Interspeech 2022
18-22 September 2022, Incheon, Korea

Weakly-Supervised Neural Full-Rank Spatial Covariance Analysis for a Front-End System of Distant Speech Recognition

Yoshiaki Bando[1*], *Takahiro Aizawa*[1,2*], *Katsutoshi Itoyama*[2], *Kazuhiro Nakadai*[2]

[1]National Institute of Advanced Industrial Science and Technology, Japan
[2]Tokyo Institute of Technology, Japan

y.bando@aist.go.jp, {aizawa,itoyama,nakadai}@ra.sc.e.titech.ac.jp

Abstract

This paper presents a weakly-supervised multichannel neural speech separation method for distant speech recognition (DSR) of real conversational speech mixtures. A blind source separation (BSS) method called neural full-rank spatial covariance analysis (FCA) can precisely separate multichannel speech mixtures by using a deep spectral model without any supervision. The neural FCA, however, requires that the number of sound sources is fixed and known in advance. This requirement complicates its utilization for a front-end system of DSR for multi-speaker conversations, in which the number of speakers changes dynamically. In this paper, we propose an extension of neural FCA to handle a dynamically changing number of sound sources by taking temporal voice activities of target speakers as auxiliary information. We train a source separation network in a weakly-supervised manner using a dataset of multichannel audio mixtures and their voice activities. Experimental results with the CHiME-6 dataset, whose task is to recognize conversations at dinner parties, show that our method outperformed a conventional BSS-based system in word error rates.

Index Terms: multichannel source separation, distant speech recognition, deep speech prior, neural source separation

1. Introduction

Speech separation and enhancement are essential for distant speech recognition (DSR) of speech mixtures contaminated by other speakers' speech and environmental noise [1–5]. As represented by commercial smart speakers, the single-speaker DSR of voice command or reading speech has achieved excellent performance in the last decade [4, 5]. In contrast, multi-speaker DSR, especially for daily conversations, still has a lot of difficulties to overcome, as revealed in the CHiME-5 [1] and -6 [2] challenges. There are many challenging problems with not only the recognition of conversational speech but also the separation of noisy, reverberant, and overlapped speech.

Multichannel blind source separation (BSS) has been utilized for front-end systems of DSR in situations where sufficient supervised training data are not available [3, 4, 6, 7]. The complex angular central Gaussian mixture model (cACGMM) [8], for example, has frequently been utilized in single-speaker DSR systems to estimate the time-frequency (TF) mask of a target speaker from a noisy speech signal [4]. A guided source separation (GSS) [3] was proposed to separate conversational speech mixtures by extending the cACGMM to take the temporal voice activities of target speakers as auxiliary information.

While the cACGMM-based separation has yielded great success in DSR, its separation performance is limited by its strong assumption that each TF bin contains only one dominant source. Multichannel nonnegative matrix factorization (MNMF) [9] represents each TF bin as a sum of source signals and is reported to improve the separation performance compared with a clustering-based method. Since MNMF represents source spectra with a linear NMF-based model, its nonlinear extension called neural full-rank spatial covariance analysis (FCA) has been proposed [10]. This method jointly trains its nonlinear source model and a source separation network only from multichannel mixture recordings without any supervision. At the test time, speech signals are separated from a mixture signal by estimating latent speech embeddings and corresponding spatial covariance matrices (SCMs) on the fly. An experimental result with numerically-simulated speech mixtures reported that the neural FCA outperformed existing BSS methods including the cACGMM [8] and FastMNMF [11].

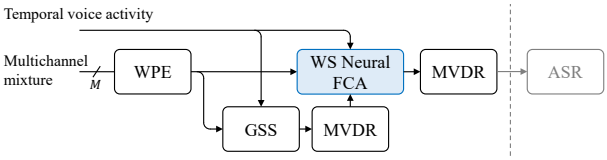

Figure 1: *Overview of proposed front-end system based on weakly-supervised neural FCA.*

In this paper, we propose a front-end system of DSR based on an extension of neural FCA to handle a dynamically changing number of sound sources (Fig. 1). The original neural FCA assumes that the number of sound sources is fixed and known in advance. Since the number of speakers changes dynamically in multi-speaker conversations, it is difficult to apply the original method to such recordings naïvely. The proposed method extends the original neural FCA to take the temporal voice activities as auxiliary information. To this end, we formulate a nonlinear generative model of a multichannel mixture signal conditioned by the temporal voice activities. Based on this generative model, an objective function is derived to train neural source separation on a weakly-supervised dataset of multichannel audio mixtures and time annotations of voice activities.

The main contribution of this paper is to extend a nonlinear multichannel BSS method to work for real-world audio recordings. We evaluate the proposed weakly-supervised neural FCA using the CHiME-6 dataset, which provides actual multichannel recordings of dinner parties. In the CHiME-6 challenge (Track 1), many DSR systems utilized the GSS-based source separation for their front-end systems [12–14]. In this paper, we show that a separation network can be directly trained from actual multichannel conversation recordings and their voice activities and improves the word error rate (WER) compared to a conventional GSS-based system.

*These two authors contributed equally to this work.

10.21437/Interspeech.2022-11077

2. Related Work

This section first reviews the existing front-end systems of DSR and then introduces a series of blind source separation methods.

2.1. Front-end systems of distant speech recognition

Multichannel DSR has been investigated to robustly recognize noisy reverberant speech signals [6]. One approach is to jointly train a neural beamforming front-end and an ASR back-end in an end-to-end manner [15, 16]. This framework trains the front-end system to minimize the WER for the training data. When the amount of transcribed training data is limited, a front-end system is constructed separately from the back-end. A typical system first performs dereverberation with the weighted prediction error (WPE) method [17], then estimates the spatial statistics (e.g., SCMs) of the target speech and other interfering signals, and finally obtains the enhanced speech by applying an adaptive beamformer. If we have a sufficiently large amount of isolated signals of speech and noise, the spatial statistics can be estimated by a neural network trained in a supervised manner [18, 19]. Several studies have trained an enhancement network with pseudo-isolated signals generated by separating or enhancing noisy speech mixtures with conventional methods [20–22]. Multichannel BSS methods have also been utilized, especially when we cannot obtain any isolated signals. The cACGMM [4, 8] and its extension, GSS [3], have been actively utilized because of their low computational complexity and high robustness against diffuse noise.

2.2. Blind source separation

Recent BSS methods are based on probabilistic generative models, and they can be categorized into two types: mixture and factor models. The mixture models (including cACGMM [8]) are designed for clustering the TF bins of a mixture signal into individual source components. The mixture models are valid as long as the source signals are sparse enough not to overlap in the TF domain. The factor models [9–11, 23], in contrast, assume that each TF bin follows a multivariate Gaussian distribution with the sum of SCMs of all the sources. Since the noisy speech mixture signals are the sums of source signals, the factor models are more suitable representations than the mixture models. A popular factor model called MNMF [9, 24] assumes that each source signal follows a Gaussian distribution whose power spectral density (PSD) is represented by NMF. The NMF-based source model can be replaced with a nonlinear generative model based on a variational autoencoder (VAE) [25]. Its powerful representation enables us to precisely represent speech source spectra by pretraining the network from isolated signals [26–28]. The VAE-based source model has recently gained the ability to be trained only from multichannel mixture signals in an unsupervised (or blind) manner. This method, called neural FCA [10], trains the source generative model and a source separation model as a large VAE for the multichannel mixture signals.

3. Unsupervised Neural FCA

This section describes the original neural FCA [10], which is extended to a weakly-supervised version in the next section. Let $f = 1, \ldots, F$ and $t = 1, \ldots, T$ be the frequency and time frame indices, respectively. The neural FCA is based on a probabilistic generative model that represents an M-channel observed mixture signal $\mathbf{x}_{ft} \in \mathbb{C}^M$ as a sum of N source signals $s_{nft} \in \mathbb{C}$ $(n = 1, \ldots, N)$:

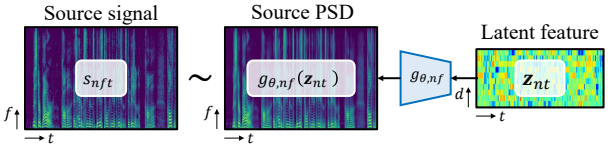

Figure 2: *Overview of deep spectral model.*

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{a}_{nf} s_{nft}, \tag{1}$$

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ is the steering vector for source n. Each source signal is represented by a deep spectral model [26, 29] as in Fig. 2. In this model, the source signal s_{nft} follows a zero-mean Gaussian distribution characterized by D-dimensional latent vectors $\mathbf{z}_{nt} \in \mathbb{R}^D$:

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, g_{\theta,nf}(\mathbf{z}_{nt})\right), \tag{2}$$

where $g_{\theta,nf} : \mathbb{R}^D \to \mathbb{R}_+$ is a neural network with parameters θ that transforms the latent representation $\mathbf{z}_{nt}$ to the PSD. From these equations, the likelihood function is derived as follows:

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} g_{\theta,nf}(\mathbf{z}_{nt}) \mathbf{H}_{nf}\right), \tag{3}$$

where $\mathbf{H}_{nf} = \mathbb{E}[\mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}}] \in \mathbb{S}_+^{M\times M}$ is an SCM of source n. The full-rankness of $\mathbf{H}_{nf}$ can handle small source movements and reverberations [23].

The source generative model (or decoder) $g_{\theta,nf}$ is trained from multichannel mixture signals in an unsupervised manner. This training is derived as in the autoencoding variational Bayesian inference [25] by assuming $\mathbf{z}_{nt}$ to follow the standard Gaussian distribution:

$$\mathbf{z}_{nt} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right). \tag{4}$$

This prior distribution helps to solve the frequency permutation ambiguity by encouraging the statistical independency among the source signals [29]. Let $\mathbf{X} \triangleq \{\mathbf{x}_{ft}\}_{f,t=1}^{F,T}$ be a mixture spectrogram in the training data. The neural FCA introduces a separation (or encoder) network with parameters ϕ to predict the posterior distribution of source latent vectors $q_\phi(\mathbf{z}_{nt} \mid \mathbf{X})$, and the encoder and decoder networks are trained jointly to maximize the following evidence lower bound (ELBO):

$$\mathcal{L} = \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{H})] - \mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X}) \mid p(\mathbf{Z})]. \tag{5}$$

where $\mathbb{E}_{q_\phi}[\cdot]$ is the expectation by the posterior $q_\phi(\mathbf{z}_{nt} \mid \mathbf{X})$, and $\mathcal{D}_{\mathrm{KL}}[q \mid p]$ is the Kullback-Leibler (KL) divergence between q and p. The SCM $\mathbf{H}_{nf}$ is obtained by an expectation-maximization (EM) update rule [23] to maximize the ELBO. The network parameters are updated by the stochastic gradient descent with the backpropagation technique.

The neural FCA can be considered as nonlinear BSS performed on the training mixture signals. It can obtain the powerful nonlinear source model and its inference network in an unsupervised manner. Once the two networks are trained, they can be used to separate unknown mixture signals by estimating the PSD of source spectra. This method, however, is difficult to apply to real-world audio recordings because the number of sound sources N should be fixed for all of the training and test data and known in advance. The evaluation in the literature [10] was performed only with numerically-simulated mixtures of two speech signals.

4. Weakly-Supervised Neural FCA

We propose an extension of neural FCA designed for a front-end system of DSR for conversational speech recordings. It is generally difficult to collect a sufficiently large amount of isolated signals for real conversational speech mixtures. The proposed method trains a source separation network directly from a set of multichannel noisy speech mixtures and their time annotations of voice activity. Since we use the time annotations, the resulting method is a weakly-supervised approach.

4.1. Generative model of multichannel mixture signal

Using the voice activities of N_{spk} speakers $u_{nt} \in \{0, 1\}$ ($n = 1, \ldots, N_{\mathrm{spk}}$), the observed mixture signal $\mathbf{x}_{ft}$ is represented by a sum of speech sources $s_{nft} \in \mathbb{C}$ ($n = 1, \ldots, N_{\mathrm{spk}}$) active in the corresponding time frame and N_{noi} noise signals $s_{nft} \in \mathbb{C}$ ($n = N_{\mathrm{spk}} + 1, \ldots, N_{\mathrm{spk}} + N_{\mathrm{noi}}$) as follows:

$$\mathbf{x}_{ft} = \sum_{n \in \mathfrak{N}_t} \mathbf{a}_{nf} s_{nft}, \tag{6}$$

where $\mathfrak{N}_t = \{n | u_{nt} = 1\} \cup \{N_{\mathrm{spk}} + 1, \ldots, N_{\mathrm{spk}} + N_{\mathrm{noi}}\}$ is a set of source indices active in time frame t. The source signals are represented by the deep spectral model as in the original neural FCA, and we obtain the following likelihood:

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \sum_{n \in \mathfrak{N}_t} g_{\theta,nf}(\mathbf{z}_{nt}) \mathbf{H}_{nf} \right). \tag{7}$$

Note that this naïve formulation may lead that the noise sources represent target speech signals because they are always active. To avoid this problem, we reduce the number of feature dimensions for noise D_{noi} compared to that of speech D_{spk}. In other words, the speech spectra have high degrees of freedom but limited activity, while the noise spectra are always active but have limited degrees of freedom.

4.2. Weakly-supervised training

We train the source model $g_{\theta,nf}$ from a set of multichannel mixture signals and temporal voice activities in a weakly-supervised manner based on the autoencoding paradigm. More specifically, the source model is regarded as the decoder, and we introduce a separation (encoder) model $q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})$ that estimates the latent source vectors $\mathbf{Z} \triangleq \{\mathbf{z}_{nt}\}_{n,t=1}^{N,T}$ from the mixture $\mathbf{X} \triangleq \{\mathbf{x}_{ft}\}_{f,t=1}^{F,T}$ and time activities $\mathbf{U} \triangleq \{u_{nt}\}_{n,t=1}^{N,T}$:

$$q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U}) = \prod_{n,t,d} \mathcal{N}\left(z_{ntd} \mid \mu_{\phi,ntd}(\mathbf{C}), \sigma^2_{\phi,ntd}(\mathbf{C})\right), \tag{8}$$

where z_{ntd} is the d-th element of $\mathbf{z}_{nt}$, and $\mu_{\phi,ntd}(\mathbf{C}) \in \mathbb{R}$ and $\sigma^2_{\phi,ntd}(\mathbf{C}) \in \mathbb{R}_+$ are outputs of a separation network that has model parameters ϕ and takes feature vectors $\mathbf{C}$ as input. The feature $\mathbf{C}$ is calculated from $\mathbf{X}$ and $\mathbf{U}$ (described in Sec. 4.3). We train these two networks to maximize the following ELBO:

$$\mathcal{L} = \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{X}|\mathbf{Z}, \mathbf{H}, \mathbf{U})] - \mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U}) | p(\mathbf{Z})]. \tag{9}$$

In this optimization, the parameters θ and $\mathbf{H}_{nf}$ are updated to maximize the log-marginal likelihood, and the parameter ϕ is updated to minimize the KL divergence $\mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U}) \mid p_\theta(\mathbf{Z}|\mathbf{X}, \mathbf{H}, \mathbf{U})]$. Since the first term of the ELBO is intractable, we calculate it with the Monte-Carlo approximation:

$$\begin{aligned} &\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{H}, \mathbf{U})] \\ &\quad \approx - \sum_{f,t=1}^{F,T} \log |\mathbf{Y}_{:ft}| - \sum_{f,t=1}^{F,T} \mathbf{x}_{ft}^{\mathsf{H}} \mathbf{Y}_{:ft}^{-1} \mathbf{x}_{ft}, \end{aligned} \tag{10}$$

where $\mathbf{Y}_{:ft} = \sum_{n \in \mathfrak{N}_t} \mathbf{Y}_{nft} \in \mathbb{S}_+^{M \times M}$ is a sum of source images $\mathbf{Y}_{nft} \triangleq g_{\theta,nf}(\mathbf{z}^*_{nt}) \mathbf{H}_{nf}$, and $\mathbf{z}^*_{nt} \sim q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})$ is a sample from the encoder output $q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})$. The network parameters θ and ϕ are updated with stochastic gradient descent, while the SCMs $\mathbf{H}_{nf}$ are obtained as constant values by iteratively performing the following update rules [30]:

$$\mathbf{H}_{nf} \leftarrow \mathbf{B}_{nf}^{-\frac{1}{2}} \left(\mathbf{B}_{nf}^{\frac{1}{2}} \mathbf{A}_{nf} \mathbf{B}_{nf}^{\frac{1}{2}} \right)^{\frac{1}{2}} \mathbf{B}_{nf}^{-\frac{1}{2}}, \tag{11}$$

$$\mathbf{A}_{nf} \triangleq \mathbf{H}_{nf} \left(\sum_t g_{\theta,nf}(\mathbf{z}^*_{nt}) \mathbf{Y}_{:ft}^{-1} \mathbf{x}_{ft} \mathbf{x}_{ft}^{\mathsf{H}} \mathbf{Y}_{:ft}^{-1} \right) \mathbf{H}_{nf}, \tag{12}$$

$$\mathbf{B}_{nf} \triangleq \sum_t g_{\theta,nf}(\mathbf{z}^*_{nt}) \mathbf{Y}_{:ft}^{-1}, \tag{13}$$

where $\mathbf{X}^{\frac{1}{2}}$ is the square root of matrix $\mathbf{X} = \mathbf{X}^{\frac{1}{2}} \mathbf{X}^{\frac{1}{2}}$. The summations in Eqs. (12) and (13) are performed over the time frames where the source n is active. Since this update rule includes $\mathbf{H}_{nf}$ itself, we update $\mathbf{H}_{nf}$ five times from an identity matrix for every one update of θ and ϕ. Note that we utilized the multiplicative update (MU) rule for $\mathbf{H}_{nf}$ instead of the expectation-maximization (EM) rule in the original neural FCA. We utilized the MU update rule for its low computational complexity. While the EM rule requires the inversions of $\mathcal{O}(FTN)$ matrices in each iteration, the MU update rule requires only the inversions of $\mathcal{O}(F(T + N))$ matrices and the square roots of $\mathcal{O}(FN)$ matrices, where N is $N_{\mathrm{spk}} + N_{\mathrm{noi}}$.

4.3. Input feature

The selection of the input feature is vital for the separation network because using only the naïve mixture spectrogram makes it too difficult to find the clues of sound source separation. In this paper, we utilized the enhancement results by the GSS-based clustering as the input feature. The input feature $\mathbf{C}$ is the concatenation of the log-power mixture spectrogram, those of GSS results for all the speakers, and the voice activities.

4.4. Front-end system of distant speech recognition

Once the separation and generative models are trained, they are used for the front-end system of DSR as shown in Fig. 1. Following the existing DSR systems, we first perform dereverberation on the observed mixture by using the weighted prediction error (WPE) method [17]. We then extract the feature $\mathbf{C}$ from the dereverberated mixture by using the GSS [3] and input it to the encoder network to obtain the PSD of source signals $g_{\theta,nf}(\boldsymbol{\mu}_{\phi,nt}(\mathbf{C}))$. The SCMs corresponding to the sources are estimated using Eq. (11). Finally, we obtain the speech spectrograms by performing minimum-variance distortionless response (MVDR) beamforming [3,31] with the estimated SCMs.

5. Experimental Evaluation

We evaluated the proposed method with the WER of conversation recordings provided by the CHiME-6 dataset [2].

5.1. Dataset

The CHiME-6 dataset [2] provides multichannel audio recordings of 20 dinner parties each of which was recorded in a different home. Each party had $N_{\mathrm{spk}} = 4$ participants and lasts at least two hours. The recordings were captured by six or five four-channel microphone arrays (Microsoft Kinect v2), which were placed in three different locations in the home: kitchen,

Table 1: *Word error rates for dev and eval sets of the CHiME-6 dataset.*

Method	Dev set				Eval set			
	Avg.	Dining	Kitchen	Living	Avg.	Dining	Kitchen	Living
GSS + MVDR (official implementation)	51.8	53.8	53.9	48.6	51.3	44.7	61.2	50.3
GSS + MVDR ($M = 16$, 30-s context)	49.8	51.6	52.3	46.4	51.1	45.0	60.8	49.7
WS Neural FCA + MVDR ($N_{\mathrm{noi}} = 2$, IPD spec.)	55.7	56.5	59.6	51.7	52.9	46.1	60.6	53.9
WS Neural FCA + MVDR ($N_{\mathrm{noi}} = 1$, GSS spec.)	49.4	51.4	51.7	46.1	49.4	43.3	57.4	49.5
WS Neural FCA + MVDR ($N_{\mathrm{noi}} = 2$, GSS spec.)	48.6	51.2	50.8	45.1	49.0	43.2	56.7	48.9

dining, and living. The participants were free to converse on any topics without any artificial scenario-ization. The dataset provides the transcriptions of participants' utterances with their start and end timestamps. Note that this dataset also provides the recordings by worn microphones for reference, but we did not use it for training or inference of our method. The audio signals were recorded in 16-kHz and 16-bit sampling. The 20 parties were split into train, dev, and eval sets, which have 16, 2, and 2 parties, respectively. Since the recordings were not synchronized over the multiple Kinects, array synchronization was performed by an official CHiME-6 toolkit. Following the existing studies [3, 21], we suppressed the late reverberation by WPE-based dereverberation [17] to all of the microphones.

5.2. Experimental condition

The network architectures of the proposed method were experimentally determined as follows. The encoder network is based on an existing separation network [32] consisting of multiple U-Net-like blocks. The input frames were first converted to 512-channel vectors by a 1×1-convolutional layer, and then we stacked eight U-Net blocks with residual connections. Each block consists of eight 1024-channel depth-wise convolutional layers with the kernel size of 5 and the activation functions of parametric rectified linear units (PReLUs). The outputs $\mu_{\phi,ntd}(\mathbf{C})$ and $\sigma^2_{\phi,ntd}(\mathbf{C})$ were obtained by 1×1-convolutional layers. The decoder $g_{\theta,nf}$ consists of three 1D-convolutional layers with 512 latent channels, the kernel size of 3, Swish activations, and residual connections. The nonnegativity of $\sigma^2_{\phi,ntd}$ and $g_{\theta,nf}$ was obtained by the softplus activation function.

The networks were trained for 200 epochs by an Adam optimizer [33] with the learning rate of 1.0×10^{-3}. The spectrograms were obtained by the short-time Fourier transform with the window size of 1024 samples and the hop length of 256 samples. The number of noise sources N_{noi} was set to 2. The training mixture signals were split into 30-s clips, and the networks were trained by feeding 64 clips as one mini-batch. The dimensions of the source latent variables D_{spk} was 50 and that of noise D_{noi} was 20. Following the original neural FCA [10], we performed the cyclic annealing of the KL term in Eq. (9) with the maximum weight of 5.0. To reduce the computational cost, we performed the training with $M = 12$ channels having the largest powers in the all 24 channels of the mixture signals. At the inference phase, the latent features $\mathbf{z}_{nt}$ were estimated by the encoder network, and the SCMs $\mathbf{H}_{nf}$ were updated ten times. The beamforming was performed by $M = 16$ channels having the largest powers in all of the microphones. Target utterances were separated with surrounding 30 s of mixture recordings for obtaining the spatial context [3].

Our method is evaluated with WER obtained by using an ASR back-end. We utilized the official baseline pre-trained model for the Kaldi-based ASR system [34] provided by the organizers of the CHiME-6 challenge [2]. The acoustic model of this system is based on the factorized time-delay neural network (TDNN-F), and its language model is based on a 3-gram language model. As a baseline, we compared the proposed method with our implementation of the GSS-based front-end system [3]. As in the proposed method, the GSS is performed by $M = 16$ channels with 30-s context. The number of iterations was 20. We also compared with a publicly available implementation of GSS that is the official CHiME-6 baseline [2].

5.3. Experimental results

The WERs for the dev and eval sets of the CHiME-6 dataset are listed in Table 1. We first see that the WER of GSS is improved by our implementation with $M = 16$ channels and 30-s context compared to the official implementation ($M = 12$ with 20-s context). The WER of the proposed weakly-supervised (WS) neural FCA was significantly improved by taking the log-power spectrograms of GSS results compared with that replacing them with the inter-channel phase differences (IPDs) of a multichannel mixture signal. While the original neural FCA utilized the IPDs for an input feature, this result indicates that the IPDs of mixtures were not effective for actual noisy reverberant recordings. The WER was slightly improved by increasing the number of noise sources N_{noi} to 2 from 1. This gain might be because the spatial flexibility for noise signals was increased. The resulting WER of our method was clearly improved from both the two implementations of the GSS-based front-end systems.

6. Conclusion

In this paper, we proposed a weakly-supervised neural FCA for a front-end system of DSR that recognizes real conversations at dinner parties. We extended the unsupervised neural FCA to take the voice activities of target speakers as input. This extension is derived by formulating a new nonlinear generative model of multichannel mixture signals to handle the dynamically changing number of sound sources. The proposed method can precisely represent audio mixture signals with its nonlinear source generative model and full-rank SCM model. The experimental results with the CHiME-6 dataset showed that the proposed front-end system outperformed the official baseline system of the CHiME-6 challenge (Track 1) in WER. The future work includes jointly training the source separation and source number counting without any supervision. Since our method is based on a probabilistic generative model, it could be integrated with the conventional Bayesian frameworks that can estimate the number of sound sources [35, 36].

7. Acknowledgments

This work was supported in part by NEDO and the JST ACT-X under Grant JPMJAX200N. We thank Dr. Sei Ueno and Mr. Keitaro Tanaka for their valuable discussion.

8. References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[2] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[3] C. Boeddecker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. of CHiME-5 Workshop*, 2018, pp. 35–40.

[4] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM TASLP*, vol. 25, no. 4, pp. 780–793, 2017.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. of IEEE ASRU*. IEEE, 2015, pp. 504–511.

[6] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel nmf-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM TASLP*, vol. 27, no. 5, pp. 960–971, 2019.

[7] L. Drude, J. Heymann, and R. Haeb-Umbach, "Unsupervised training of neural mask-based beamforming," in *Proc. of Interspeech*, 2019, pp. 1253–1257.

[8] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. of EUSIPCO*, 2016, pp. 1153–1157.

[9] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE TASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[10] Y. Bando, K. Sekiguchi, Y. Masuyama, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Neural full-rank spatial covariance analysis for blind source separation," *IEEE SPL*, vol. 28, pp. 1670–1674, 2021.

[11] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM TASLP*, vol. 28, pp. 2610–2625, 2020.

[12] J. Du, Y.-H. Tu, L. Sun, L. Chai, X. Tang, M.-K. He, F. Ma, J. Pan, J.-Q. Gao, D. Liu *et al.*, "The USTC-NELSLIP systems for CHiME-6 challenge," in *Proc. of CHiME-6 Workshop*, 2020, pp. 1–5.

[13] H. Chen, P. Zhang, Q. Shi, and Z. Liu, "The IOA systems for CHiME-6 challenge," in *Proc. of CHiME-6 Workshop*, 2020, pp. 1–4.

[14] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny *et al.*, "The STC system for the CHiME-6 challenge," in *Proc. of CHiME-6 Workshop*, 2020, pp. 1–6.

[15] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proc. of ICML*, 2017, pp. 2632–2641.

[16] A. S. Subramanian, C. Weng, M. Yu, S.-X. Zhang, Y. Xu, S. Watanabe, and D. Yu, "Far-field location guided target speech extraction using end-to-end speech recognition objectives," in *Proc. of IEEE ICASSP*, 2020, pp. 7299–7303.

[17] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE TASLP*, vol. 20, no. 10, pp. 2707–2720, 2012.

[18] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. of IEEE ICASSP*, 2016, pp. 196–200.

[19] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks." in *Proc. of Interspeech*, 2016, pp. 1981–1985.

[20] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome, "Unsupervised training for deep speech source separation with Kullback-Leibler divergence based probabilistic loss function," in *Proc. of IEEE ICASSP*, 2020, pp. 56–60.

[21] Y. Tu, J. Du, L. Sun, F. Ma, J. Pan, and C.-H. Lee, "A space-and-speaker-aware iterative mask estimation approach to multi-channel speech recognition in the CHiME-6 challenge." in *Proc. of Interspeech*, 2020, pp. 96–100.

[22] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *Proc. of IEEE ICASSP*, 2019, pp. 695–699.

[23] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[24] 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 TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[26] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," *arXiv preprint arXiv:1808.00892*, 2018.

[27] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM TASLP*, vol. 27, no. 12, pp. 2197–2212, 2019.

[28] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. of IEEE ICASSP*, 2019, pp. 101–105.

[29] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. of IEEE ICASSP*, 2018, pp. 716–720.

[30] K. Yoshii, "Correlated tensor factorization for audio source separation," in *2018 Proc. of IEEE ICASSP*, 2018, pp. 731–735.

[31] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE TASLP*, vol. 18, no. 2, pp. 260–276, 2009.

[32] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *Proc. of IEEE MLSP*, 2020, pp. 1–6.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, 2011, pp. 1–4.

[35] J. Taghia and A. Leijon, "Variational inference for watson mixture model," *IEEE TASLP*, vol. 38, no. 9, pp. 1886–1900, 2015.

[36] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with Dirichlet prior," in *Proc. of IEEE ICASSP*, 2009, pp. 33–36.