Direction-aware target speaker extraction with a dual-channel system based on conditional variational autoencoders under underdetermined conditions

Rui Wang, Li Li and Tomoki Toda

Nagoya University, Nagoya, Japan
E-mail: {rui.wang, li.li}@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—In this paper, we deal with a dual-channel target speaker extraction (TSE) problem under underdetermined conditions. For the dual-channel system, the generalized sidelobe canceller (GSC) is a commonly used structure for estimating a blocking matrix (BM) to generate interference, and geometric source separation (GSS) can be used as an implementation of BM estimation utilizing directional information. However, the performance of the conventional GSS methods is limited under underdetermined conditions because of the lack of a powerful source model. In this paper, we propose a dual-channel TSE method that combines the ability of target selection based on geometric constraints, more powerful source modeling, and nonlinear postprocessing. The target directional information is used as a geometric constraint, and two conditional variational autoencoders (CVAEs) are used to model a single speaker's speech and interference mixture speech. For the postprocessing, an ideal ratio Time–Frequency (T-F) mask estimated from the separated interference mixture speech is used to extract the target speaker's speech. The experimental results demonstrate that the proposed method achieves 6.24 dB and 8.37 dB improvements compared with the baseline method in terms of signal-to-distortions ratio (SDR) and source-to-interferences ratio (SIR) respectively under strong reverberation for 470 ms.**

***Index Terms*—multichannel source separation, target speaker extraction, multichannel variational autoencoder (MVAE)**

I. Introduction

Target speaker extraction (TSE) aims to extract a target speaker's voice from mixed signals, which is desired for numerous applications like speech recognition systems and smart home devices. Over the past few decades, various methods based on blind source separation (BSS) [1] [2] have been applied to this task. Most BSS methods are often designed for determined conditions where the number of microphones M is equal to the number of sources N ($M = N$), and their performance depends on the number of microphones. However, in realistic applications, many devices are often equipped with only a few microphones because of the hardware limitation or efficiency, and many real-world TSE tasks need to deal with scenarios with many speakers, which is insufficient to meet the determined condition. Therefore, TSE under underdetermined cases has become interesting and challenging research.

Generalized sidelobe canceller (GSC) [3] [4] is one of the implementations of underdetermined TSE, which can be interpreted as three main components: a fixed beamformer for target enhancement, a blocking matrix (BM) that provides an estimation of interference by suppressing the target signal only, and an adaptive canceller that serves as a postfilter. A good estimation of BM is important for TSE [5]. Imposing a Geometric Constraint (GC) based on spatial information like direction-of-arrival (DOA) on the BSS method is effective in the BM estimation, which yielded many geometric source separation (GSS) methods [6]–[11]. For example, Geometrically constrained independent vector analysis (GCIVA) [11] [12] is a well-known method, which combines linear GC derived from prior spatial information with independent vector analysis (IVA) [13] [14]. However, these GSS methods are designed for determined cases, and their performance for underdetermined TSE is limited. Recently, a Bayesian framework-based GCIVA has been proposed, which introduces a Background (BG) source model derived from the Independent Vector Extraction (IVE) [15] that allows for underdetermined cases to extract the Source Of Interest (SOI) [16]. This method models all background signals except the SOI, including speech, white noise, and diffuse sound fields. On the other hand, it is not straightforward to accurately model various types of background signals.

For TSE under underdetermined cases, a powerful source model is required because the source model not only needs to deal with the target speaker's voice but also needs to deal with the multi-speaker interference mixture. Many efforts have been done in developing the source model of the speech signal. Independent low-rank matrix analysis (ILRMA) applied a flexible source model of nonnegative matrix factorization (NMF) decomposition in the IVA framework, which yielded a better modeling power of complex spectral structures than former IVA with a Laplace distribution-based source model [17]. Most recently, the deep neural network (DNN) has been used to model the source spectral characteristics owing to its powerful modeling capability [18] [19]. The multichannel variational autoencoder (MVAE) method [20] utilizes the conditional variational autoencoder (CVAE) [21] as the generative source model in an IVA framework under determined cases and has attracted attention.

In this paper, we propose a novel dual-channel TSE method under underdetermined conditions, which combines GC-based

TSE, the CVAE-based source model, and an ideal ratio time-frequency (T-F) mask-based postprocessing. Inspired by MVAE, we innovatively model the target and interference mixture separately using two types of CVAE to better handle underdetermined cases. We design an iterative TSE algorithm based on the GSC structure with linear GCs and the target and interference mixture source modeling with CVAEs to provide a good estimate of interference mixture when only the direction of the target is known. Experimental results show that our trained CVAE is powerful in modeling clean speech and mixed speech and the combination of GC and CVAE-based source model leads to a significant improvement in TSE tasks under underdetermined conditions.

II. Problem formulation of GSS

Let us consider a TSE problem where a dual-channel microphone array is used. Let $\mathbf{s}(f,n)$ and $\mathbf{x}(f,n)$ be the short-time Fourier transform (STFT) coefficients of the source and observed signals, where f and n are the frequency and time indices. We represent:

$$\mathbf{s}(f,n) = [s_1(f,n), s_2(f,n)]^T, \tag{1}$$
$$\mathbf{x}(f,n) = [x_1(f,n), x_2(f,n)]^T, \tag{2}$$

where $s_1(f,n)$ is the target with a known DOA and $s_2(f,n)$ is the interference mixture except the target. $x_1(f,n)$ and $x_2(f,n)$ are the observed signals of two input channels. We use a separation system as

$$\mathbf{s}(f,n) = \mathbf{W}^{\mathrm{H}}(f)\mathbf{x}(f,n), \tag{3}$$
$$\mathbf{W}(f) = [\boldsymbol{w}_1(f), \boldsymbol{w}_2(f)], \tag{4}$$

where $\mathbf{W}(f)$ is called a demixing matrix and $\mathbf{s}(f,n)$ is an estimate of the target and interference mixture. $\boldsymbol{w}_1(f)$ is used to enhance the target while $\boldsymbol{w}_2(f)$ is used to estimate the interference by suppressing the target.

Let us assume that source signals follow the local Gaussian model (LGM) [23], i.e., $s_j(f,n)$ independently follows a zero-mean complex Gaussian distribution with the variance $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$. We further assume that $s_1(f,n)$ and $s_2(f,n)$ are independent of each other. $\mathbf{s}(f,n)$ then follows:

$$\mathbf{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|\mathbf{0}, \mathbf{V}(f,n)), \tag{5}$$

where $\mathbf{V}(f,n) = \mathrm{diag}[v_1(f,n), v_2(f,n)]$. From Eqs. (3) and (5), we can show that $\mathbf{x}(f,n)$ follows:

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, (\mathbf{W}^{\mathrm{H}}(f))^{-1}\mathbf{V}(f,n)\mathbf{W}(f)^{-1}). \tag{6}$$

The log-likelihood of $\mathcal{W} = \{\mathbf{W}(f)\}_f$ is given by:

$$\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) \stackrel{c}{=} 2N\sum_f \log|\det\mathbf{W}(f)| - \sum_{f,n,j}(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)|^2}{v_j(f,n)}), \tag{7}$$

where $\stackrel{c}{=}$ denotes equality up to constant terms and source model parameters are represented as $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$.

Now, let us consider geometric constraints [6] restricts the far-field response of the jth demixing filter in the target direction α, which is described as

$$J_b(\mathcal{W}) = \sum_j \lambda_j \sum_f |\boldsymbol{w}_j^{\mathrm{H}}(f)\boldsymbol{d}(f,\alpha) - b_j|^2, \tag{8}$$

where $\boldsymbol{d}(f,\alpha)$ is a steering vector toward α, and λ_j is a weighting parameter on the geometric constraints in the jth channel. $b_j \geq 0$ is the parameter to control the beam pattern. If $b_j = 1$, the corresponding $\boldsymbol{w}_j(f)$ is estimated to form a delay-and-sum (DS) beamformer [24] toward α to preserve the target. While b_j with a small value can generate a null beamformer to suppress the target, which produces a good estimate of the interference mixture. The overall objective function is

$$J(\mathcal{W},\mathcal{V}) = -\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) + J_b(\mathcal{W}) \tag{9}$$

III. Direction-aware TSE method under underdetermined cases

A. Overview

Under underdetermined cases, the source model needs to handle not only the target speaker's voice but also the interference mixture. Inspired by MVAE and GSS, we proposed a target extraction method that combines a well-designed GC-based GSC structure and the powerful modeling ability of CVAE.

Figure 1 shows the framework. DOA of the target speaker is used to design $J_b(\mathcal{W})$, which creates a null beamformer towards the direction of the target speaker on the interference channel. Such a spatial filter can serve as a blocking matrix (BM) that suppresses the target source and preserve all the other interferences. Whereas in the target channel, a preliminary estimate of the target can be obtained with a generated DS beamformer. Two CVAEs are used to model sources and iteratively update $\mathcal{V}$. The demixing matrix $\mathcal{W}$ can be updated based on the updated $\mathcal{V}$. After that, an ideal ratio Time-Frequency (T-F) mask is calculated using the extracted interference mixture and the observed mixture. Finally, the target signal can be extracted by calculating the product of the T-F mask and target channel output.

B. CVAE-based target and interference models

The research on MVAE shows that CVAE is powerful in source modeling [20]. To extract the target speaker in the underdetermined case of multiple interfering speakers, it is

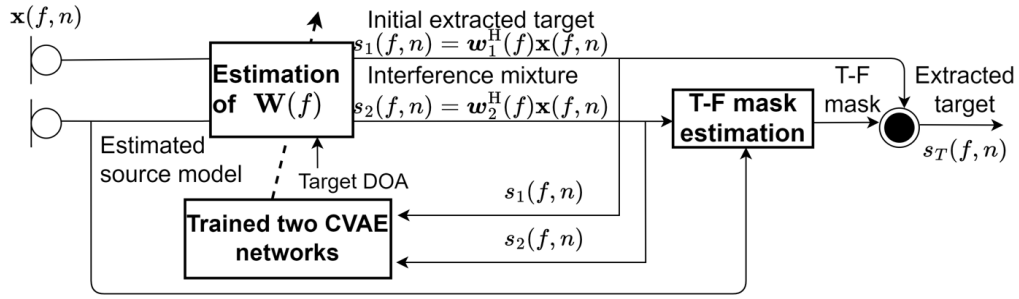


Fig. 1: Framework of the proposed method.

desired to accurately model the single target speech and mixed interfering speech. In this paper, we use two CVAEs to model these two parts. We call these two CVAEs target CVAE (TarCVAE) and interference CVAE (IntCVAE).

Figure 2 shows an illustration of CVAE. Let $\boldsymbol{S} = \{\mathrm{s}(f,n)\}_{f,n}$ be the complex spectrogram of a particular sound source and $\boldsymbol{c}$ be the class label of that source. The encoder network generates a set of parameters for the conditional distribution $q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})$ of a latent space variable $\boldsymbol{z}$ given input data $\boldsymbol{S}$, whereas the decoder network generates a set of parameters for the conditional distribution $p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})$. The network parameters ϕ and θ are trained jointly using labeled samples $\{\boldsymbol{S}_m,\boldsymbol{c}_m\}_{m=1}^M$, where $\boldsymbol{c}_m$ is a one-hot vector that denotes the corresponding class label indicating to which class the spectrogram $\boldsymbol{S}_m$ belongs. In TarCVAE, we use a single target speaker's clean speech to train the model, and the condition $\boldsymbol{c}$ is the class label associated with the IDs of various speakers in a training dataset. For the IntCVAE, mixed interfering speech is used to train the model, and the class label represents the number of interfering speakers in the mixed speech remaining except the target.

In the separation, only the decoder is used to model the source spectrogram by estimating distribution parameters. The decoder can output the variance matrix of sources, which can be used in the estimation of the demixing matrix.

The following objective function is used to train the encoder and decoder networks:

$$\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},\boldsymbol{c})\sim p_D(\boldsymbol{S},\boldsymbol{c})}[\mathbb{E}_{z\sim q_\phi(z|\boldsymbol{S},\boldsymbol{c})}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})] - \mathrm{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})||p(\boldsymbol{z})]], \tag{10}$$

where $\mathbb{E}_{(\boldsymbol{S},\boldsymbol{c})\sim p_D(\boldsymbol{S},\boldsymbol{c})}[\cdot]$ represents the sample mean over the labeled data set and $\mathrm{KL}[\cdot||\cdot]$ is the Kullback–Leibler divergence. The output distribution of the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})$ and the prior distribution of $\boldsymbol{z}$ are given by Gaussian distributions:

$$q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c}) = \prod_k \mathcal{N}(\boldsymbol{z}(k)|\mu_\phi(k;\boldsymbol{S},\boldsymbol{c}),\sigma_\phi^2(k;\boldsymbol{S},\boldsymbol{c})), \tag{11}$$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I}), \tag{12}$$

where $\boldsymbol{z}(k)$, $\mu_\phi(k;\boldsymbol{S},\boldsymbol{c})$, and $\sigma_\phi^2(k;\boldsymbol{S},\boldsymbol{c})$ denote the kth element of $\boldsymbol{z}$, $\mu_\theta(\boldsymbol{S},\boldsymbol{c})$, and $\sigma_\theta^2(\boldsymbol{S},\boldsymbol{c}))$, respectively. The decoder's output distribution $p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g)$ is designed to be a complex Gaussian distribution:

$$p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g) = \prod_{f,n}\mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|0,v(f,n)), \tag{13}$$

$$v(f,n) = g\cdot\sigma_\theta^2(f,n;\boldsymbol{z},\boldsymbol{c}), \tag{14}$$

where $\sigma_\theta^2(f,n;\boldsymbol{z},\boldsymbol{c})$ represents the (f,n)th element of the decoder output $\sigma_\theta^2(\boldsymbol{z},\boldsymbol{c})$ and g is a global-scale parameter of the generated spectrogram.

C. Demixing matrix estimation with target DOA

In the iteratively demixing matrix estimation, the source model $v(f,n)$ of single speech and mixed interference speech estimated by CVAE is used in the first term of the objective function Eq. (9) given by Eq. (7). The update rule for optimizing $\mathbf{W}(f)$ is derived on the basis of the idea adopted in vectorwise coordinate descent (VCD), which is noteworthy for its fast convergence, low computational cost, and nonrequirement of the step-size parameter. We omit the derivation (see [25] for details) here owing to the space limitation. The derived update rules are summarized as

$$\boldsymbol{u}_j = \boldsymbol{D}_j^{-1}\mathbf{W}(f)^{-1}\boldsymbol{e}_j, \tag{15}$$

$$\hat{\boldsymbol{u}}_j = \lambda_j b_j \boldsymbol{D}_j^{-1}\boldsymbol{d}_j, \tag{16}$$

$$h_j = \boldsymbol{u}_j^H \boldsymbol{D}_j \boldsymbol{u}_j, \tag{17}$$

$$\hat{h}_j = \boldsymbol{u}_j^H \boldsymbol{D}_j \boldsymbol{u}_j, \tag{18}$$

$$\boldsymbol{w}_j(f) = \begin{cases} \frac{1}{\sqrt{h_j}}u_j + \hat{u}_j & (\text{if } \hat{h}_j = 0), \\ \frac{\hat{h}_j}{2h_j}[-1+\sqrt{1+\frac{4h_j}{|\hat{h}_j|^2}}]\boldsymbol{u}_j + \hat{\boldsymbol{u}}_j & (\text{o.w.}). \end{cases} \tag{19}$$

where $\boldsymbol{D}_j = \mathbb{E}[\mathbf{x}(f,n)\mathbf{x}^{\mathrm{H}}(f,n)/v_j] + \lambda_j \boldsymbol{d}_j\boldsymbol{d}_j^{\mathrm{H}}$ and $\boldsymbol{e}_j$ is the jth column of the identity matrix. TarCVAE and IntCVAE are used to output the variances v_j as in Eq. (7), whereas their source model parameters are updated using backpropagation. The global scale parameter $\mathcal{G} = \{g_j\}_j$ is updated as

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z},\boldsymbol{c})}. \tag{20}$$

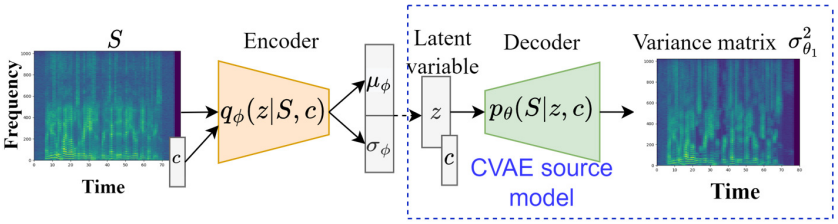


Fig. 2: Illustration of CVAE used in MVAE.

The proposed algorithm is thus summarized as follows:
1. Train θ and ϕ using Eq. (10) in advance.
2. Initialize $\mathcal{W}$ and $\Psi = \{\boldsymbol{z}, \boldsymbol{c}\}$.
3. Iterate the following steps for each j:
   (a) Update $\boldsymbol{w}_j(f)$ using Eqs. (15) to (19).
   (b) Update $\Psi = \{\boldsymbol{z}, \boldsymbol{c}\}$ using backpropagation.
   (c) Update g_j by Eq. (20).
   (d) Update v using Eq. (14).

D. Postprocessing based on T-F mask

Owing to the null constraint towards the target direction, which can serve as a BM, the interference mixture except the desired target can be extracted with high quality at the corresponding channel, whereas the remaining part of the mixture is extracted by a linear calculation, which is considered as the initial extraction of the target. However, despite the good quality of the interference mixture signal obtained, the remainder of the original mixed signal obtained by linear subtraction is not a good extraction result of the target signal. Therefore, we designed a postprocessing method based on a T-F mask to extract the target, which represents the ratio of spectrogram energy of the interference mixture to the observed mixture. The extracted target $s_T(f,n)$ is calculated as

$$IRM = s_1(f,n)(1 - \frac{|s_2(f,n)|^2}{|\mathbf{x}(f,n)|^2}). \quad (21)$$

IV. Experiment Evaluation

A. Experimental conditions

1) Traing of CVAEs: The training data was from the Wall Street Journal (WSJ0) corpus [27]. We used the WSJ0 folder si_tr_s (around 25 h) to train TarCVAE, which contains 101 speakers with 141 sentences per speaker. Speaker identities were considered as the label $\boldsymbol{c}$, which was represented as a 101-dimensional one-hot vector. Whereas for the training of IntCVAE, the training data was generated by mixing clean speech. We used nine groups of mixture speech of 2 to 10 speakers with 200 utterances per group (around 9 h). The label was represented as a nine-dimensional one-hot vector to indicate the number of speakers of the mixture.

2) Evaluation of the modeling power of trained CVAEs: To evaluate the modeling ability of our trained CVAEs on single speech signals and mixed speech signals, we took the clean signal of one speaker and the mixed signal of two speakers as the inputs of TarCVAE and IntCVAE and calculated the source-to-distortions ratio (SDR) between the output reconstructed signal and the original signal. SDR is usually considered to be an overall measure of how good a source quality is. The higher the SDR is, the more similar the CVAE output signal is to the original signal. In the evaluation of the modeling ability of single speech, we randomly selected 50 utterances as test signals from the WSJ0 folders si_dt_05 and si_et_05 where the number of speakers was 18. In the evaluation of mixed speech modeling ability, 50 test signals mixed by two different randomly selected speakers were generated.

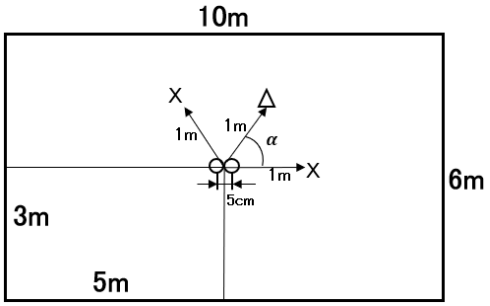
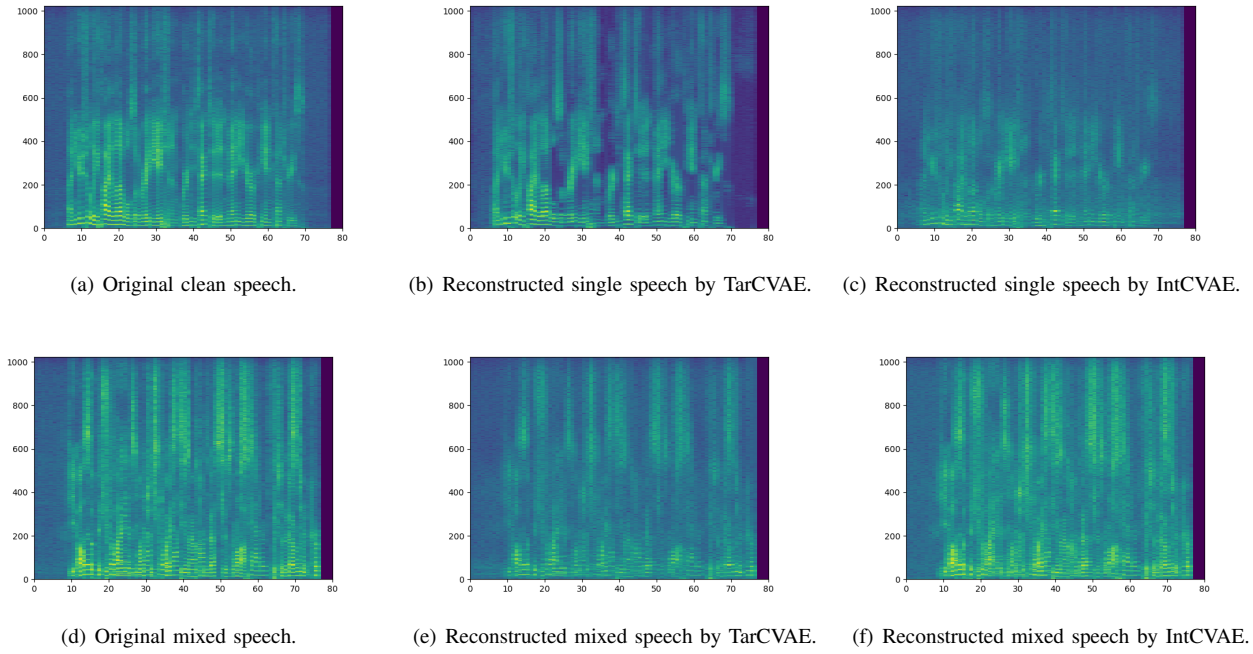


Fig. 3: Configurations of sources and microphones, where △ and × denote the target position and two interference positions, respectively, and α is the DOA of the target relative to the microphone array.

TABLE I: Comparison between baselines and the proposed method.

Method	Application scenario	Source model	Target selection	Processing type
GCIVA	Determined	Laplace	✓	Linear
NL-GCIVA	Underdetermined	Laplace	✓	Nonlinear
MVAE	Determined	CVAE	✗	Linear
Proposed	Underdetermined	CVAE	✓	Nonlinear

3) Evaluation of TSE under underdetermined cases: In the evaluation, test mixture signals were generated by simulating two-channel recordings of three sources where the room impulse responses (RIRs) were synthesized by the image source method (ISM) [26]. Figure 3 shows an example of the relative position of three sources and two microphones. The interval of microphones was set at 5 cm. The evaluations were conducted under three different reverberant conditions with reverberation times (RT_{60}) of 28 ms (an-echoic), 200 ms, and 470 ms. We performed evaluations in three different relative positions of



(a) Original clean speech. (b) Reconstructed single speech by TarCVAE. (c) Reconstructed single speech by IntCVAE.

(d) Original mixed speech. (e) Reconstructed mixed speech by TarCVAE. (f) Reconstructed mixed speech by IntCVAE.

Fig. 4: Magnitude spectrograms of reference sources and reconstructed sources by CVAEs.

sources: the target was located on the left, middle, and right of two interference speakers. In the evaluation of each relative position, three speakers were randomly selected from the WSJ0 folders si_dt_05 and si_et_05. The speakers were randomly located at angles from 0 to 360 degrees. We mixed the images of three speakers with SIR uniformly. We conducted 20 tests at each relative position with different RT_{60}. The average length of the test utterance was 10 seconds.

We chose GCIVA, and MVAE as the baseline methods. To evaluate the effectiveness of the CVAE-based source model, we also applied our designed T-F mask as nonlinear postprocessing to GCIVA as a baseline because our proposed method is similar to GCIVA in terms of framework except for the source model and postprocessing. We call this baseline of nonlinear GCIVA (NL-GCIVA), which could be used under underdetermined cases because of the usage of our designed T-F mask. Table I shows the differences between the baseline methods and our proposed method.

We computed the SDR, source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) between the extracted target and the reference to evaluate the extraction performance. SIR represent the amount of other sources that can be heard in a source estimate, while SAR represent the amount of the true source has with relation to unwanted artifacts. Higher SDR, SIR, and SAR mean better extraction performance. The alignment of the extracted target and the reference signal is important in the evaluation. Since the DOA of the desired speaker α was known, we set the signal in direction α as the reference signal. For our method and other GC-based baselines, the output at the corresponding channel was used as the extracted target. For baselines without GC-based target selection, we evaluated all separated signals and selected the one with the best evaluation result as the extracted target.

TABLE II: Average SDR [dB] of clean signal and mixed signal output by different CVAEs.

	single speech	mixed speech
TarCVAE	18.25	13.65
IntCVAE	15.57	17.74

B. Results

Table II lists the average SDR of reconstructed signals of different CVAE for the input clean signal and mixed signal. The results show that TarCVAE has a better modeling ability for single speech signals than IntCVAE while IntCVAE surpasses TarCVAE in the modeling for the mixed speech signal. Figure 4 shows examples of the CVAE source model fitted to the spectrogram of original clean speech and mixed speech. As these examples show, the CVAE source model was able to express single speech and mixed speech, and TarCVAE is better in modeling the single speech signal while IntCVAE is better in modeling the mixed speech signal.

Table III shows a summary of the evaluation results of extraction performance. Average SDR, SIR, and SAR show that our proposed method exceeded all the baseline methods, especially in terms of SDR and SIR. The comparison between GCIVA and NL-GCIVA reveals that, without improving the source model, the improvement effect of the T-F mask on GCIVA is very limited. The comparison between the proposed

TABLE III: Average SDR, SIR, and SAR [dB] of three-speakers case.

Method	Anechoic		
	SDR	SIR	SAR
GCIVA	9.65	12.67	12.25
NL-GCIVA	9.98	13.05	12.38
MVAE	12.05	13.18	13.06
Proposed	**15.65**	**23.39**	**12.65**
Method	$RT_{60} = 200ms$		
	SDR	SIR	SAR
GCIVA	8.64	11.75	11.80
NL-GCIVA	9.14	12.16	11.97
MVAE	10.84	12.28	12.02
Proposed	**14.32**	**20.28**	**12.37**
Method	$RT_{60} = 470ms$		
	SDR	SIR	SAR
GCIVA	6.34	10.37	9.97
NL-GCIVA	7.13	11.45	10.07
MVAE	8.67	11.68	9.80
Proposed	**12.58**	**18.74**	**11.76**

method and NL-GCIVA reveals that a more powerful source model makes a significant improvement in the extraction performance. The comparison between the proposed method and MVAE implied that the combination of directional information and the CVAE source model has successfully contributed to improving the extraction performance. By comparing the performance of the proposed method with those of other baseline methods under long $RT_{60} = 470$ ms, we found that the proposed method still has high SDR and SIR scores, which demonstrates that our proposed method has high robustness under the condition of strong reverberation.

V. Conclusion

In this paper, we proposed a TSE method under underdetermined cases, which combines geometric constraints and the CVAE-based source model. The key features are that (1) the designed framework can solve the underdetermined problem by applying geometric constraints, and (2) this method takes full advantage of the strong representation power of CVAE to model the source of the target speech and interference mixture. Experimental results revealed that our trained CVAEs could represent the single source and mixture sources, and our proposed method achieved better performance than the conventional GCIVA method and MVAE under the underdetermined condition.

Acknowledgement

This work was partially supported by JST CREST JPMJCR19A3, Japan.

References

[1] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement.* Hoboken, NJ, USA: Wiley, 2018.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* New York, NY, USA: Wiley, 2001.

[3] L. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag*, vol. 30, no. 1, pp. 27–34 ,1982.

[4] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process*, vol. 49, no. 8, pp. 1614–1626, 2001.

[5] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ICA-based blocking matrix for improved noise estimation," *EURASIP J. Adv. Signal Process*, pp. 1–24, 2014.

[6] L.C. Parra and C.V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. SAP*, vol. 10, no. 6, pp. 352–362, 2002.

[7] M. Knaak, S. Araki, and S. Makino, "Geometrically Constrained Independent Component Analysis," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 15, no.2, pp. 715–726, Feb. 2007.

[8] W. Zhang and B. D. Rao, "Combining independent component analysis with geometric information and its application to speech processing," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 3065–3068.

[9] K. Reindl, S. Meier, H. Barfuss, and W. Kellermann, "Minimum Mutual Information-Based Linearly Constrained Broadband Signal Extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1096– 1108, June 2014.

[10] H. Barfuss, K. Reindl, and W. Kellermann, "Informed Spatial Filtering Based on Constrained Independent Component Analysis," in *Audio Source Separation*, Shoji Makino, Ed., pp. 237–278. Springer International Publishing, Cham, 2018

[11] A. Khan, M. Taseska, and E. A. P. Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Proc. Int. Conf. Latent Variable Anal. Signal Separat.* Cham, Switzerland: Springer, pp. 396–403, 2015.

[12] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in Proc. *ICASSP*, pp. 846–850, 2020.

[13] T. Kim, T. Eltoft, and T.W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in Proc. *ICA*, pp. 165–172, 2006.

[14] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in Proc. *ICA*, pp. 601–608, 2006.

[15] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex nonGaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1050–1064, Feb. 2019

[16] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 3545–3558, 2020.

[17] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H.Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Transactions on Audio Speech & Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[18] A.A. Nugraha, L. Antoine, and V. Emmanuel, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.

[19] D. Kitamura, H. Sumino, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono, "Experimental evaluation of multichannel audio source separation based on IDLMA," *IEICE Tech. Rep*, vol. 117, no. 515, pp. 13–20, 2018.

[20] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Comput.*, vol. 31, no. 9, pp. 1891–1914, 2019.

[21] D.P. Kingma, S. Mohamed, D.J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Adv. NIPS*, pp. 3581–3589, 2014.

[22] L. Li, H. Kameoka, and S. Makino, "Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier," in *Proc. of ICASSP'19*, pp. 546–550, 2019.

[23] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time–frequency Gaussian source models,"

in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '05)*, Mohonk, NY, USA, pp. 78–81, Oct. 2005.
[24] K. Buckley, "An adaptive generalized sidelobe canceller with derivative constraints," *IEEE Trans. AP*, vol. 34, no. 3, pp. 311–319, 1986.
[25] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in *Proc. ICASSP*, pp. 746–750, 2018.
[26] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[27] D.B. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop on Speech and Natural Language*. ACL, pp. 357–362, 1992.