Dual-Channel Target Speaker Extraction Based on Conditional Variational Autoencoder and Directional Information

Rui Wang, Li Li, *Member, IEEE*, and Tomoki Toda, *Senior Member, IEEE*

***Abstract*—Target speaker extraction (TSE) has become an attractive research topic in recent years. However, TSE under the underdetermined conditions is still a challenge. In this paper, we deal with a dual-channel TSE problem under underdetermined conditions. Geometric source separation (GSS) is used to be a solution to the TSE problem, but the performance of conventional GSS methods is limited under underdetermined conditions because of the lack of a powerful source model. We propose a dual-channel TSE method with the combined capabilities of target selection based on geometric constraints, more powerful source modeling, and nonlinear postprocessing. A geometric constraint (GC) on the target direction of arrival (DOA) is applied to select the target, and two conditional variational autoencoders (CVAEs) are used to model a single speaker's speech and interference mixture speech. For 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. Moreover, to overcome the impact of DOA estimation errors, we improve the objective function so that the target DOA information can be modified. 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-distortion ratio (SDR) and source-to-interference ratio (SIR), respectively, under medium reverberation for 470 ms. Furthermore, through the analysis of experimental results, we found that the improvement method is robust against DOA estimation errors.**

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

I. Introduction

The human brain has a remarkable capacity to selectively direct auditory attention to a specific sound amidst various interferences. This is known as selective auditory attention. However, machines have yet to achieve the same target speaker selection capability as humans [1], [2], [3]. Many efforts have been spent on its engineering solution, which yields the research on target speaker extraction (TSE). TSE has become an important research topic in the field of signal processing and has many practical applications such as speech enhancement, speech recognition, and hearing aids [4].

One of the important techniques related to TSE is blind source separation (BSS) [5], [6], [7], [8], [9], [10]. BSS aims to recover all signals from a mixed signal. Over the past few decades, various BSS methods have been proposed. As the first class of BSS methods, independent component analysis (ICA), which is based on the linear mixing and demixing processing and the independence of the assumed source model of a single source [11], has been well studied in the fields of statistics and information theory. The frequency domain (FD) ICA (FD-ICA) has the advantage of faster convergence than time domain deconvolution [12], [13], [14], [15]. However, it suffers from a permutation issue, which refers to the inconsistency of output channels among frequency bins. Independent vector analysis (IVA) is a multivariate extension of ICA. It tackles the convolutedly mixed BSS problem in a way that the permutation problem is avoided by employing a multivariate source prior to short-time Fourier transform (STFT) components [16], [17].

In realistic implementations, recovering all mixed sources is often unnecessary, as only one or a few specific sources are needed. Conventional BSS methods face challenges in selectively extracting specific sound sources, whereas TSE is tailored to output a single source. Several efforts have been dedicated to TSE. Early works such as those in [18] and [19] present target-dependent TSE methods, which are capable of extracting speech from a specific speaker but lacking generalizability to other speakers. To achieve speaker-adaptive methods, prior knowledge or auxiliary information is required to specify the target. Recently extracting speaker information from audio samples has proven to be an effective approach. Some frequency-domain methods like SpeakerBeam [20] and VoiceFilter [21] have been proposed to isolate the target speaker from a mixture using the target speaker information extracted from an adaptation utterance or a reference signal. Some time-domain methods such as SpEx+ [22] have also attracted attention by adapting speaker encoder. In addition, the visual feature is another feature widely applied in many frameworks. Some visual-based methods have been proposed to leverage visual cues from the target speaker, such as lip movements [23], [24], [25], [26] or cropped facial frames [27], [28].

Another effective approach is using spatial properties of sound sources. There are several frameworks have been studied for

Manuscript received 8 August 2023; revised 9 February 2024; accepted 5 March 2024. Date of publication 14 March 2024; date of current version 25 March 2024. This work was supported in part by JST CREST under Grant JPMJCR19A3, Japan, and in part by Nagoya University Research Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jong Won Shin. *(Corresponding author: Rui Wang.)*

Rui Wang is with the Graduate School of Informatics, Nagoya University, Aichi 464-8601, Japan (e-mail: rui.wang@g.sp.m.is.nagoya-u.ac.jp).

Li Li and Tomoki Toda are with Information Technology Center, Nagoya University, Aichi 464-8601, Japan (e-mail: li.li@g.sp.m.is.nagoya-u.ac.jp; tomoki@icts.nagoya-u.ac.jp).

Digital Object Identifier 10.1109/TASLP.2024.3376154

decades. Traditionally, the spatial filter is achieved by a beamformer that aims at suppressing signal components from other than the target direction. With the development of deep neural networks (DNN), some research has shown that the DNN-based nonlinear spatial filter outperform traditional beamformers plus postfiltering schemes [29]. Recently, some studies on neural beamformer has achieved remarkable results, demonstrating the effectiveness of DNN on traditional beamformer methods [30].

On the other hand, some studies have also shown the potential of combining spatial information with traditional BSS frameworks. Geometric source separation (GSS) [31], [32], [33], [34], [35] has been proposed by adding spatial information into the BSS framework. For example, in geometrically constrained independent vector analysis (GCIVA), a linear geometric constraint (GC) based on the direction of arrival (DOA) of the target is introduced into an IVA framework [36]. GCIVA utilizes a generalized sidelobe canceller (GSC) [37], [38] structure to generate a fixed beamformer that enhances the target signal and a null beamformer [39] for estimating interferences by suppressing the target, where the number of beams generated is limited by the available number of microphones. Compared to the non-linear spatial filter and neural beamformer, the GSS method based on the BSS framework uses signal independence as the basis for separating different sound sources, which relies less on prior spatial knowledge. Most GSS methods do not require a large amount of training data, and prior spatial information or environmental information is not needed in the training. It only requires target spatial information to generate geometric constraints to achieve target person selection in the inference stage.

Most GSS and localized TSE methods, including GCIVA, are designed for determined cases, where the number of microphones M is equal to the number of sources N. However, in realistic applications, hardware limitations often lead to underdetermined conditions, posing challenges for traditional GSS methods. A major challenge to applying GSS under underdetermined conditions is the limitation of a source model. In cases under underdetermined conditions, a more powerful source model is required because the source model needs to deal not only with the target speech but also with the mixture of interference speakers. Many efforts have been made in developing the source model of a speech signal. In independent low-rank matrix analysis (ILRMA), a flexible source model of nonnegative matrix factorization (NMF) decomposition was applied in the IVA framework, which yielded a higher modeling power of complex spectral structures than the former IVA with a Laplace distribution-based source model [40]. Recently, a Bayesian framework-based method has been proposed to introduce a background source (BG) model derived by independent vector extraction (IVE) [41] that allows for underdetermined cases to extract the source of interest (SOI) [42]. However, these methods rely on fixed statistical models and have limited flexibility when dealing with different numbers of speakers. Most recently, a deep neural network (DNN) has been used to model source spectral characteristics owing to its powerful modeling capability [43], [44]. The multichannel variational autoencoder (MVAE) method [45] utilizes the conditional variational autoencoder (CVAE) [46] as the generative source model in an IVA framework in determined conditions and has attracted attention. An MVAE trains a CVAE using power spectrograms of clean speech samples and the corresponding speaker index (ID) as an auxiliary label input so that the trained decoder output distribution can be used as a universal generative model of source signals, which has shown an impressive performance under determined conditions owing to its representation power.

To solve the underdetermined problem, we previously proposed an innovative and robust dual-channel TSE method under underdetermined conditions, which combines GC-based on prior spatial information, the CVAE-based source model, and time–frequency (T–F) mask-based postprocessing [47]. On the basis of an MVAE, we innovatively modeled the target and interference mixture separately using two types of CVAE to better handle underdetermined cases. In our method, we focused on the BSS framework and designed an iterative TSE algorithm on the basis of the GSC structure with linear GCs and the target and interference mixture source modeling with CVAEs. In the implementation of the proposed method, two CVAEs were trained by single speech and multi-speaker mixed speech, excluding any prior spatial information. The DOA of the target is assumed to be a piece of known prior information, which will be manually provided or estimated in the test stage.

However, there are some remaining issues in our previous proposed method. Due to GSS being an algorithm based on signal statistical independence in the BSS framework, the method of selecting target speakers is based on GCs, which require the target DOA to be known. Therefore the inaccuracy of spatial information, such as errors caused by DOA estimation, often leads to the severe degradation of GSS [48], [49]. This problem also exists in our previous proposed method. In addition, under underdetermined conditions, the impact of this error will be more pronounced because of the influence of the number of microphones on the generated beams in GSS. In addition, the impact of the angle between sources and the distance between sources and microphones is still unclear.

In this paper, to overcome the negative impact of DOA estimation errors, we propose a robust TSE algorithm against DOA estimation errors based on the former framework. We improve the original objective function by adding a new variable as the target DOA and an L2-NORM regularize so that DOA information can be modified during the process of estimating the demixing matrix. Note that this paper is an extended full-paper version of our conference paper [47]. The additional contributions of this paper are as follows.

1) In order to further evaluate the performance of the previously proposed method under different sound field environments, we conducted a series of extended experiments to explore the impact of angles between different speakers and the distance between speakers and microphones.
2) We introduced an L2-norm constraint based on the given DOA into the demixing matrix estimation's objective function, leading to an enhanced algorithm. Different from directly using the target DOA in [30] and [36], this improvement allows for DOA modification, thereby reducing the adverse effects of DOA estimation errors.

The remainder of the paper is structured as follows. Problem formulation is presented in Section II. In Section III, we review the related works of MVAEs and CVAEs. Subsequently, we describe completely and in detail the proposed framework in Section IV of our work in [47]. After that, we propose our new method to address the problem of DOA errors in Section V. Experimental results are presented in Section VI. Finally, we make a conclusion in Section VII.

II. Problem Formulation

Consider a TSE problem under the underdetermined condition where a dual-channel microphone array is used. Let $\mathbf{s}(f,n)$ and $\mathbf{x}(f,n)$ be the STFT coefficients of the source signals and a set of microphone signals, where f and n are the frequency and time indices, respectively. We express them as

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

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

where $s_1(f,n)$ is the target with a known DOA and $s_2(f,n)$ is the interference mixture excluding the target. $x_1(f,n)$ and $x_2(f,n)$ are the observed signals of two input microphones. 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 the 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, whereas $\boldsymbol{w}_2(f)$ is used to estimate the interference by suppressing the target. Due to the challenge of suppressing interference mixtures with a linear filter in underdetermined conditions, estimating the target accurately in such scenarios is difficult [50]. On the other hand, it is still possible to suppress the target using the linear filter to estimate the interference mixture.

Let us assume that source signals follow the local Gaussian model (LGM), 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 (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

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

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

Now, let us consider that GCs [31] restrict the far-field response of the jth demixing filter in the target DOA α, which is described as

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

$$\boldsymbol{d}(f,\alpha) = \exp[-j(\mathbf{p}/c)f\cos(\alpha)], \tag{9}$$

where $\boldsymbol{d}(f,\alpha)$ is the steering vector toward α, $\mathbf{p} = [p_1, p_2]$ are the positions of two microphones, and c is the wave propagation speed. λ_j is a weighting parameter and $b_j \geq 0$ is the parameter for controlling the beam pattern. This concept has been used in the linearly constrained minimum variance (LCMV) beamformer [51]. If $b_j = 1$, the corresponding $\boldsymbol{w}_j(f)$ is estimated to form a delay-and-sum (DS) beamformer [52] toward α to preserve the target. On the other hand, a small b_j value can generate a null beamformer to suppress the target, which produces a good estimate of the interference mixture. The overall objective function is

$$J(\mathcal{W},\mathcal{V}) = -\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) + J_{gc}(\mathcal{W}). \tag{10}$$

III. Related Works

A. Geometric Constraint-Based IVA

In the traditional source separation method, the demixing matrix-based source separation processing aims to just separate the observed mixture signals into the individual source signals. However, real-world applications often require additional information to select target speech post-separation, addressing output-channel permutation issues. As we mentioned in the introduction section, there are three main frameworks for utilizing spatial information to implement TSE, in which BSS has shown its own merit in incorporating signal independence and spatial information. In our paper, we focus on the BSS framework-based method.

Incorporating spatial information into BSS demixing filters has been approached in two main ways. One involves using spatial data as prior knowledge to optimize the demixing matrix, such as the Bayesian framework-based IVA method utilizing a spatially informed prior [42], which extends the original algorithm of IVA to the maximum *a posteriori* (MAP) method. The other is GSS which integrates GCs into traditional BSS methods. A notable example is GSS like GCIVA, which merges a linear GC with the IVA framework. The upcoming section will delve into the specifics of the GCIVA algorithm.

The fundamental framework in GCIVA is IVA, which assumes that sources follow a multivariate super Gaussian distribution; thus, dependences over frequency bins can be utilized to avoid the inner permutation. The objective function for estimating the demixing matrix $\mathcal{W}$ in IVA is given by

$$J_{\mathrm{IVA}}(\mathcal{W}) = \sum_{f,n}\mathbb{E}[G(s_j(f,n))] - 2\sum_f \log|\det\mathbf{W}(f)|. \tag{11}$$

Here, $\mathbb{E}[\cdot]$ denotes the expectation operator and $G(s_j(f,n)) = -\log p(s_j(f,n))$ is the contrast function, where $p(s_j(f,n))$ represents a multivariate probability density function of the jth source. The objective function of the GCIVA method is

$$J_{\mathrm{GCIVA}}(\mathcal{W}) = J_{\mathrm{IVA}}(\mathcal{W}) + J_{gc}(\mathcal{W}), \tag{12}$$

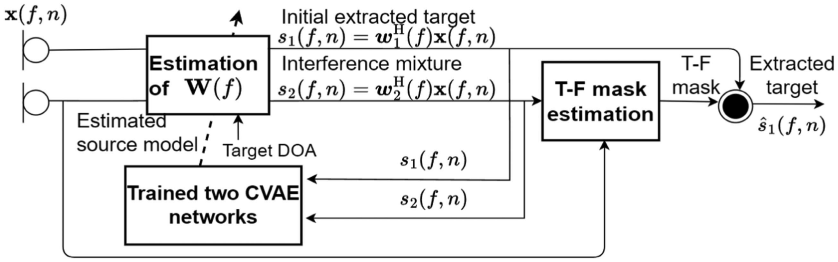


Fig. 1. Framework of the proposed method.

where the GC is given by (8) to restrict the far-field response of the jth estimated demixing filter by using the target DOA.

B. VAE-Based Methods

Methods like IVA, GCIVA, and IVE, despite integrating spatial information, face limitations under underdetermined conditions since traditional source models such as the Laplacian distribution in GCIVA and the SOI and BG models in IVE are not powerful enough in modeling complex spectrogram structures, such as a mixture of multi speakers. To overcome these limitations, deep generative models such as VAEs and GANs, as highlighted in recent studies [53], [54], [55], offer advanced solutions. These models excel in learning complex data distributions, which traditional source models struggle to represent. Innovations in this area, as demonstrated by Bando et al. [56] and others [57], [58], [59], include the application of VAEs for enhanced noise modeling and speech separation, merging them with techniques like NMF and class supervision to boost performance.

The use of conditional VAEs, where the decoder network is conditioned on additional information, has also been explored and shown to improve separation performance in certain scenarios. The research on MVAE first introduced the CVAE model in multi channel speech separation. An MVAE trains a conditional VAE (CVAE) [46] using power spectrograms of clean speech samples and the corresponding speaker ID as auxiliary label inputs so that the trained decoder output distribution can be used as a universal generative model of source signals. Although an MVAE showed impressive performance in determined cases, it is still limited in underdetermined cases.

IV. Direction-Aware TSE Method Under Underdetermined Conditions

A. Overview

In this section, we detail our proposed direction-aware TSE approach for underdetermined conditions, addressing two main challenges. Firstly, we incorporate linear GCs based on the target's DOA to select the target in the underdetermined TSE problem. Secondly, to handle both target speech and interference in complex scenarios, we introduce a novel CVAE, named Interference CVAE (IntCVAE). IntCVAE is designed to effectively model mixed speech signals, particularly in situations involving varying numbers of speakers.

Fig. 1 shows the framework. The DOA of the target is used to design $J_{gc}(\mathcal{W})$ on two channels. On channel 1, also called the target channel, the parameter b_1 in the GC given by (8) is set to 1 to create a delay-and-sum (DAS) beamformer, which yields a spatial beamformer towards the direction of the target. So the GC on the target channel is shown as

$$J_{gc}^{\text{one}}(\mathcal{W}) = \lambda_1 \sum_f |\boldsymbol{w}_1^{\mathrm{H}}(f)\boldsymbol{d}(f,\alpha) - 1|^2. \tag{13}$$

A preliminary estimation of the target can be obtained by calculating

$$s_1(f,n) = \boldsymbol{w}_1^{\mathrm{H}}(f)\mathbf{x}(f,n). \tag{14}$$

On channel 2 on the other hand, also called the interference channel, $b_2 = 0$ is set in (8) to generate a null beamformer, which serves as a blocking matrix (BM) to suppress the target source and preserves all the other interferences. The GC on the interference channel is given as

$$J_{gc}^{\text{null}}(\mathcal{W}) = \lambda_2 \sum_f |\boldsymbol{w}_2^{\mathrm{H}}(f)\boldsymbol{d}(f,\alpha)|^2. \tag{15}$$

We can obtain the interference mixture from the output of the interference channel as

$$s_2(f,n) = \boldsymbol{w}_2^{\mathrm{H}}(f)\mathbf{x}(f,n). \tag{16}$$

Two CVAEs are used to model sources. The set of demixing matrices $\mathcal{W}$ can be updated on the basis of the updated $\mathcal{V}$. Subsequently, an ideal ratio 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

To extract the target speaker in the underdetermined case of multiple interfering speakers, it is desired to accurately model the single target speaker's speech and interference mixture speech. We use two CVAEs to model these two parts, which are called target CVAE (TarCVAE) and IntCVAE.

Figs. 2 and 3 show illustrations of TarCVAE and IntCVAE. TarCVAE has been applied in MVAE [45]. Let $\boldsymbol{S} = \{\mathbf{s}(f,n)\}_{f,n}$ be the complex spectrogram of an input sound source and $\boldsymbol{c}$ be the conditional variable of that source. In TarCVAE, $\boldsymbol{S}$ represents the clean speech of one single speaker, and $\boldsymbol{c}$ represents this speaker's identity. In IntCVAE on the other hand, the input $\boldsymbol{S}$ is a mixture of different speakers, and $\boldsymbol{c}$ represents the number

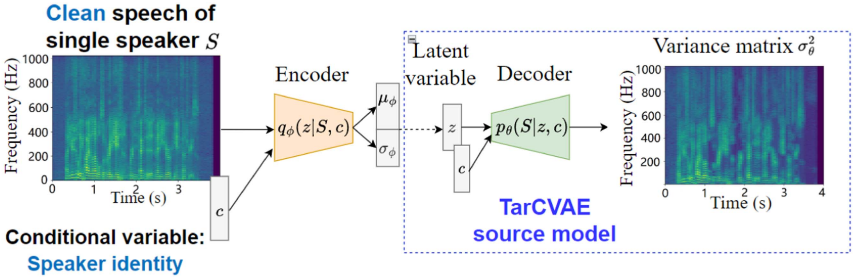


Fig. 2. Illustration of TarCVAE.

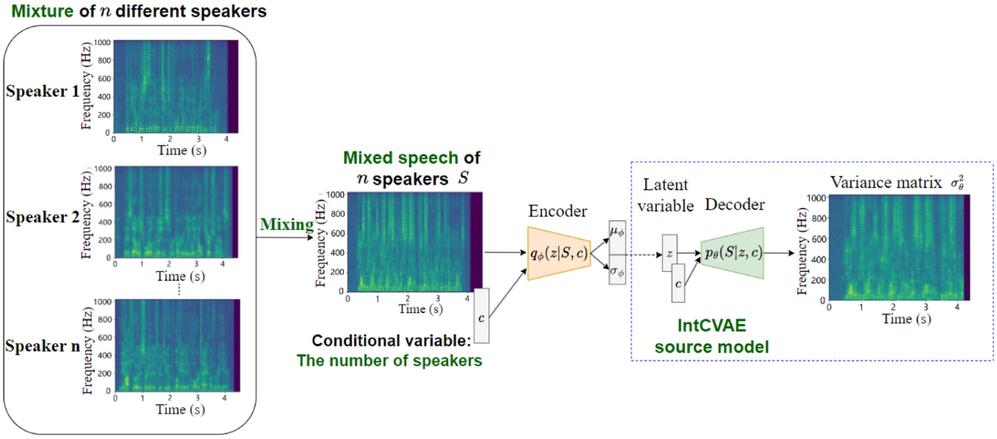


Fig. 3. Illustration of IntCVAE.

of speakers in this mixture. 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 the 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 the separation, only the decoder is used to model the source spectrogram by estimating the latent space variable z and the conditional variable c as the source model 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:

$$\begin{aligned}\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},\boldsymbol{c})\sim p_D(\boldsymbol{S},\boldsymbol{c})}[\mathbb{E}_{z\sim q_\phi(z|\boldsymbol{S},\boldsymbol{c})}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})] \\ - \mathrm{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})||p(\boldsymbol{z})]],\end{aligned} \tag{17}$$

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:

$$q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c}) = \prod_k \mathcal{N}(\boldsymbol{z}(k)|\mu_\phi(k;\boldsymbol{S},\boldsymbol{c}),\sigma_\phi^2(k;\boldsymbol{S},\boldsymbol{c})), \tag{18}$$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I}), \tag{19}$$

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 $\boldsymbol{z}$, the mean vector $\mu_\theta(\boldsymbol{S},\boldsymbol{c})$, and the variance vector $\sigma_\theta^2(\boldsymbol{S},\boldsymbol{c})$, respectively. The decoder's output distribution $p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g)$ is designed to be a complex Gaussian distribution:

$$p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g) = \prod_{f,n} \mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|0,v(f,n)), \tag{20}$$

$$v(f,n) = g\cdot\sigma_\theta^2(f,n;\boldsymbol{z},\boldsymbol{c}), \tag{21}$$

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 a single target speaker's speech and interference mixture's speech obtained by CVAE is used in the first term of the objective function, which is given by (7).

The update rule for $\mathbf{W}(f)$ is derived based on 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 [60] for details) here owing to space limitations. The derived update rules are summarized as

$$\boldsymbol{u}_j = \boldsymbol{D}_j^{-1}\mathbf{W}(f)^{-1}\boldsymbol{e}_j, \tag{22}$$

$$\hat{\boldsymbol{u}}_j = \lambda_j b_j \boldsymbol{D}_j^{-1}\boldsymbol{d}_j, \tag{23}$$

$$h_j = \boldsymbol{u}_j^H\boldsymbol{D}_j\boldsymbol{u}_j, \tag{24}$$

$$\hat{h}_j = \boldsymbol{u}_j^H\boldsymbol{D}_j\hat{\boldsymbol{u}}_j, \tag{25}$$

Algorithm 1: CVAE-Based TSE.

```
Require: Network parameters $\theta$ and $\phi$ of two CVAEs
  trained using (17), observed mixture signal $\mathbf{x}(f,n)$,
  iteration number $L$
randomly initialize $\mathcal{W}$ and $\Psi = \{\boldsymbol{z}, \boldsymbol{c}\}$
initialization: update $\mathcal{W}$ using a BSS method such as
  ILRMA
for $l = 1$ to $L$ do
  for $j = 1$ to 2 do
    $s_j(f,n) = \boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)$
    (updating parameters of source model)
    initialize $g_j$ using (27)
    for $k = 1$ to 100 do
      update $\Psi$ using BP with $\log p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})$
       while keeping $\theta$ fixed
    end for
    compute $v(f,n)$ using (21)
    (updating demixing matrices)
    update $\boldsymbol{w}_j(f)$ using (22) to (26)
  end for
end for
```

$$\boldsymbol{w}_j(f) = \begin{cases} \frac{1}{\sqrt{h_j}}\boldsymbol{u}_+\hat{\boldsymbol{u}}_j & (\text{if } \hat{h}_j = 0), \\ \frac{\hat{h}_j}{2h_j}\left[-1 + \sqrt{1 + \frac{4h_j}{|\hat{h}_j|^2}}\right]\boldsymbol{u}_j + \hat{\boldsymbol{u}}_j & (\text{o.w.}), \end{cases} \tag{26}$$

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

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z},\mathbf{c})}. \tag{27}$$

The proposed algorithm is thus summarized as follows.

D. Postprocessing Based on T–f Mask

By applying a null constraint toward the target direction, we can extract the non-target interference with high quality. However, the residual signal obtained after subtracting this interference isn't an effective extraction of the target speech. To improve this, we've developed a postprocessing technique using a T-F mask, which calculates the ratio between the spectrogram energies of the interference and the observed mixtures, thereby enhancing the extraction of the target signal. The extracted target $\hat{s}_1(f,n)$ is

$$\hat{s}_1(f,n) = s_1(f,n)\left(1 - \frac{|s_2(f,n)|^2}{|\mathbf{x}(f,n)|^2}\right). \tag{28}$$

V. Improved TSE Method Against DOA Errors

A. Impact of DOA Errors

In an acoustic environment, whether for GSS or other TSE methods based on spatial information, the DOA is one of the most prevalent and critical information in calculating geometric constraints or generating beamformers. Accurate DOA information is required for such systems. However, estimating the DOA of the speaker is not simple. Researchers found that in many practical applications, the error of DOA information will bring significant errors to the steering vector, which is the main reason for the degradation of the performance in many systems [48], [49]. Especially under underdetermined conditions, where the number of generated beams in GSS is limited by the number of microphones, errors of the given DOA will lead to the wrong steering vector. In the field of robust adaptive beamformer, the challenge of inaccurate DOA information is commonly addressed as a DOA mismatch issue.

Over the years, several attempts have been made to address this issue. Some significant research has been made in developing robust adaptive beamformer methods, particularly in enhancing their resilience to steering vector inaccuracies [61]. Among them, the imposition of multiple linear constraints along with minimum variance beamformer has been considered a useful method [62], [63], [64], [65]. These methods are designed to widen the main beam in the beampattern, compensating for uncertainties in the DOA information. However, adding these extra constraints reduces the beamformer's degrees of freedom, limiting its capability to suppress unwanted signal components. The error in DOA remains in the calculation of the steer vector. As long as there is a fixed error in the DOA that is given to the system and it cannot be modified in the process of estimating the beam, such DOA mismatch will inevitably bring errors to the calculation of the steel vector.

Therefore, to address this problem, in this section, we propose a robust TSE algorithm against DOA estimation errors based on the former framework. We improved our objective function of the estimation of the demixing matrix to enable the given DOA can be updated in this processing.

B. Improved Method With DOA Modification

In estimating the demixing matrix, the objective function is shown by (10). In the second part of this equation, which is the linear GC, the DOA α is fixed. In this case, if the given α is different from the true direction of the target, the steering vector $\boldsymbol{d}(f,\alpha)$ is forced toward the wrong location instead of the desired target in the direction α. This mismatch will cause the extracted source in this direction to contain residues of other audio sources. To solve this problem, we improve the original proposed objective function by adding the L2-NORM of the target DOA as the regularizer. The term of the L2-NORM of the target DOA α is calculated as

$$J_c(\alpha|\alpha_0) = \lambda_\alpha ||\alpha - \alpha_0||_2^2. \tag{29}$$

In this regularizer, α_0 is the estimated result of the target DOA, which is known in advance as prior information in our system, whereas α is the DOA target used to calculate the geometric constraint $J_{gc}(\mathcal{W},\alpha)$, which is set as a variable and can be updated in the process of estimating the demixing matrix. The improved objective function is shown as

$$\begin{aligned}\mathcal{L}(\mathcal{W},\mathcal{V},\alpha) \\ = -\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) + J_{gc}(\mathcal{W},\alpha) + J_c(\alpha|\alpha_0).\end{aligned} \tag{30}$$

Algorithm 2: CVAE-Based TSE With DOA Modification.

```
Require: Network parameters $\theta$ and $\phi$ of two CVAEs
  trained using (17), observed mixture signal $\mathbf{x}(f,n)$,
  iteration number $L$, estimated DOA $\hat{\alpha}$ is given.
randomly initialize $\mathcal{W}$ and $\Psi = \{\boldsymbol{z}, \boldsymbol{c}\}$
initialization: update $\mathcal{W}$ using a BSS method such as
  ILRMA
for $l = 1$ to $L$ do
  for $j = 1$ to 2 do
    $s_j(f,n) = \boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)$
    (updating parameters of source model)
    initialize $g_j$ using (27)
    for $k = 1$ to 100 do
      update $\Psi$ using BP with $\log p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})$
       while keeping $\theta$ fixed
    end for
    compute $v(f,n)$ using (21)
    (updating demixing matrices)
    update $\boldsymbol{w}_j(f)$ using (22) to (26)
    (updating DOA)
    for $h = 1$ to 100 do
      update DOA $\alpha$ using GD with (30), (8),
      and (9)
  end for
end for
```

To obtain the optimal DOA, we adopt a GD (gradient descent)-based algorithm to update α. In this algorithm, the DOA after each iteration will be used to correct the geometric constraints in the next iteration. Based on the updating rule of the demixing matrix in (26), the improved algorithms with variable DOAs are summarized as follows,

VI. Experimental Evaluation

A. Dataset for Training

The training data was from the Wall Street Journal (WSJ0) corpus [66]. 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 label $\boldsymbol{c}$, which was presented by a 101-dimensional one-hot vector. Whereas for the training of IntCVAE, the training data was generated by linearly mixing clean speeches without additional background noise. We used nine groups of a mixture of speeches of 2 to 10 speakers with 200 utterances per group (around 9 h). The label was presented by a nine-dimensional one-hot vector to indicate the number of speakers of the mixture. In these mixtures, each source's energy was kept equal, ensuring a linear and uniform mixing of the speech signals. This method maintains consistent energy levels across all sources, resulting in an evenly balanced audio mix where no single speaker's voice dominates the composite signal.

B. Evaluation of the Reconstruction Power of Trained CVAEs

To evaluate the reconstruction ability of our trained CVAEs on single speech and mixed speech signals, we took the clean signals of one speaker and the mixed signals of two speakers as the inputs of TarCVAE and IntCVAE and calculated the source-to-distortions ratio (SDR) of the output reconstructed signal to the original signal. The higher the SDR is, the more similar the CVAE output signal is to the original signal. In the evaluation of the reconstruction capability of a 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 the mixed-speech reconstruction capability, 50 test signals mixed from two different randomly selected speakers were generated.

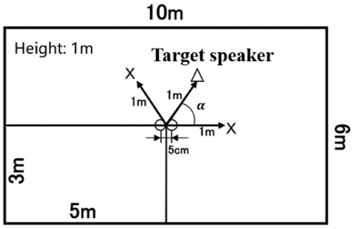


Fig. 4. Configuration of evaluation, where △ and × denote the target and interferences, respectively, and α is the DOA of the target.

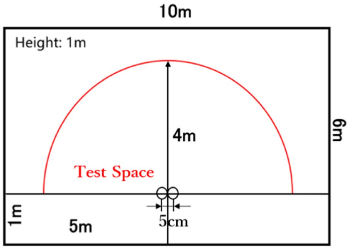


Fig. 5. Configurations of the test space.

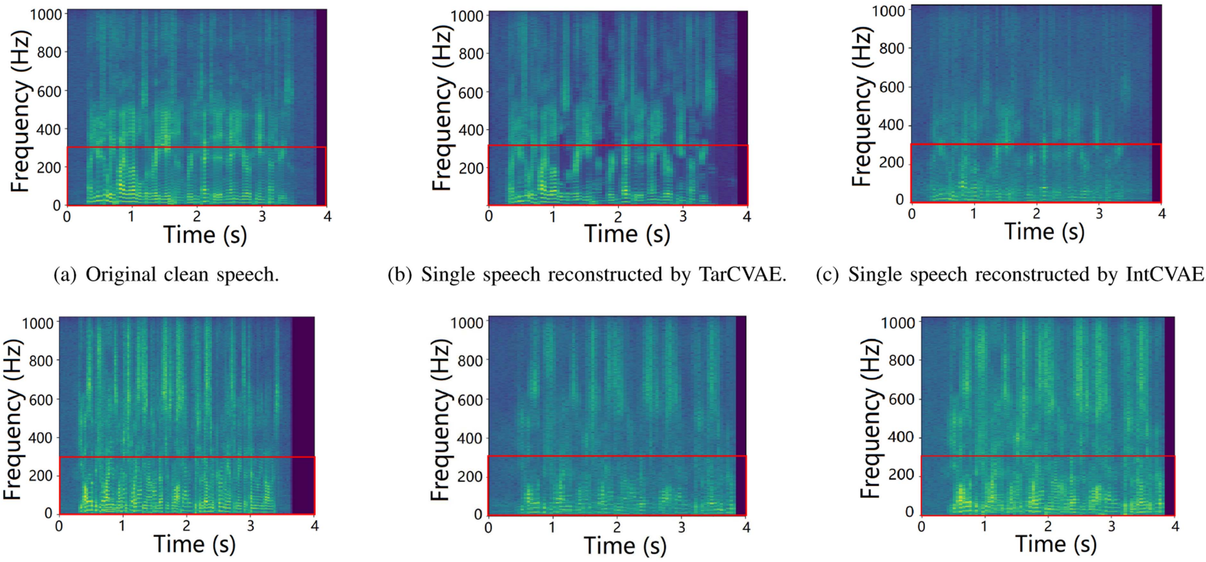
Table II shows the average SDRs of signals reconstructed by different CVAEs for the input clean and mixed signals. The results show that TarCVAE has a better reconstruction capability for single speech signals than IntCVAE, whereas IntCVAE surpasses TarCVAE in the reconstruction for mixed speech signals. Fig. 6 shows examples of the CVAE source model fitted to the spectrogram of the original clean and mixed speech. As shown by Fig. 6(a)–(c), we can see that spectral structures of the single speech especially in the low-frequency range are more precisely reconstructed by TarCVAE than those by IntCVAE. As for Fig. 6(c), (d), and (e), we can see that IntCVAE can reconstruct the spectral structures of the mixed speech more precisely than TarCVAE. In contrast to the MVAE method, which solely employs TarCVAE, our approach's inclusion of IntCVAE is particularly beneficial for effectively handling mixed speech under underdetermined conditions.

C. Evaluation of TSE in Underdetermined Cases

In the evaluation, test mixture signals were generated by simulating two-channel recordings of three sources where room impulse responses (RIRs) were synthesized by the image source method (ISM) [67]. Fig. 4 shows an example of the relative position of three sources and two microphones. The interval of microphones was set at 5 cm. The evaluation was conducted

TABLE I
COMPARISON BETWEEN BASELINE METHODS AND PROPOSED METHODS

Method	Application scenario	Source model	Target selection	Post filter	DOA modification
GCIVA	Determined	Laplace	✓	Linear	N/A
NL-GCIVA	Underdetermined	Laplace	✓	Nonlinear	N/A
MVAE	Determined	TarCVAE	N/A	Linear	N/A
NL-MVAE	Underdetermined	TarCVAE	N/A	Nonlinear	N/A
Proposed method	Underdetermined	TarCVAE + IntCVAE	✓	Nonlinear	N/A
Proposed method with DOA modification	Underdetermined	TarCVAE + IntCVAE	✓	Nonlinear	✓



(a) Original clean speech.

(b) Single speech reconstructed by TarCVAE.

(c) Single speech reconstructed by IntCVAE.

(d) Original mixed speech.

(e) Mixed speech reconstructed by TarCVAE.

(f) Mixed speech reconstructed by IntCVAE.

Fig. 6. Magnitude spectrograms of reference sources and sources reconstructed by CVAEs.

TABLE II
AVERAGE SDRS [DB] OF CLEAN SIGNAL AND MIXED SIGNAL OUTPUTS OBTAINED BY DIFFERENT CVAES

	single speech	mixed speech
TarCVAE	18.25	13.65
IntCVAE	15.57	17.74

under three different reverberant conditions with reverberation times (RT_{60}) of 28 ms (anechoic), 200 ms, and 470 ms. Three speakers were randomly selected from the WSJ0 folders si_dt_05 and si_et_05. Three speakers were randomly located at angles from 0° to 180°, in different directions, with the minimum angle between speakers set to 10 degrees. We mixed the images of three speakers using SIR uniformly. We conducted 60 tests under each reverberation condition. The average length of the test utterance was 10 seconds.

We selected GCIVA and MVAE as the baseline methods, and to conduct an ablation study on different components of our proposed methods, we incorporated our designed T–F mask into GCIVA and MVAE for nonlinear postprocessing, resulting in two additional baselines, namely, nonlinear GCIVA (NL-GCIVA) and nonlinear MVAE (NL-MVAE). These nonlinear variant methods can be utilized in underdetermined cases owing to the designed T–F mask. Table I presents a comparison between the baseline and proposed methods.

We computed the SDR, SIR, and source-to-artifact ratio (SAR) of the extracted target to the reference signal to evaluate the 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 the direction α as the ground truth. 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 III shows the evaluation results of the extraction performance. Our proposed method outperforms all baseline methods, particularly in terms of SDR and SIR. By comparing GCIVA with NL-GCIVA and MVAE with NL-MVAE, we can clearly see that the T–F mask's improvement effect on performance is limited without enhancing the source model. By comparing NL-MVAE with NL-GCIVA and the proposed method, we find that a more powerful source model significantly improves the extraction performance. The proposed method, combining directional information and the CVAE source model, successfully enhances the extraction performance, as observed in its comparison with all baseline methods. Furthermore, when considering a medium reverberation time of $RT_{60} = 470$ ms, the proposed method

TABLE III
AVERAGE SDR, SIR, AND SAR [DB] OF THREE-SPEAKER 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
NL-MVAE	12.26	14.75	13.31
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
NL-MVAE	11.34	13.25	12.54
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
NL-MVAE	9.33	12.05	10.12
Proposed	**12.58**	**18.74**	**11.76**

maintains its high SDR and SIR, demonstrating its robustness against reverberations.

D. Evaluation of the Impact of the Angle Between Sources and Distance Between Sources and Microphones on the Performance of the Proposed Method

In Section VI-C, we confirmed the effectiveness of the proposed method under the underdetermined conditions. Considering that the selection of targets depends on the spatial information, the observed spatial properties of audio signals always depend on the spatial distribution of a sound source, the sound scene acoustics, and the distance between the source and the microphones. In particular, one potential problem of the proposed method is its limited discriminative capability when any of the interference speakers shares a close position with the target speaker in space, even if they are far apart, referred to as the spatial overlap issue [68], [69]. Moreover, the distance between the source and microphones may have played some role in directional TSE. The farther the sources are from the microphones, the lower the sound pressure level will be, which may lead to a challenging situation. In this evaluation, we evaluated the impact of the angle between the desired target and the nearest interference, which can be considered as the spatial resolution of the GC-based TSE method.

1) Evaluation Condition: For evaluation, we simulated RIRs by ISM for the same room shown in Fig. 4 with a reverberation time of $RT_{60} = 150$ ms. Three speakers were randomly located in the range of 0°–180°. The test dataset was the same as those in Section VI-C. In the test space, all speakers were randomly located in different positions with the angle between the target and the nearest interference speaker of four ranges: 5°–15°, 15°–30°, 30°–50°, 50°–70°, and 70°–90°. All sources kept the same distance to the center point of the dual-microphone array of 0.5–4.0 m with a resolution of 0.5 m. The test space in the simulated room is shown in Fig. 5.

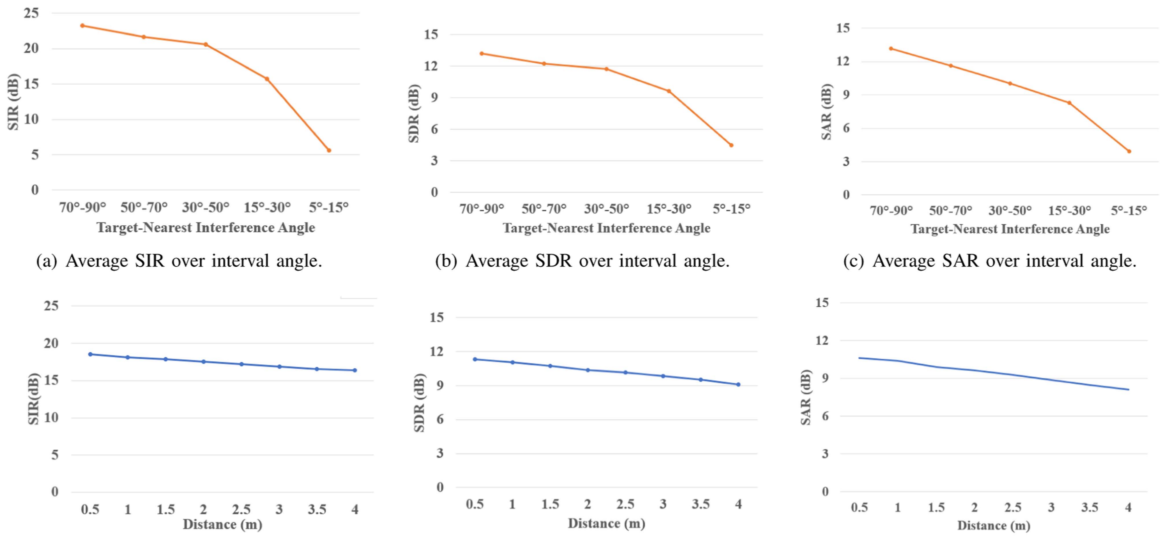
2) Evaluation Results: Fig. 8 shows a summary of the evaluation results of the performance of the proposed method at different interval angles between the target and the nearest interference and different distances between the source and the center of the microphone array. To analyze the impact of these two variables, the average performance over all angle ranges at each distance and the average performance at all distances in different angle ranges are summarized in Fig. 7(a), (b), and (c). As expected, the proposed method showed reduced performance when the source directions were closer. Particularly when there is interference within 30 degrees around the target, the performance will decline significantly. We also notice that when the angle is less than 15 degrees, such a decline trend becomes more significant.

As in the analysis of the interval angle, the average performance over all angle ranges at different distances is summarized in Fig. 7(d), (e) and (f). The result shows that the performance is stable even with increasing distance in the simulated room. This supports the benefit of the robustness of the proposed method in near and far fields.

E. Evaluation of DOA Modification

In this section, we evaluated the impact of DOA errors on our proposed method and the robustness of the proposed method against DOA errors. Note that the impact of DOA errors is closely related to the angle between sources. For example, when the interval angle between the target speaker and the nearest interference source is large, even if the estimated DOA has some errors, the impact on the result is relatively limited. Particularly when the error is within 0.5 times the interval angle, that is, the estimated target DOA was biased to the target side in the space between the target speaker and the interference source, the spatial filter calculated by geometric constraints will tend to extract the target signal from the mixed signal. Therefore, instead of using the absolute error for evaluation, we used the relative DOA error and compared it with the interval angle. For example, if the angle between the target and the interference is 60° and the estimated target DOA error is 12°, then the relative DOA error is considered 0.2. The test range in this evaluation was set from 0.1 to 1.0 relative DOA errors. All sources were randomly located at angles from 0 to 180 degrees in the simulated room shown in Fig. 4. The distance from all sources to the center of the dual-microphone array was 1 m. To prevent the impact of multiple interference sources, the relative DOA error in each experiment was biased towards the interference source nearest to the target. The evaluation was carried out in the simulated room shown in Fig. 5 with $RT_{60} = 150$ ms. Three sources were located randomly with an interval angle between the target and the nearest interference of 30°–90°, and the distance of each source from the center of the microphone array was 1 m.

Fig. 9 shows the SIR and SDR results with different relative DOA errors of the proposed method and the proposed method with DOA modification. The orange and blue lines represent the average performance of the two methods in different relative



(a) Average SIR over interval angle.

(b) Average SDR over interval angle.

(c) Average SAR over interval angle.

(d) Average SIR over distance between sources and center of microphone array.

(e) Average SDR over distance between sources and center of microphone array.

(f) Average SAR over distance between sources and center of microphone array.

Fig. 7. Average performance of the proposed method with different interval angles and distances between sources and center of the microphone array.

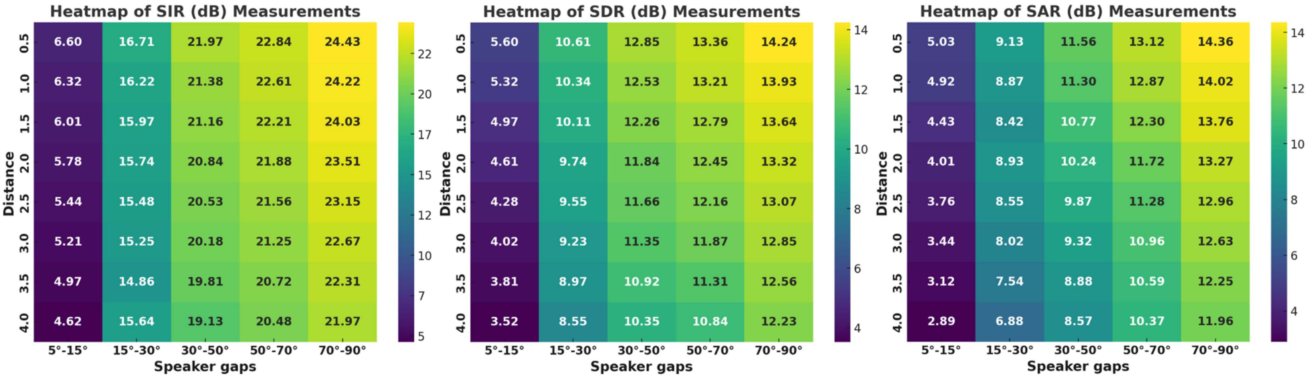
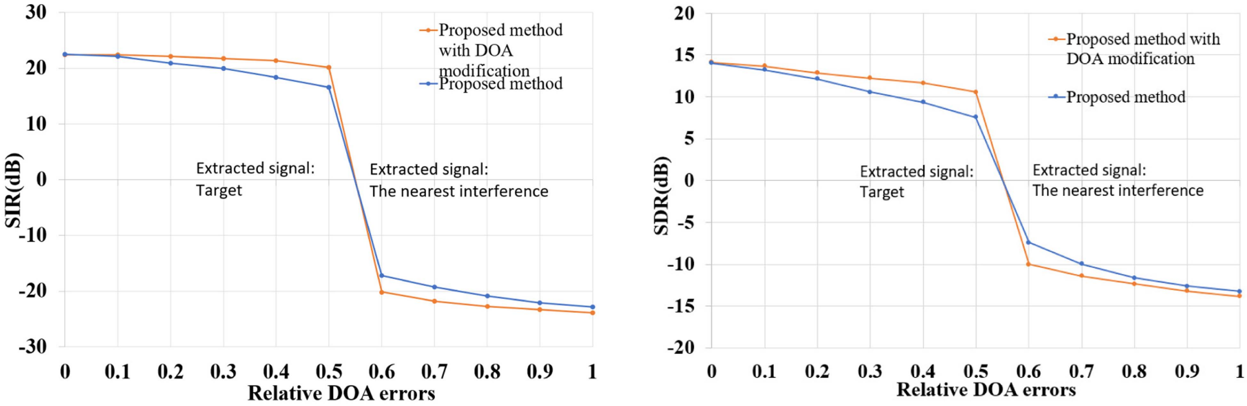


Fig. 8. Average SDR, SIR, and SAR of proposed method in 3-speaker case.



(a) The average SIR with different relative DOA errors.

(b) The average SDR with different relative DOA errors.

Fig. 9. Average SIR and SDR with different relative DOA errors of two proposed methods.

DOA error ranges. For example, a point with a horizontal axis of 0.3 represents the average result of DOA estimation error within 0.2–0.3. With increasing relative DOA error, the two proposed methods show different degrees of extraction performance. Unlike the proposed method, the proposed method with DOA modification is more stable as error increases and has a smaller performance reduction, which means that it is more robust to DOA errors than method 1. Additionally, we observe from Fig. 9 that the SIR and SDR of these two methods become negative when the relative error exceeds 0.5 times the interval angle. Since the DOA given to the system is skewed to the nearest interference side, the extracted signal changes from the target speaker to the nearest interference speaker.

VII. Conclusion

In this paper, we presented a dual-channel geometrically constrained TSE method for underdetermined conditions based on the CVAE source model. Our dual-channel algorithm designed based on the GSC structure can effectively utilize the spatial information of the target speaker. As the main novelty of this research, we first utilize CVAE to model the mixed speech signal, which overcomes the limitations of the source model in the traditional BSS algorithm under underdetermined conditions. As another contribution, we newly proposed the TSE algorithm with DOA modification to overcome the negative impact of DOA estimation errors.

Our experimental results demonstrated the following. (1) The proposed IntCVAE source models effectively represent mixed speech under the underdetermined conditions. (2) Compared with baselines, our proposed TSE method achieved significant improvement under the underdetermined conditions, even in the presence of strong reverberation. (3) Owing to our algorithm's dependence on spatial information, the performance is affected by the interval angle of sources. However, it is less affected by the distance between the source and the microphone array, and (4) the proposed method with DOA modification significantly reduced the negative impact of DOA estimation errors. It is worth noting that this paper only focuses on the processing of speech signals. Whether this method has generalization for non-speech signals such as background noise is an interesting research direction, and we will further investigate it in our future work.

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC, 2013.

[3] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.

[4] M. Elminshawi, W. Mack, S. Chakrabarty, and E. A. Habets, "New insights on target speaker extraction," 2022, *arXiv:2202.00733*.

[5] C. Cherry and J. A. Bowles, "Contribution to a study of the cocktail party problem," *J. Acoustical Soc. Amer.*, vol. 32, no. 7, pp. 884–884, 1960.

[6] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley-IEEE Press, 2006.

[7] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix FACtorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[8] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 11, 2003, Art. no. 569270.

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

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

[11] J. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[12] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1/4, pp. 1–24, 2001.

[13] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[14] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 666–678, Mar. 2006.

[15] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of binwise separated signals for permutation alignment in frequency-domain BSS," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2007, pp. 3247–3250.

[16] T. Kim, T. Eltoft, and T. W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation*, 2006, pp. 165–172.

[17] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation*, 2006, pp. 601–608.

[18] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. Int. Conf. Signal Process.*, 2014, pp. 473–477.

[19] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to singlechannel speech separation via high-resolution deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1424–1437, Aug. 2016.

[20] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5554–5558.

[21] Q. Wang et al., "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. INTERSPEECH*, 2019, pp. 2728–2732.

[22] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx: A complete time domain speaker extraction network," in *Proc. Interspeech*, 2020, pp. 1406–1410.

[23] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. Interspeech Conf.*, 2018, pp. 1170–1174.

[24] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.- M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.

[25] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech Conf.*, 2018, pp. 3244–3248.

[26] J. Wu et al., "Time domain audio visual speech separation," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2019, pp. 667–673.

[27] A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, Jul. 2018.

[28] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *Proc. Interspeech Conf.*, 2019, pp. 4295–4299.

[29] K. Tesch and T. Gerkmann, "Spatially selective deep nonlinear filters for speaker extraction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[30] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-SpEx: Localized target speaker extraction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7287–7291.

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

[32] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 715–726, Feb. 2007.

[33] W. Zhang and B. D. Rao, "Combining independent component analysis with geometric information and its application to speech processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 3065–3068.

[34] K. Reindl, S. Meier, H. Barfuss, and W. Kellermann, "Minimum mutual information-based linearly constrained broadband signal extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 6, pp. 1096–1108, Jun. 2014.

[35] H. Barfuss, K. Reindl, and W. Kellermann, "Informed spatial filtering based on constrained independent component analysis," in *Audio Source Separation*, Shoji Makino, ed. Berlin, Germany: Springer, 2018, pp. 237–278.

[36] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 846–850.

[37] 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, Jan. 1982.

[38] 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, Aug. 2001.

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

[40] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

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

[42] 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.

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

[44] N. Makishima et al., "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, Oct. 2019.

[45] 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.

[46] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3581–3589.

[47] R. Wang, L. Li, and T. Tomoki, "Direction-aware target speaker extraction with a dual-channel system based on conditional variational autoencoders under underdetermined conditions," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2022, pp. 347–354.

[48] S. Chakrabarty and E. A. P. Habets, "A Bayesian approach to informed spatial filtering with robustness against DOA estimation errors," *IEEE/ACM Trans. Audio Speech Lang. Process*, vol. 26, no. 1, pp. 145–160, Jan. 2018 .

[49] S. Sivasankaran, E. Vincent, and D. Fohr, "Analyzing the impact of speaker localization errors on speech separation for automatic speech recognition," in *Proc. 28th Eur. Signal Process. Conf.*, 2019, pp. 346–350.

[50] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 650–664, May 2009.

[51] J. Bourgeois and W. Minker, Eds. *Linearly Constrained Minimum Variance Beamforming*. Berlin, Germany: Springer, 2009, pp. 27–38.

[52] K. Buckley, "An adaptive generalized sidelobe canceller with derivative constraints," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 311–319, Mar. 1986.

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

[54] M. Pariente, A. Deleforge, and E. Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders," in *Proc. Interspeech*, 2019, pp. 3158–3162.

[55] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[56] 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. Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 716–720.

[57] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process.*, 2018, pp. 1–6.

[58] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder," in *Proc. ISCA Interspeech*, 2020, pp. 2437–2441.

[59] E. Karamatli, A. T. Cemgil, and S. Kırbız, "Audio source separation using variational autoencoders and weak class supervision," *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 1349–1353, Sep. 2019.

[60] 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. Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 746–750.

[61] J. Li and P. Stoica, *Robust Adaptive Beamforming*. New York, NY, USA: Wiley, 2005.

[62] K. Takao, M. Fujita, and T. Nishi, "An adaptive antenna array under directional constraint," *IEEE Trans. Antennas Propag.*, vol. AP-24, no. 5, pp. 662–669, Sep. 1976.

[63] S. Applebaum and D. Chapman, "Adaptive arrays with main beam constraints," *IEEE Trans. Antennas Propag.*, vol. AP- 24, no. 5, pp. 650–662, Sep. 1976.

[64] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

[65] B. Van Veen, "Minimum variance beamforming with soft response constraints," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 1964–1972, Sep. 1991.

[66] D. B. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 357–362.

[67] 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.

[68] R. Gu et al., "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. Interspeech*, 2019, pp. 4290–4294.

[69] L. Chen, M. Yu, D. Su, and D. Yu, "Multi-band pit and model integration for improved multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 705–709.