5], and independent low-rank matrix analysis (ILRMA) [6] have been proposed. Some of these methods have a source model representing a source signal, and the parameters of the source model and the demixing matrix are alternately updated to reduce the same objective function.

Recently, multichannel BSS methods using deep neural networks (DNNs) has been actively investigated [7]–[9]. The high expressive power of DNNs suggests that they may be effective in separating speech signals. In particular, the multichannel variational autoencoder (MVAE) [10], which introduces a variational autoencoder (VAE) as a source model, is effective in supervised source separation tasks. MVAE is a method that uses the decoder distribution of Conditional VAE (CVAE) as a source model for each sound source. CVAE is trained with the spectrogram of a single speaker signal and the corresponding speaker ID as condition class variables. This training allows CVAE to be used as a speech source model for the speakers in the training data. During source separation, iterative projection (IP) updates the demixing matrix. And, backpropagation of the CVAE decoder is used to update the source model, searching for latent variables that maximize the independence of the separated signals. In this time, if the step size can be adjusted appropriately, monotonic non-decreasing and convergence of the likelihood function in the update are guaranteed.

However, each iterative update requires a long processing time for backpropagation. This makes MVAE unsuitable for real-time processing. To address this, a fast algorithm called the FastMVAE [11] has subsequently been proposed. The idea is to replace the process of updating the latent variables with the forward propagation of the trained encoder. This replacement eliminates the time-consuming backpropagation steps and significantly reduces processing time. However, the encoder output is only an approximation of the update destination that decreases the objective function. Therefore, the objective function is not guaranteed to monotonically decrease and converge. For this reasons, the literature [11] reports that while FastMVAE is indeed faster than MVAE, it falls short of MVAE in separation performance.

To solve this problem, we propose a method to train the encoder to find latent variables that decrease the objective function during the source separation algorithm, while following the idea of updating latent variables using only the forward propagation of the encoder. Specifically, after training the network (encoder and decoder) in conventional FastMVAE, only the encoder is re-trained using the likelihood function as the loss function. Experimental results show that the network trained by the proposed method reduces the decrease of the likelihood function and achieves higher separation performance than the conventional method.

II. Blind Source Separation

A. Problem formulation

Let N be the number of sound sources and microphones. The speech signal, the mixed signal, and the separated signal at each time-frequency are denoted as follows

$$\boldsymbol{s}_{f,t} = (s_{f,t,1}, \ldots, s_{f,t,n}, \ldots, s_{f,t,N})^{\mathrm{T}} \in \mathbb{C}^N, \quad (1)$$

$$\boldsymbol{x}_{f,t} = (x_{f,t,1}, \ldots, x_{f,t,n}, \ldots, x_{f,t,N})^{\mathrm{T}} \in \mathbb{C}^N, \quad (2)$$

$$\boldsymbol{y}_{f,t} = (y_{f,t,1}, \ldots, y_{f,t,n}, \ldots, y_{f,t,N})^{\mathrm{T}} \in \mathbb{C}^N, \quad (3)$$

where $f = 1, \ldots, F$, $t = 1, \ldots, T$ and $n = 1, \ldots, N$ denote the frequency bins, time frames, and sources or channel indices, respectively. $^{\mathrm{T}}$ indicates transposition. Let $\boldsymbol{S}_n \in \mathbb{C}^{F \times T}$ be the complex spectrogram matrix of the nth speech signal, $s_{f,t,n}$ over all time frequencies, and let $\boldsymbol{X}_n \in \mathbb{C}^{F \times T}$ be the complex spectrogram matrix of the mixed signal as well. Assuming that the mixing system is linear time-invariant and is

978-616-590-477-3 ©2022 APSIPA

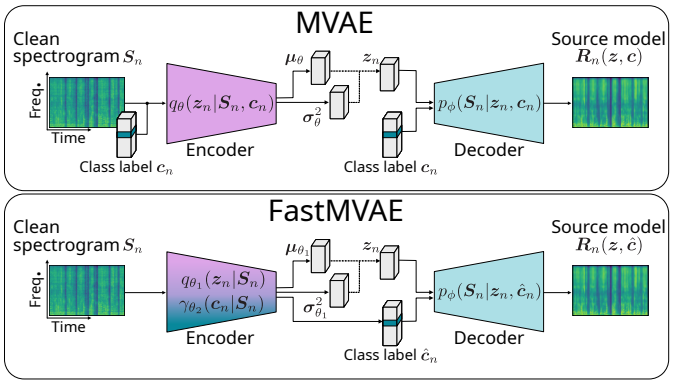

Fig. 1. Overview of MVAE method (upper) and FastMVAE method (lower).

a complex instantaneous mixing in the time-frequency domain, the separated signal can be estimated by $\boldsymbol{y}_{f,t} = \boldsymbol{W}_f \boldsymbol{x}_{f,t}$. where $\boldsymbol{W}_f = (\boldsymbol{w}_{f,1}, \cdots, \boldsymbol{w}_{f,N})^{\mathrm{H}} \in \mathbb{C}^{N\times N}$ is the demixing matrix and $^{\mathrm{H}}$ is the Hermitian transpose.

Under the above assumptions, the speech signal $s_{f,t,n}$ is defined as a random variable following a complex normal distribution with mean 0 and variance $r_{f,t,n} = \mathbb{E}[|s_{f,t,n}|^2]$ as follows

$$p(s_{f,t,n}|r_{f,t,n}) = \mathcal{N}_{\mathbb{C}}(s_{f,t,n}|0, r_{f,t,n}). \tag{4}$$

Assuming that each speech signal ($s_{f,t,n}$ and $s_{f,t,n'}, n \neq n'$) is statistically independent, the negative log-likelihood function $\mathcal{J}$ for the mixed signal can be expressed as

$$\begin{aligned}\mathcal{J}(\mathcal{W},\mathcal{V}) = & -2T\sum_f \log|\det \boldsymbol{W}_f| \\ & + \sum_{f,t,n}\left(\log v_{f,t,n} + \frac{|\boldsymbol{w}_{f,n}^{\mathrm{H}}\boldsymbol{x}_{f,t}|^2}{v_{f,t,n}}\right),\end{aligned} \tag{5}$$

where $\mathcal{W} = \{\boldsymbol{W}_f\}_f$ is a set of all the separation matrices, $\mathcal{V} = \{v_{f,t,n}\}_{f,t,n}$ is a set of $v_{f,t,n}$ for all sources and time-frequencies. Here, when the variance $r_{f,t,n}$ is unconstrained, the (5) is computed independently for each frequency f, thus occurring arbitrariness in the order of the separated signals. This problem is called the permutation problem and requires permutation resolution as a post-processing step to obtain the correct separation signal. Recently, a method has been proposed to achieve source separation while avoiding the permutation problem by introducing constraints on the source model $r_{f,t,n}$. ILRMA and MVAE methods are examples of such techniques.

B. MVAE and FastMVAE

1) MVAE: The MVAE method uses CVAE to represent the source model in (5). CVAE is trained to match the posterior distribution $p_\Phi(\boldsymbol{z}_n, |\boldsymbol{S}_n, \boldsymbol{c}_n) \propto p_\Phi(\boldsymbol{S}_n|\boldsymbol{z}_n, \boldsymbol{c}_n)p(\boldsymbol{z}_n)$ derived from $p_\Phi(\boldsymbol{S}_n|\boldsymbol{z}_n, \boldsymbol{c}_n)$ and $q_\theta(\boldsymbol{S}_n|\boldsymbol{z}_n, \boldsymbol{c}_n)$ as closely as possible and has the structure shown in the upper part of Fig.1.

θ and Φ denote the encoder and decoder weight parameters, respectively. Where $\boldsymbol{c}_n$ is a time-invariant latent variable vector representing the category class of the sound source and $\boldsymbol{z}_n$ is a latent variable representing the time variation of the spectrum.

Now, for simplicity, let these latent variables be represented collectively as $\xi_n = \{\boldsymbol{z}_n, \boldsymbol{c}_n\}$, and the decoder output of CVAE represents the following source model

$$\boldsymbol{R}_n(\xi_n) = g_n \cdot \mathrm{Dec}(\xi_n, \Phi), \tag{6}$$

where $\mathrm{Dec}(\xi_n, \Phi) \in \mathbb{R}^{F\times T}$ is the decoder output, and its f,t component is $\sigma^2_{\Phi,f,t,n}(\xi_n)$. g_n is a variable representing the scale. This source model is the same as in (4), representing the local Gaussian source model as follows

$$p_\Phi(\boldsymbol{S}_n|\boldsymbol{z}_n, \boldsymbol{c}_n, g_n) = \prod_{f,t}\mathcal{N}_{\mathbb{C}}(s_{f,t,n}|0, r_{f,t,n}(\xi_n)), \tag{7}$$

$$r_{f,t,n}(\xi_n) = g_n \cdot \sigma^2_{\Phi,f,t,n}(\xi_n). \tag{8}$$

In the following, $\boldsymbol{R}_n(\xi_n)$ is called the CVAE source model. The CVAE source model $\boldsymbol{R}_n(\xi_n)$ represents various single sources in the training dataset by the latent variable ξ_n.

IP [5] can be used to minimize (5) with respect to $\boldsymbol{W}_f$ as in ILRMA. Since (5) is differentiable with respect to ξ_n, ξ_n can be updated to decrease (5) using backpropagation. The scale parameter g_n can be updated to minimize (5) under fixed ξ_n and $\boldsymbol{W}_f$ as follows

$$g_n \leftarrow \frac{1}{FT}\sum_{f,t}\frac{|\boldsymbol{w}_{f,n}^{\mathrm{H}}\boldsymbol{x}_{f,t}|^2}{\sigma^2_{\Phi,f,t,n}(\xi_n)}. \tag{9}$$

2) FastMVAE: MVAE updates the parameters $\boldsymbol{W}_f, \xi_n$ and g_n so that the objective function (5) becomes smaller at each iteration, which has the advantage of guaranteeing convergence of the objective function to the stopping point. On the other hand, the huge computational costs when determining where to update the CVAE source model parameters ξ_n is an issue. FastMVAE was proposed to avoid this problem. Note that we use FastMVAE2 [11], the latest version of the FastMVAE method, and all descriptions below are based on this version.

FastMVAE decomposes the posterior distribution $p_\theta(\boldsymbol{z}_n, \boldsymbol{c}_n|\boldsymbol{S}_n)$ into a product of two conditional distributions, such as $p_\theta(\boldsymbol{z}_n|\boldsymbol{S}_n)p_\theta(\boldsymbol{c}_n|\boldsymbol{S}_n)$, and learns $q_\Phi(\boldsymbol{z}_n|\boldsymbol{S}_n)$ and $\gamma_\Phi(\boldsymbol{c}_n|\boldsymbol{S}_n)$ to approximate each distribution. As a result, the search for ξ_n, which is used to apply the backpropagation method to (5) , can be substituted by the forward propagation of $q_{\theta_1}(\boldsymbol{z}_n|\boldsymbol{S}_n)$ and $\gamma_{\theta_2}(\boldsymbol{c}_n|\boldsymbol{S}_n)$. Therefore, the backpropagation method is not required, and fast source model updating is possible. Let $\hat{\xi}_n$ denote the outputs $\boldsymbol{z}_n(\boldsymbol{S}_n, \theta_1), \boldsymbol{c}_n(\boldsymbol{S}_n, \theta_2)$ of the encoder with weight parameter $\Theta = \{\theta_1, \theta_2\}$, and denote

$$\hat{\xi}_n = \mathrm{Enc}(\boldsymbol{S}_n, \Theta). \tag{10}$$

The encoder output $\hat{\xi}_n$ in the FastMVAE method is not directly calculated to reduce the objective function of source separation (5). Also, the input signal to the encoder during source separation is a mixed signal $\boldsymbol{X}_n$, or an imperfectly separated signal $\boldsymbol{Y}_n$ during iterative estimation. Thus, one of the problems is that these signals are very different from the clean single speech signals in training phase. For the above

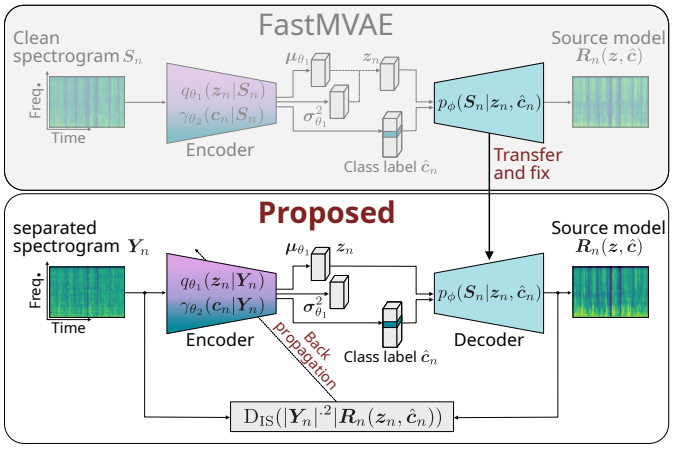

Fig. 2. Overview of the proposed method. Note that during re-training in the proposed method, only the encoder is re-trained through back-propagation. The decoder is fixed with the parameters trained in the previous stage (training as FastMVAE method).

reasons, there is no guarantee that the objective function (5) will decrease in the source model update in the FastMVAE method. It may increase.

III. Proposed method

A. Motivation

The previous work [11] has reported that the FastMVAE method runs as fast as the conventional fast BSS methods but shows slightly worse separation performance than the MVAE method. The main difference between the MVAE and lies in the way the latent variable ξ_n is updated. As explained in section II-B, the latent variable is updated through the forward propagation of the encoder in the FastMVAE method. However, the decrease of the objective function is not guaranteed in this process. Therefore, if the encoder can be trained to consistently find ξ_n that decreases (5), we expect to achieve both the high separation performance of the MVAE method and the fast processing speed of the FastMVAE method.

To achieve this, we propose to re-train only the encoder using (5) as the loss function after network (encoder and decoder) training in the conventional FastMVAE method.This re-training allows the encoder to find the update destinations so that the objective function (5) is reduced under the constraint that the decoder distribution can only represent a single speaker signal. This is very similar to the parameters of the source model in ILRMA. which minimizes the objective function (5) under the constraint that the source spectrogram is represented as a low-rank matrix. Namely, the re-trained encoder can be interpreted as corresponding to the update equation for the source model parameters in ILRMA.

B. Training of encoder parameter Θ

Consider updating $\hat{\xi}_n$ to decrease the objective function (5) as described above. Minimizing the objective function (5) for the source model parameters can be expressed as the Itakura-Saito divergence between the power spectrum of the separated signal $|\boldsymbol{Y}_n|^{\cdot 2}$ estimated in the previous step and the source model $\boldsymbol{R}_n(\xi_n)$ of the decoder output. Therefore, the objective here is to estimate the latent variable $\hat{\xi}_n$ as follows

$$\hat{\xi}_n = \underset{\xi_n}{\operatorname{argmin}}\ \mathrm{D}_{\mathrm{IS}}(|\boldsymbol{Y}_n|^{\cdot 2}|\ \boldsymbol{R}_n(\xi_n)) \tag{11}$$

$$= \underset{\xi_n}{\operatorname{argmin}}\ \mathrm{D}_{\mathrm{IS}}(|\boldsymbol{Y}_n|^{\cdot 2}|\ g_n \cdot \mathrm{Dec}(\xi_n, \Phi)) \tag{12}$$

$$= \underset{\xi_n}{\operatorname{argmin}}\ \sum_{f,t}\left(\frac{|y_{f,t,n}|^2}{g_n \cdot \sigma^2_{\Phi,f,t,n}(\xi_n)} - \log \frac{|y_{f,t,n}|^2}{g_n \cdot \sigma^2_{\Phi,f,t,n}(\xi_n)} - 1\right), \tag{13}$$

where $|\cdot|^{\cdot 2}$ denotes the element-wise square of a matrix, and $\mathrm{D}_{\mathrm{IS}}(\cdot|\cdot)$ is the Itakura-Saito divergence between the matrix elements.

Now, we want the encoder to find the best latent variable $\hat{\xi}_n$ from $\boldsymbol{Y}_n$. Note that $\boldsymbol{Y}_n$ is an imperfectly separated signal since it is obtained during the iterative estimation process. By substituting the expression $\xi_n = \mathrm{Enc}(\boldsymbol{Y}_n, \Theta)$ into (12), the function to be minimized can be written as $\mathrm{D}_{\mathrm{IS}}(|\boldsymbol{Y}_n|^{\cdot 2}|\ g_n \cdot \mathrm{Dec}(\mathrm{Enc}(\boldsymbol{Y}_n, \Theta), \Phi)$. Taking the expectation with respect to $\boldsymbol{Y}_n$, we obtain the following objective function

$$\mathcal{L} = \mathbb{E}_{(\boldsymbol{Y}_n)}\left[\mathrm{D}_{\mathrm{IS}}(|\boldsymbol{Y}_n|^{\cdot 2}|\ g_n \cdot \mathrm{Dec}(\mathrm{Enc}(\boldsymbol{Y}_n, \Theta), \Phi)\right]. \tag{14}$$

Thus, to obtain an encoder that outputs the best latent variable ξ_n for a given $\boldsymbol{Y}_n$, the encoder parameter Θ should be determined to minimize the expression (14). We can see that it is a form of the classical autoencoder (non-variational) training using imperfectly separated signals $\boldsymbol{Y}_n$ as training data. Namely, the difference from the FastMVAE method is that it learns as AE instead of VAE. In addition, the training data is not a clean single speech signal but an imperfectly separated signal obtained during the iterative estimation process.

Although (14) can be seen as AE training using imperfectly separated signals. our goal, however, is not to have the encoder and decoder reconstruct $\boldsymbol{Y}_n$ as it is. Rather, our goal is to let the encoder and decoder find the closest match to $\boldsymbol{Y}_n$ among the possible spectrograms of a single source. This is similar to ILRMA, which updates the source model parameters as to approximate the spectrograms of the imperfectly separated signals obtained during the iterative process using low-rank matrices. In ILRMA, if the number of basis spectra becomes too large, the source model becomes so flexible that it can represent any spectrogram, resulting in poor separation performance. By analogy, we consider it desirable to fix the decoder to the one trained by the FastMVAE method.

Therefore, we propose to train the encoder and decoder first by the FastMVAE method. Then, while keeping the decoder parameter Φ fixed, only the encoder parameter Θ is updated through the backpropagation of the (14). An overview diagram of network training is shown in Fig. 2.

C. Model architecture

In this paper, the network structure used in the first and second stages is the same as the FastMVAE method.

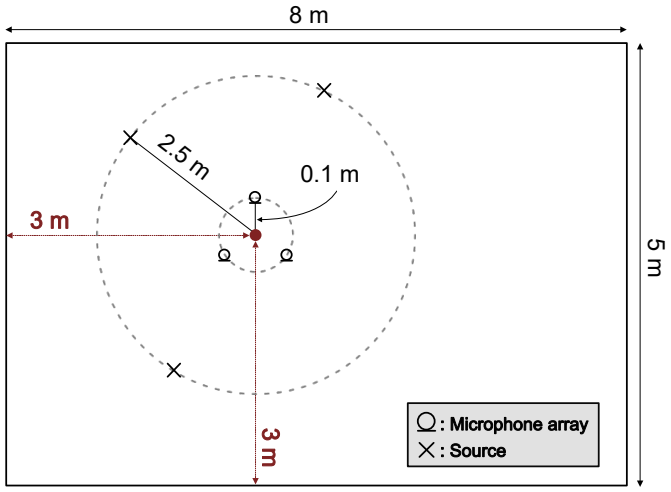

Fig. 3. The room size in the simulation was $8\ \mathrm{m} \times 5\ \mathrm{m} \times 3\ \mathrm{m}$.The microphone array was arranged in a circle with a radius of $0.1\ \mathrm{m}$ centered on the red point. The sound sources were randomly arranged in a circle with a radius of $2.5\ \mathrm{m}$ centered on the red point.

The encoder and classifier layers consist of convolutional layers, LayerNormalization (LN) and Sigmoid Linear Unit (SiLU). The decoder layer consists of an inverse convolution layer, LN and SiLU.

IV. Experiments

A. Conditions

To evaluate the performance of the proposed method, we conducted source separation experiments using speech signals. In this experiment, we applied BSS to the mixed signals and compared the separation performance and how monotonically decreasing the objective function is. ILRMA and the FastMVAE method were employed as comparison methods. The number of bases in ILRMA was set to two.

The WSJ0 [12] dataset was used as training data for the FastMVAE and the proposed methods. For the FastMVAE method, we used single-speaker signals with simulated reverberation by Pyroomacoustics [13]. The signal length was set to $10\ \mathrm{s}$. While for the proposed method, we first selected multiple speech signals from the WSJ0 dataset and simulated the mixing. Then, we applied the FastMVAE method to the mixtures and stored the imperfectly separated signals at each step from 0 to 30 iterations. They were used for re-training the encoder by the proposed method.

The test data used the JVS [14] dataset, and the simulated speech signal of $10\ \mathrm{s}$ was used. The number of sources N was two and three, and the 50 mixed signals to test were created for each source number.

Fig. 3 shows the room configuration, microphone and sound source layout in the simulation. The room size in the simulation was $5\ \mathrm{m} \times 8\ \mathrm{m} \times 3\ \mathrm{m}$. The microphone array was arranged in a circle with a radius of 0.1 m, centered on the red point, as shown in Fig. 3. The sound sources were randomly arranged in a circle with a radius of 2.5 m, centered on the red point. The reverberation time was about $300\ \mathrm{ms}$.

The sampling frequency, STFT length and shift length were set to $16\ \mathrm{kHz}$, $128\ \mathrm{ms}$ and $64\ \mathrm{ms}$, respectively. The amount of improvement in SI-SDR [15] was used as the evaluation index.

B. Results

The average SI-SDR improvement for each method is shown in Fig. 4. The proposed method showed high separation performance for all source numbers. It can also be seen that FastMVAE and the proposed method do not show a significant decrease like ILRMA in separation performance with an increase in the number of sources.

The evolution of the objective functions for each method is shown in Fig. 5. ILRMA updates the source model parameters so that the objective function is always decreasing. Therefore, the objective function of ILRMA has not increased in Fig. 5. On the other hand, FastMVAE often shows an increasing objective function. Although the proposed method is based on FastMVAE, the increase in the objective function was suppressed, and the convergence was similar to ILRMA.

V. Conclusion

In this paper, we proposed a new encoder training method on FastMVAE to improve the estimation of the latent variables at each iteration. After training the encoder and decoder by the conventional FastMVAE method, only the encoder is re-trained using imperfectly separated signals to minimize the objective function of BSS. Experimental results showed that the proposed method suppressed the increase of the objective function at iterative updates and improved the source separation performance.

Acknowledgments

This work was supported by JST CREST JPMJCR19A3.

References

[1] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e12, 2019.

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

[3] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 70–79, 2006.

[4] H. Atsuo, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Independent Component Analysis and Blind Signal Separation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 601–608.

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

[6] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, pp. 1626–1641, 2016.

[7] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1601–1615, 2019.

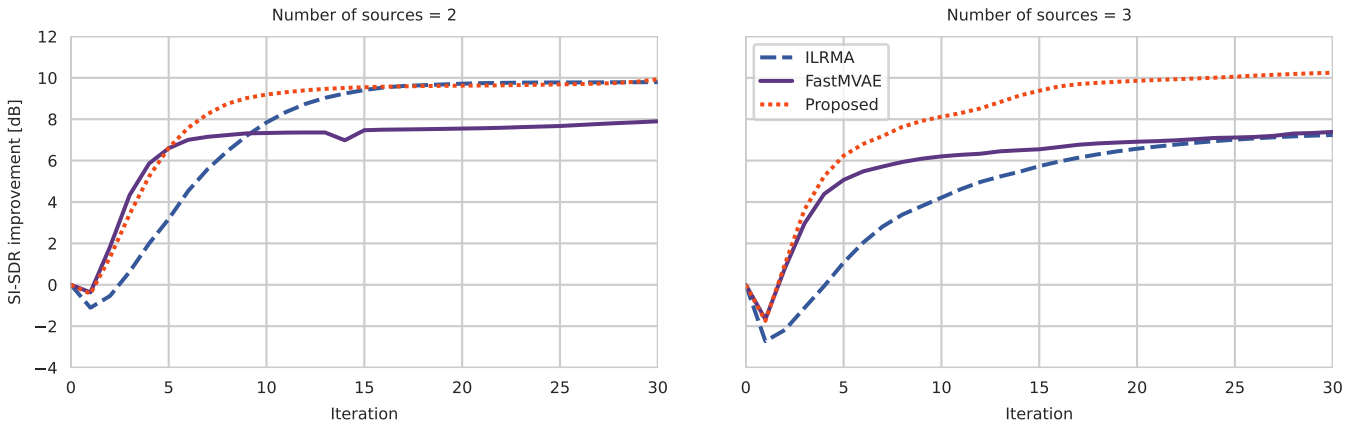

Fig. 4. The SI-SDR improvements with iterations

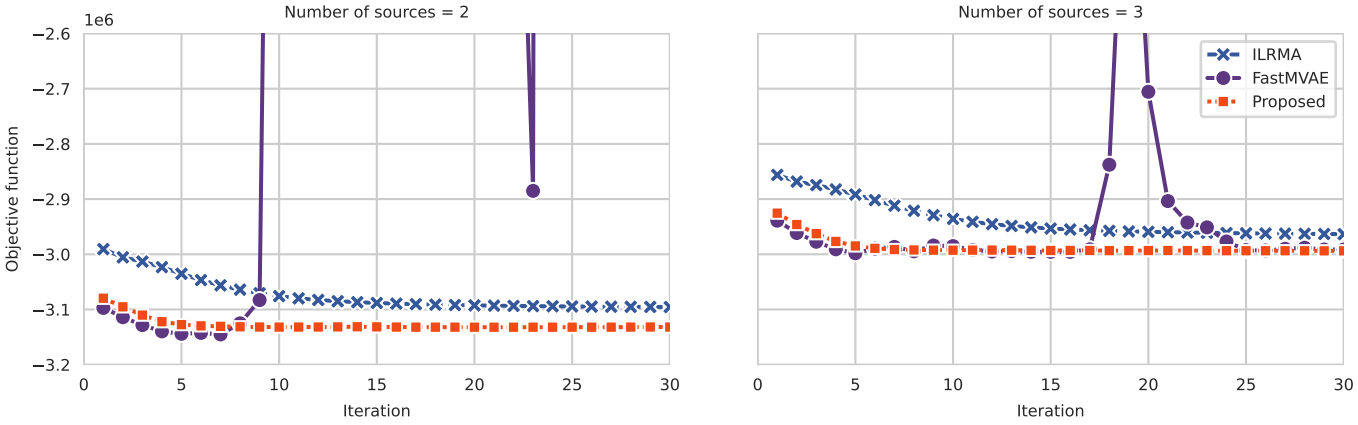

Fig. 5. The objective functions with iterations.

[8] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 10, pp. 1652–1664, Jun. 2016.
[9] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Underdetermined source separation based on generalized multichannel variational autoencoder," *IEEE Access*, vol. 7, pp. 168 104–168 115, 2019.
[10] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Computation*, vol. 31, no. 9, pp. 1891–1914, Sep. 2019.
[11] L. Li, H. Kameoka, S. Inoue, and S. Makino, "FastMVAE2: On improving and accelerating the fast variational autoencoder-based source separation algorithm for determined mixtures," *ArXiv*, vol. abs/2109.13496,v1, 2021.
[12] J. S. Garofolo, G. David, P. Doug, and P. David, "CSR-I (WSJ0) complete LDC93S6A," *Philadelphia: Linguistic Data Consortium*, 1993.
[13] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, Sep. 2018, pp. 351–355.
[14] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free japanese multi-speaker voice corpus," *arXiv preprint*, 2019.
[15] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.