APSIPA Transactions on Signal and Information Processing, 2025, 14, e2

Original Paper

Target Speaker Extraction under Noisy Underdetermined Conditions Using Conditional Variational Autoencoder, Global Style Token, and Neural Postfilter

Rui Wang[1*], Takuya Fujimura[1] and Tomoki Toda[2]

[1] *Graduate School of Informatics, Nagoya University.*
[2] *Information Technology Center, Nagoya University.*

ABSTRACT

Target speaker extraction (TSE) acts as a front-end processing technology for various speech applications, such as automatic speech recognition. However, TSE has long faced challenges from underdetermined environments and in the presence of noise. In this paper, we propose a dual-channel system for directional TSE under noisy underdetermined conditions. In our approach, we utilize two source models that integrate conditional variational autoencoders (CVAEs) with global style tokens (GSTs) to learn representations of the noisy single speech and the noisy mixed speech within a geometric source separation framework, where GSTs generate conditional variables for CVAEs. To address residual noise in the extracted target signal under various noisy conditions, we introduce a conditional neural postfilter with a GST to estimate a complex Time-Frequency (T-F) mask for denoising. Additionally, we propose a joint network, where a conditional neural postfilter is

*Corresponding author: rui.wang@g.sp.m.is.nagoya-u.ac.jp. This work was partially supported by JST CREST JPMJCR19A3.s

Received 01 September 2024; revised 12 November 2024; accepted 05 January 2025
ISSN 2048-7703; DOI 10.1561/116.20240067

jointly trained with a CVAE and a shared GST module. The experimental results demonstrate that our proposed dual-channel TSE method achieves better performance under noisy underdetermined conditions.

Keywords: Target speaker extraction, multichannel source separation, conditional variational autoencoder (CVAE), speech enhancement

1 Introduction

In daily speech communication, numerous factors such as interfering speakers and background noise often hinder the clarity of the desired sound. Although the human brain can naturally focus on sounds of interest [25, 36], significant efforts have been dedicated to developing engineering solutions, leading to research on target speaker extraction (TSE). TSE is a specialized form of speech separation that aims to isolate and enhance the speech of a specific individual from a complex acoustic mixture. TSE has become a crucial front-end processing technique in speech signal processing applications, such as automatic speech recognition and speaker verification [8].

Unlike traditional blind speech separation (BSS), TSE only focuses on the desired target in the observed mixture. This technology is essential under complex multi speaker conditions where the clarity of a specific speaker's voice is critical, despite the presence of other interference speakers, background noise, or various sound sources. The TSE process often involves several stages, beginning with the identification of the target speaker, often using prior knowledge or auxiliary information of the target, followed by feature computation and separation processes. Among various types of auxiliary information, extracting speaker information from audio samples has proven effective. Frequency-domain methods such as SpeakerBeam [7] and VoiceFilter [28] isolate the target speaker using an adaptation utterance or a reference signal. Time-domain methods such as SpEx+ [14] adapt speaker encoders. Additionally, visual features like lip movements [12, 17, 2, 41] and facial frames [9, 1] are widely used to enhance speaker isolation.

In a spatial sound field, the spatial information of sound sources like the direction of arrival (DOA) is a clearly distinguishable feature of different sources. Utilizing spatial information is an effective way of target identification, which has shown the potential of TSE in combination with traditional BSS frameworks. Geometric source separation (GSS) [30, 23, 42, 33, 4] is an implementation that use geometric constraints (GCs) in BSS frameworks to separate a target source from the mixture. Classical methods like geometrically constrained independent vector analysis (GCIVA) [24] have been proposed. In

GCIVA, a linear GC based on the target's DOA is integrated into the IVA framework [20]. GCIVA employs a generalized sidelobe canceller (GSC) [15, 13] structure, utilizing a fixed beamformer to enhance the target signal and a null beamformer to suppress the target and estimate interferences. GSS methods do not require a large amount of training data and do not need prior spatial audio information during training. It only requires target spatial information to generate GCs to achieve target person selection in the inference stage.

Most classical methods such as GSS and spatial TSE methods, including GCIVA, are designed for determined and clean conditions, where the number of microphones is equal to that of sources, and the background diffuse noise is not considered. In contrast to these ideal conditions, real-world applications often face noisy underdetermined environments owing to hardware limitations and environmental noise. A major challenge under such conditions is the limitation of a source model. Traditional source models like the Laplace distribution in the IVA framework, are usually used for modeling a single source. Under underdetermined conditions, a more robust source model is essential, as it must handle not only the target speech but also a mixture of interfering speakers. Several efforts have been made to enhance source modeling. For independent Low-Rank Matrix Analysis (ILRMA), a flexible non-negative matrix factorization was introduced to improve the model's capacity [22]. Furthermore, in a recently proposed Bayesian-based GSS method, a background source model derived by independent vector extraction (IVE) was utilized to isolate the source of interest [5]. Recently, deep neural networks (DNNs) have been leveraged to model spectral features owing to their robust capabilities [29, 26]. Notably, the multichannel variational autoencoder (MVAE) [18] approach employs a conditional variational autoencoder (CVAE) [21] as the source model within an IVA framework, which was proved effective for determined cases. In contrast, in [37], a dual-channel TSE method was proposed for underdetermined cases, by introducing the target CVAE (TarCVAE) and interference CVAE (IntCVAE) within a GC-based framework. TarCVAE is designed to model a target speaker's speech and trained on clean, individual speech samples with one-hot labels representing the speaker's identity. IntCVAE, on the other hand, models mixed speech using the number of speakers as a condition, and is trained on clean mixed speech with one-hot labels indicating the number of speakers. The reliance on clean speech data and discrete labels limits their effectiveness in noisy environments.

To overcome this limitation, we have proposed a noise-robust source model for modeling noisy mixed speech signals by introducing global style tokens (GSTs) embedded within IntCVAE, which is called GIntCVAE [38]. GSTs is a set of embeddings that captures global acoustic features and can be trained in an unsupervised manner [39]. In the GIntCVAE framework, a GST functions as an embedding layer to capture the latent representation of noisy mixed

speech. The GST output serves as the conditional variable for the CVAE. During training, the GST and CVAE are jointly trained across varying numbers of speakers and noise environments, forming a robust source model for noisy interference mixtures.

Although our proposed GIntCVAE has already demonstrated its noise robustness under noisy underdetermined conditions, some issues remain. In our evaluations, it has been observed that under lower signal-to-noise ratio (SNR) conditions with background diffuse noise, the final extracted target still contains some residual noise. This is because, within the GSS framework, the target selection is based on the generated beamformer. When the number of microphones is limited, the diffusion noise from the target speaker in the same direction and within a certain range around the speaker will inevitably be retained with the generated beamformer. This is a common issue in research on directional speech extraction, separation, and enhancement [44, 16]. Generally, this problem can be solved by using a postfilter. In our previously proposed framework, we designed an ideal ratio mask (IRM) by utilizing the estimated interference mixture as the postfilter to enhance the final extracted target. Such a simple Time-Frequency (T-F) mask is however weak in dealing with diffuse noise. In addition, for our trained GIntCVAE, the types of information and representation have been learned by its GST module as an embedding layer during the training of the entire network are still unknown.

To overcome the limitation in our previous work, we proposed a noise-robust TSE method for noisy underdetermined conditions based on our former framework. A research has shown that the complex mask estimated by a neural postfilter significantly outperforms the traditional T–F mask [19]. Recently, a DNN-based speech enhancement method that employs a CNNBLSTM network to generate a complex T–F mask has been shown to be effective [11]. In this paper, we still focus on the dual-channel GSS framework using the minimum number of microphones to leverage spatial information, where we assume the target DOA is known as prior information. We introduce a neural postfilter to estimate a complex T–F mask as instead of the IRM-based T–F mask in the framework. To better model the initial extracted target signal with noise, we jointly train a conditional neural postfilter with a new TarCVAE source model called GTarCVAE, where a GST module is shared by the TarCVAE and neural postfilter. Note that this paper is an extended full journal version of the previous conference paper [38]. Our new contributions are as follows:

(1) We investigated how the number of speakers and SNR affect the discrimination capability of GST in mixed speech and what the GST module learns by visualizing the embedding space generated by the GST in GIntCVAE.

(2) We introduced a conditional neural postfilter to estimate a denoising complex T–F mask to reduce the residual diffuse noise in the extracted target signal. This postfilter is jointly trained with a new TarCVAE and has a shared GST module for the TarCVAE and postfilter.

Compared with our previous method [38], which incorporated GST only in the interference channel through GIntCVAE, this study extends GST to model both the interference and target channels, enhancing the representation of noisy target speech. Unlike the discrete one-hot vector used previously, GST generates a continuous latent representation, serving as a more effective conditional variable for CVAE under noisy underdetermined conditions. Additionally, the IRM-based postfilter previously used has limited performance in handling diffuse noise. The newly introduced CNNBLSTM-based neural postfilter leverages complex T-F masking, providing superior noise suppression and improving robustness in challenging environments. These advancements collectively strengthen target speaker extraction in noisy underdetermined conditions.

In addition, to provide a clear comparison between this work and our previous methods [37] and [38], we have summarized the key differences in Table 1.

Table 1: Comparison between previous methods and the proposed method.

Method	Source model for target	Source model for interference mixture	Postfilter
Previous method 1 [37]	TarCVAE	IntCVAE	IRM
Previous method 2 [38]	TarCVAE	GIntCVAE	IRM
Proposed	GTarCVAE	GIntCVAE	Neural postfilter

The rest of the paper is organized as follows. The problem formulations of the underdetermined GSS and CVAE are described in Sections 2. After that, we review the related work for noisy underdetermined conditions in Sections 3, in which Sectionss 3.2 and 3.3 provide a complete and detailed description of the proposed framework and source model for noisy underdetermined conditions. In Sections 4, we propose our conditional neural postfilter for complex T–F mask estimation and the joint network with TarCVAE. Experimental evaluations are presented in Sections 5. The conclusion is made in Sections 6.

2 Directional TSE Based on Dual-Channel System

2.1 Problem formulation

Let us consider a TSE problem using a dual-channel microphone array. The STFT coefficients of the source and observed signals are denoted as $\mathbf{s}(f,n)$ and $\mathbf{x}(f,n)$, where f and n are the frequency and time indices respectively. The vectors containing $s_1(f,n), s_2(f,n)$ and $x_1(f,n), x_2(f,n)$ at two channels can be respectively represented 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, including other interference speakers and additional noise. $x_1(f,n)$ and $x_2(f,n)$ are the observed signals of two input microphones, and $(\cdot)^T$ denotes transpose. We use a separation system based on the demixing matrix $\mathbf{W}(f)$ 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 $(\cdot)^{\mathrm{H}}$ denotes a Hermitian transpose. Here, $\boldsymbol{w}_1(f)$ is a linear filter for enhancing the target and $\boldsymbol{w}_2(f)$ is a linear filter for estimating the interference by suppressing the target. Under underdetermined conditions, using a linear filter to estimate the interference mixture by suppressing the target is feasible.

Next, we assume that source signals follow the local Gaussian model, i.e., $s_j(f,n)$ follows a zero-mean complex Gaussian distribution with the variance $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$, where $j = 1, 2$. The source signal $\mathbf{s}(f,n)$ then follows

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)). \tag{5}$$

Furthermore, we assume that $s_1(f,n)$ and $s_2(f,n)$ are independent of each other. $\mathbf{s}(f,n)$ then follows

$$\mathbf{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|\mathbf{0}, \mathbf{V}(f,n)), \tag{6}$$

where $\mathbf{V}(f,n) = \mathrm{diag}[v_1(f,n), v_2(f,n)]$. From Eqs. (3) and (6), we can show that $\mathbf{x}(f,n)$ follows

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, (\mathbf{W}^{\mathrm{H}}(f))^{-1}\mathbf{V}(f,n)\mathbf{W}(f)^{-1}). \tag{7}$$

The log-likelihood of the demixing matrices $\mathcal{W} = \{\mathbf{W}(f)\}_f$ and source model parameters $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$ for the observed mixture signals $\mathcal{X} = \{\mathbf{x}(f,n)\}_{f,n}$ is given by

$$\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) \stackrel{c}{=} 2N \sum_f \log|\det \mathbf{W}(f)| - \sum_{f,n,j} (\log v_j(f,n) + \frac{|\mathbf{W}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)|^2}{v_j(f,n)}), \tag{8}$$

where $\stackrel{c}{=}$ denotes equality up to constant terms. It means the equation holds except for an irrelevant constant, which does not affect the outcome of the optimization.

Note that Eq. (8) is a common objective function for BSS. To achieve target speaker selection, additional cues of the target are necessary. Here, we assume that the target DOA α is known. GCs [31] restrict the far-field response of the demixing filters in the target direction given by

$$\begin{aligned} J_{gc}(\mathcal{W}) = \lambda_1 \sum_f |\boldsymbol{w}_1^{\mathrm{H}}(f)\boldsymbol{d}(f,\alpha) - 1|^2 \\ + \lambda_2 \sum_f |\boldsymbol{w}_2^{\mathrm{H}}(f)\boldsymbol{d}(f,\alpha)|^2, \end{aligned} \tag{9}$$

$$\boldsymbol{d}(f,\alpha) = \exp[-j(\mathbf{p}/c)f\cos(\alpha)], \tag{10}$$

where λ_1 and λ_2 control GCs of two channels and $\boldsymbol{d}(f,\alpha)$ is a steering vector towards α. $\mathbf{p} = [p_1, p_2]$ are the positions of two microphones, and c is the wave propagation speed. The first term on the right-hand side of Eq. (9) uses delay-and-sum (DS) beamforming to preserve the target source [6]. In contrast, the second term creates a null beamformer towards α, acting as a blocking matrix (BM) [43] that suppresses the target while estimating a mixture of all interferences. The overall objective function to be minimized is

$$J(\mathcal{W},\mathcal{V}) \stackrel{c}{=} -\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) + J_{gc}(\mathcal{W}). \tag{11}$$

2.2 GSS framework with CVAE

When a dual-channel system addresses underdetermined conditions, the signal $s_2(f,n)$ in Eq. (2) represents a mixture of multiple interfering speakers. When applying GCs based on Eq. (9), the goal is for the estimated demixing matrix $W(f)$ to effectively separate the target $s_1(f,n)$ and the interference mixture $s_2(f,n)$ across the two channels. In [37], a dual-channel GSS framework that integrates GCs, CVAE's representation capabilities, and linear postprocessing has been proposed. Figure 1 shows the illustration of the framework of this dual-channel system. On channel 1, referred to as the target channel, a preliminary estimation of the target is obtained using the DS beamforming generated by calculating Eq. (9). On channel 2, the interference channel, a BM from Eq. (9), suppresses the target and preserves all other interferences. We can respectively estimate $s_1(f,n)$ and $s_2(f,n)$ from the output of two channels as

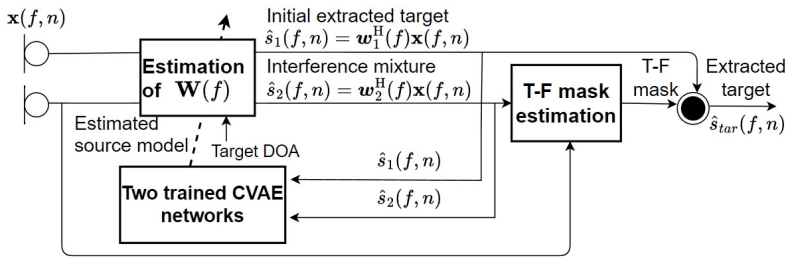

Figure 1: Proposed framework of directional target speaker extraction based on dual-channel system.

$$\hat{s}_1(f,n) = \boldsymbol{w}_1^{\mathrm{H}}(f)\mathbf{x}(f,n), \tag{12}$$

$$\hat{s}_2(f,n) = \boldsymbol{w}_2^{\mathrm{H}}(f)\mathbf{x}(f,n). \tag{13}$$

In the estimation of demixing matrix based on the objective function given by Eq. (11), the set of matrices $\mathcal{W}$ can be iteratively updated on the basis of the updated source model parameters $\mathcal{V}$. The source model is represented by a CVAE, which iteratively updates $\mathcal{V}$. After that, a T–F mask-based postprocessing is performed to obtain the final extracted target.

2.3 CVAE source model

For the TSE under underdetermined conditions, it is essential to accurately model both the target speaker's speech and the interference mixture. In [37], two CVAEs were proposed to model these two parts, which were called the target CVAE (TarCVAE) and the interference CVAE (IntCVAE). Let $\boldsymbol{S} = \{\mathbf{s}(f,n)\}_{f,n}$ be the spectrogram of a sound source. A CVAE consists of the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})$ and the decoder $p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})$, where $\boldsymbol{z}$ is a time sequence of latent feature vectors and $\boldsymbol{c}$ represents the conditional variable of $\boldsymbol{S}$. The encoder network generates a set of parameters for the conditional distribution $q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})$ of $\boldsymbol{z}$ 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 a set of labeled data samples $\{\boldsymbol{S},\boldsymbol{c}\}$. In the inference stage, only the decoder is used to produce the source model parameters of $\boldsymbol{S}$.

In the training stage, TarCVAE is trained with clean speech of different speakers, where $\boldsymbol{c}$ is a one-hot vector to represent the speaker index. IntCVAE is trained with mixed speech with $\boldsymbol{c}$ being a one-hot vector to represent the number of speakers in the mixture. In the inference stage, only the decoder is used to model the source spectrogram by estimating the latent space variable z and the conditional variable $\boldsymbol{c}$ as the source model parameters. The output of the decoder is then used in the estimation of the demixing matrix.

The goal of training these two CVAEs is for them to learn the parameters of the encoder and decoder networks so that the encoder distribution $q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})$ becomes consistent with the posterior $p_\theta(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c}) \propto p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})p(z)$. In the training of CVAEs, the objective function to be maximized is

$$\begin{aligned}\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},\boldsymbol{c})\sim p_D(\boldsymbol{S},\boldsymbol{c})}[\mathbb{E}_{z\sim q_\phi(z|\boldsymbol{S},\boldsymbol{c})}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})] \\ - \mathrm{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})||p(\boldsymbol{z})]],\end{aligned} \tag{14}$$

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. Here, $p_D(\boldsymbol{S},\boldsymbol{c})$ is approximated as the empirical distribution of sample $\boldsymbol{S},\boldsymbol{c}$. The output distribution of the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})$ and the prior distribution of $\boldsymbol{z}$ are given by the Gaussian distributions:

$$q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c}) = \mathcal{N}(\boldsymbol{z}|\mu_\phi(\boldsymbol{S},\boldsymbol{c}),\sigma^2_\phi(\boldsymbol{S},\boldsymbol{c})), \tag{15}$$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I}), \tag{16}$$

where $\mu_\phi(\boldsymbol{S},\boldsymbol{c})$ and $\sigma^2_\phi(\boldsymbol{S},\boldsymbol{c})$ are the encoder outputs denoting the mean vector and the variance vector of $\boldsymbol{z}$, respectively. The decoder output $p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g)$ is designed to be a complex Gaussian distribution:

$$p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g) = \prod_{f,n}\mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|0,v(f,n)), \tag{17}$$

$$v(f,n) = g\cdot\sigma^2_\theta(f,n;\boldsymbol{z},\boldsymbol{c}), \tag{18}$$

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

2.4 Postprocessing based on T–F mask

Under underdetermined conditions with multiple interferences, our GC-based method generate the DS beamformer that serves as the initial extraction of the target. On the other hand, the null constraint towards the target direction functions as a BM, allowing the extraction of the interference mixture excluding the desired target on the corresponding channel. However, underdetermined conditions often lead to the initial target extraction being disturbed by the presence of multiple interfering speakers. To enhance the final extraction result, a T–F mask was developed for postprocessing [37]. This T–F mask is an IRM, which calculates the ratio between the spectrogram energies of the interference and the observed mixtures. The extracted target $\hat{s}_{tar}(f,n)$ is

$$\hat{s}_{tar}(f,n) = \hat{s}_1(f,n)(1-\frac{|\hat{s}_2(f,n)|^2}{|\mathbf{x}(f,n)|^2}). \tag{19}$$

3 Related Work on TSE under Noisy Underdetermined Conditions

3.1 Overview

In realistic applications, environmental noise is a significant factor alongside interference speakers. The performance of a TSE system is easily degraded by such noise. Although the IntCVAE source model proposed in [37] has shown its effectiveness under underdetermined conditions, it still has limitations in modeling noisy mixed speech when environmental noise exists. The main reason comes from the one-hot vector-based labels in the training of IntCVAE. Under clean underdetermined conditions without environmental noise, the objective of IntCVAE training is for the IntCVAE to learn source models from mixed signals with varying numbers of speakers. The number of speakers in mixtures of clean multi speaker signals can be effectively represented by discrete one-hot vectors. However, when noise is present, its varying levels are continuous variables. In this case, it is more straightforward to use continuous representations to model mixed speech with different numbers of speakers and noise levels.

3.2 GST-IntCVAE network

To address this issue, we proposed a new source model called GST-IntCVAE (GIntCVAE for short) for modeling noisy interference mixture signals. GIntCVAE introduces GSTs [39] to generate embeddings of noisy mixed speech as conditional variables in the CVAE. A GST is a set of embeddings that captures global acoustical characteristics observed over an utterance, such as the expressiveness of speech and it is trained in an unsupervised manner. Figure 2 illustrates both TarCVAE and the proposed GIntCVAE. TarCVAE, as in [37], models a single target speaker's voice on channel 1, whereas GIntCVAE models the interference mixture with noise on channel 2. Different from the former IntCVAE, GIntCVAE incorporates a GST module to generate embeddings of noisy mixture speech. Therefore, in the training stage, there is no need to prepare labels for the training dataset. Since the GST can be trained in an unsupervised manner without additional labels, the GST and CVAE are jointly trained using only the training loss of the CVAE.

The GST module consists of a reference encoder and a noisy mixture token layer. The input audio is initially compressed into a fixed-length vector. This vector then serves as the query for the attention module in the noisy mixture token layer, which calculates a set of weights to measure its similarity to each token. The weighted sum of tokens serves as the noisy mixture embedding, which is incorporated into the encoder and decoder as the conditional variable c. This embedding captures the acoustic conditions of noisy interference mixture, such as the number of speakers and noise level.

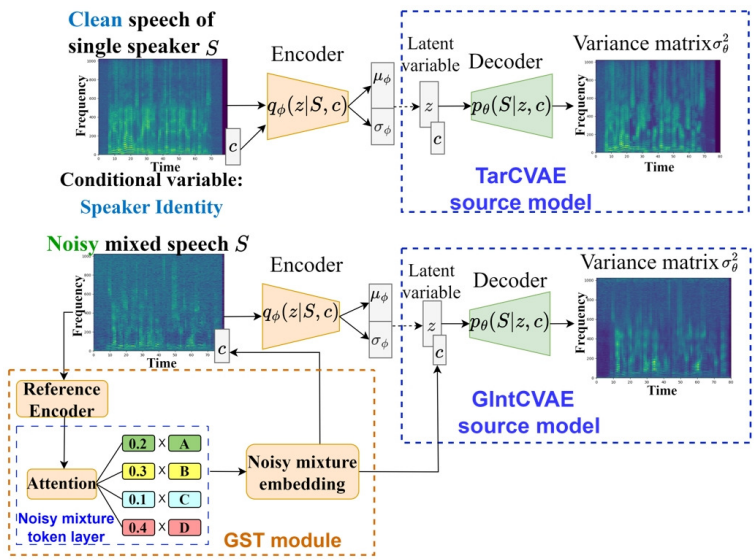

Figure 2: Illustration of TarCVAE and GIntCVAE.

In the inference stage, only the decoder is used to model the source and output distribution parameters, with GST weights being updated while fixing the noisy mixture tokens.

3.3 TSE algorithm

The crucial step in implementing TSE within the GSS framework outlined in Sections 2.2 is estimating the demixing matrix $\mathbf{W}(f)$. During the inference stage, the source model parameters on two channels are updated iteratively using the trained TarCVAE and GIntCVAE. This allows the demixing matrix to be refined iteratively through the objective function described in Eq. (8).

The update rule for $\mathbf{W}(f)$ is derived using the principles of vectorwise coordinate descent (VCD), known for its fast convergence and low computational cost, and the elimination of the need for a step-size parameter. Details of derivation can be referred to [27]. Assuming only a dual-channel case as in Sections 2.1, the derived update rules are summarized as

$$\boldsymbol{u}_j = \boldsymbol{D}_j^{-1}\mathbf{W}(f)^{-1}\boldsymbol{e}_j \qquad (j=1,2), \tag{20}$$

$$\hat{\boldsymbol{u}}_1 = \lambda_1 \boldsymbol{D}_1^{-1}\boldsymbol{d}, \tag{21}$$

$$h_j = \boldsymbol{u}_j^H \boldsymbol{D}_j \boldsymbol{u}_j \qquad (j=1,2), \tag{22}$$

$$\hat{h}_1 = \boldsymbol{u}_1^H \boldsymbol{D}_1 \hat{\boldsymbol{u}}_1, \tag{23}$$

$$\boldsymbol{w}_j(f) = \begin{cases} \frac{\hat{h}_1}{2h_1}[-1+\sqrt{1+\frac{4h_1}{|\hat{h}_1|^2}}]\boldsymbol{u}_1 + \hat{\boldsymbol{u}}_1 & (j=1), \\ \frac{1}{\sqrt{h_2}}\boldsymbol{u}_2 & (j=2), \end{cases} \tag{24}$$

where $\boldsymbol{D}_j = \mathbb{E}[\mathbf{x}(f,n)\mathbf{x}^{\mathrm{H}}(f,n)/v_j(f,n)] + \lambda_j \boldsymbol{d}\boldsymbol{d}^{\mathrm{H}}$ and $\boldsymbol{e}_j$ is the jth column of the identity matrix ($j = 1, 2$). The global-scale parameter $\mathcal{G} = \{g_j\}_j$ is updated as

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z},\boldsymbol{c})} \qquad (j=1,2). \tag{25}$$

where F and N refers to the number of frequency indices f and time indices n. In the training stage, θ, ϕ, and the parameters of GST are trained using Eq. (14). In the inference stage, the algorithm of demixing matrix estimation is summarized as follows:

1. Initialize $\mathcal{W}$ and $\Psi = \{\boldsymbol{z}, \boldsymbol{c}\}$.
2. Iterate the following steps for each j:

(a) Update $\boldsymbol{w}_j(f)$ by calculating Eqs. (20) to (24).
(b) Update $\boldsymbol{z}$ and $\boldsymbol{c}$ by backpropagation, where only GST weights are updated while fixing the noisy mixture tokens.
(c) Update g_j by calculating Eq. (25).
(d) Update v by calculating Eq. (18).

4 Proposed Method for Enhancing the Extracted Target

4.1 Overview

For a TSE problem under noisy underdetermined conditions, the observed signal can be considered a combination of three components: the desired target speaker, a mixture of interfering speakers, and environmental noise. The work in [37] focuses on clean underdetermined TSE without environmental noise, using TarCVAE and IntCVAE to model clean target speaker signals and clean interference mixtures, respectively. Building on this, GIntCVAE was proposed to model noisy interference mixtures on channel 2 while continuing to use TarCVAE as the source model for channel 1, aiming to extend the dual-channel directional TSE system to noisy underdetermined conditions.

In [37], it is assumed that the GC-based framework can divide the observed signal into two components: the clean target speaker on channel 1 and the noisy interference mixture on channel 2. The latter includes all interfering speakers and environmental noise. On the basis of this assumption, the proposed framework enables TarCVAE and GIntCVAE to model these two components separately. However, the extracted target speaker signal often contains residual noise. This issue arises because environmental noise is diffuse rather than a point source. Consequently, with a limited number of microphones, diffuse noise from sources in the same direction as the target speaker and nearby areas is inevitably retained by the beamformer on channel 1.

An effective way to reduce the noise component in the extracted signal is by applying the postfilter. In our previously proposed framework, an IRM-based T–F mask was used as the postfilter. Although effective under clean underdetermined conditions, this mask struggles with diffuse environmental noise. Additionally, a TarCVAE trained on the clean speech of different speakers is limited in modeling the noisy target speaker with residual noise on channel 1. In this research, we adopted a complex T–F mask estimation network as our neural postfilter, and we jointly trained a new GTarCVAE source model with the neural postfilter.

4.2 Neural postfilter for estimating complex T–F mask

Recent studies have shown that the complex T–F mask generated using a neural postfilter is more effective for speech enhancement than traditional T–F masks [19]. CNNBLSTM is a widely used architecture for complex T-F mask estimation in DNN-based speech enhancement tasks [10]. Recent studies, such as [11], have demonstrated its effectiveness in generating complex ideal ratio masks (cIRM) for speech enhancement [40]. Compared with the traditional IRM in our previous method, the cIRM can simultaneously enhance both the magnitude and phase responses of noisy speech. In the case of using a neural postfilter, the real and imaginary components of the cIRM are always jointly estimated by the trained DNN. In this paper, we adopt this CNNBLSTM-based T–F mask estimation network as our neural postfilter.

Here is a brief review of the neural postfilter based on our problem formulation in Sections 2.2. $\hat{s}_1(f,n)$ is the initial extracted target signal on channel 1, which contains residual interferences in channel 1. Then, the neural postfilter can be represented as

$$\hat{s}_{tar}(f,n) = \mathcal{M}(\hat{s}_1(f,n);\beta)\hat{s}_1(f,n), \tag{26}$$

where $\mathcal{M}$ is the network for estimating the complex T–F mask, and β is the set of its parameters. The objective of this neural postfilter is to estimate the

complex T–F mask from the input noisy speech signal for denoising, ensuring that the denoised signal closely resembles the original clean target signal.

4.3 Proposed joint network of TarCVAE and neural postfilter

On the other hand, for the source model on channel 1, we used to keep using TarCVAE to model a clean target speaker. In the absence of environmental noise, TarCVAE can effectively model channel 1. However, in the presence of environmental noise, the diffuse noise mixes with the single target signals, posing a challenge for TarCVAE, which is trained solely on clean speech data with one-hot labels representing only the identity of a clean speaker. Therefore, on the basis of our previously proposed GIntCVAE, a feasible approach is to introduce a GST into TarCVAE and train it using noisy speech data. This new source model for modeling a noisy single speaker is called GST-TarCVAE or GTarCVAE for short.

To model the noisy target speaker on channel 1, the new GTarCVAE source model should be trained using noisy speech signals. Similarly, the neural postfilter aims to estimate the complex T–F mask from noisy speech for denoising, using the training data of noisy speech and corresponding clean ground truth. Additionally, the GST in GTarCVAE learns the latent representation of noisy single speakers, which can be introduced into the neural postfilter to provide the conditional variable, potentially enhancing the noise robustness in various noise environments. Therefore, we propose a joint trained network of GTarCVAE and the neural postfilter with a shared GST module.

Figure 3 shows the illustration of the training process of the joint network. There are three main parts: GTarCVAE, the shared GST module, and the neural postfilter. The illustration of GTarCVAE is shown in Figure 4. Similar to GIntCVAE in Sections 3.2, the shared GST module outputs the token weight as the noisy single speaker embedding from the input of a noisy single speaker's speech, which serves as the conditional variable for both GTarCVAE and the neural postfilter. The network architecture of the CNNBLSTM-based neural postfilter is the same in [19], and we will describe it in detail latter.

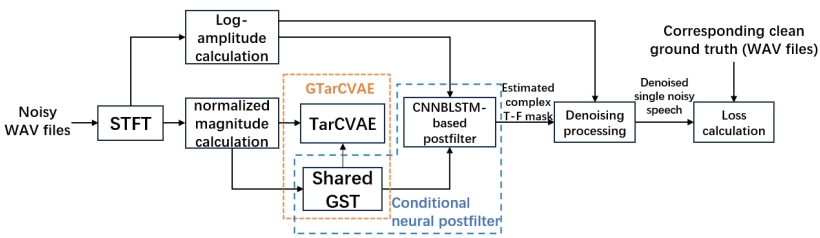

Figure 3: Illustration of joint training of GTarCVAE and postfilter.

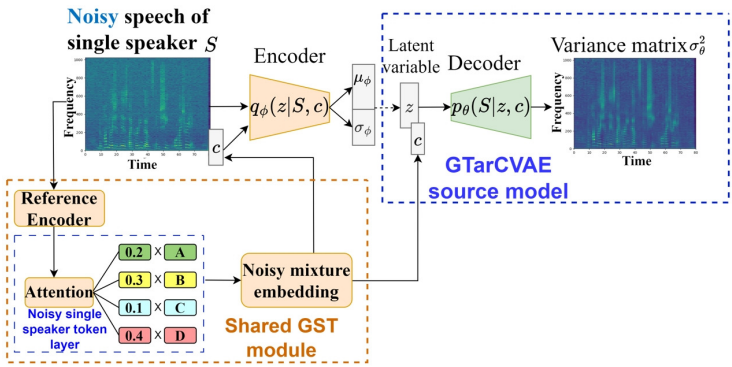

Figure 4: Illustration of GTarCVAE.

In the training of the original TarCVAE, the format of the training dataset is the normalized magnitude spectrogram, while in the training processing of the neural postfilter in [19], the format is the log-amplitude spectrogram. Therefore, we added two different calculation processes to obtain two types of training dataset from the same data source. In addition, for the neural postfilter, we adopted the clean-target training strategy, where the clean ground truth is required to calculate the loss in the training. Following the loss function Eq. (14) of the CVAE , we assume that $\boldsymbol{S}$ is the noisy training dataset and $\hat{\boldsymbol{S}}$ is the corresponding clean ground truth. The overall objective function of the training to be maximized is

$$\begin{aligned}\mathcal{J}(\phi,\theta,\beta) = \mathbb{E}_{(\boldsymbol{S},\boldsymbol{c})\sim p_D(\boldsymbol{S},\boldsymbol{c})}[\mathbb{E}_{z\sim q_\phi(z|\boldsymbol{S},\boldsymbol{c})}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})] \\ - \mathrm{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})||p(\boldsymbol{z})]] - \mathcal{D}[\mathcal{M}(\boldsymbol{S}|\beta)\boldsymbol{S},\hat{\boldsymbol{S}}],\end{aligned} \tag{27}$$

where $\mathcal{D}$ is the function that measures the difference between the denoised signal and the clean ground truth. Here, we followed [19] to set $\mathcal{D}$ as the mean-squared-error (MSE).

4.4 Inference processing of the new proposed method

In the inference stage, the decoder of the trained GTarCVAE and GIntCVAE serves as the source model on channels 1 and 2, where the source model parameters of the noisy target speaker and noisy interference mixed speakers, the weight sum of GSTs tokens, and the demixing matrix are updated as iteratively as the algorithm described in Sections 3.3. Then, the trained neural postfilter is used to denoise the initially extracted target speaker on channel 1 $\hat{s}_1(f,n)$. Different from the training stage, the conditional variable for the

postfilter is generated by the trained shared GST in the joint network with the input of the internal extracted target $\hat{s}_1(f, n)$.

5 Experimental Evaluation

5.1 Training setting of CVAEs and neural postfilter

In our experiments, we trained the joint network of GTarCVAE and the neural postfilter on 25 hrs of noisy audio data. The clean source data was obtained from the si_tr_s folder of the Wall Street Journal (WSJ0) corpus [32], which includes recordings from 101 speakers (50 male and 51 female), each contributing 141 sentences. We mixed the clean speech with four types of diffuse noise at varying SNR levels from the DEMAND dataset [34], which contains six types of diffuse noise. GIntCVAE was trained on 20 hrs of noisy mixed audio data, where the clean source for 19 groups of mixed speech with 2–20 speakers was generated by linearly mixing multiple speakers from the WSJ0 si_tr_s folder and then mixing it with the same noise sources as those used for the joint network of GTarCVAE and the neural postfilter.

The architectures of our networks are described as follows. The CVAE in both GTarCVAE and GIntCVAE has the same architecture as in [18], designed with an encoder, a latent space, and a decoder. The encoder consists of three convolutional layers: two 2D gated CNN layers followed by a regular 2D CNN layer. These layers incrementally encode the input spectrogram with a conditional variable, converting it into the latent space. The decoder mirrors the encoder with two 2D gated CNN layers and a final 2D convolutional transpose layer, enabling it to reconstruct the input spectrogram.

For the GST module, the architecture follows [39], containing a reference encoder and a token layer. The reference encoder processes the spectrogram input through six 2D CNN layers with progressively increasing channels, followed by a 128-unit GRU. The output of the reference encoder serves as the reference embedding, which is passed to the token layer to interact with the 10-token embedding bank via a multi-head attention module. The final output style embedding is a 128-dimensional vector.

The neural postfilter employs the CNNBLSTM architecture as described in [19]. The CNNBLSTM consists of an initial batch normalization and two 1D CNN layers, followed by a depthwise 2D CNN layer. A linear layer transforms the input dimension to a hidden dimension of $F \times N$. The BLSTM layers include 2 bidirectional layers. After that, a final linear layer maps the BLSTM output back to dimension of $2F \times N$. Finally, the output was divided into two $F \times N$ matrices, which constitute the real and imaginary parts of the cIRM.

The CVAE, GST, and CNNBLSTM were trained using the Adam optimizer, with learning rates of 0.0001 for the CVAE, GST and the CNNBLSTM.

GIntCVAE was trained 1000 epochs. GTarCVAE, CNN-BLSTM, and the joint network were all trained 800 epochs. All implementations are based on PyTorch 1.8.1, with hardward conditionas of a computer with Intel(R) Xeon(R) Gold 6248 CPU@ 2.50GHz, 32GB RAM, and one NVIDIA RTX 3090 GPU.

5.2 Investigation of embedding space of GIntCVAE

5.2.1 Evaluation conditions

In this evaluation, we investigated the embedding space of the trained GIntCVAE by visualizing the latent representations produced by its GST. We aim to analyse what the GST learnt regarding different aspects. Since GIntCVAE was trained on a dataset of noisy mixed multi speaker signals and models the noisy multi-interference mixture during inference, we want to determine the impact of the number of speakers and SNR on GST's capability to discriminate noisy mixed speech.

In the evaluation, we used different SNR conditions of noisy mixed signals from different numbers of speakers as inputs for the trained GIntCVAE. The dataset for this evaluation is constructed as follows. This dataset was generated by mixing different numbers of speakers and noise with different SNR conditions. There are seven categories of numbers of speakers, following a log scale: 2, 4, 8, 16, 32, 64, and 128. The SNR conditions are divided into eight categories: -20, -10, 0, 10, 20, 30, 40, and 50 dB. There are 50 samples for each number of speakers and SNR condition. Therefore, there are 8x7x50 samples in this evaluation. We used t-distributed stochastic neighbor embedding (t-SNE) [35] to compress all the 128-dimensional GST embedding outputs to 2D representations.

5.2.2 Evaluation results

We visualized the features compressed by t-SNE with the same input $8\times7\times50$ samples, using eight colors to represent different SNR conditions and seven colors to represent varying numbers of speakers. Figures 5 and 6 show the results. The GST embedding space shows clear clustering under different SNR conditions, indicating that the trained GIntCVAE effectively captures noise level information in noisy mixed speech. For varying the number of speakers, the clustering effect is less pronounced, although certain patterns can still be observed in each SNR cluster. This suggests that the trained GST finds SNR information in mixed speech easier to learn and distinguish than information on the number of speakers.

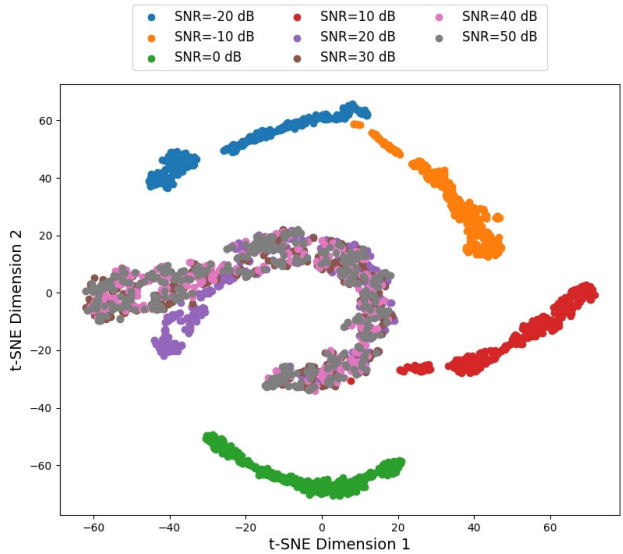

Figure 5: t-SNE visualization of GST by SNR conditions.

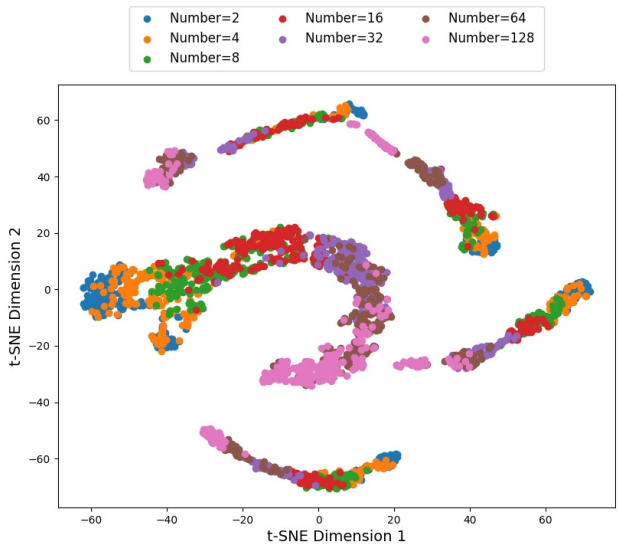

Figure 6: t-SNE visualization of GST by numbers of speakers.

5.3 Evaluation of TSE under noisy underdetermined conditions

5.3.1 Evaluation conditions

In this evaluation, we assessed our proposed method and baselines on three-source mixtures with a fixed reverberation time of 150 ms. Using the image source method (ISM) [3], we synthesized two-channel recordings of three speakers with noise in a simulated room, as shown in Figure 7. Speakers were randomly positioned at angles from 0° to 180°, with at least 10° between them, and each speaker was 1 m from the microphone array center. The ISM was chosen for its computational efficiency and ability to accurately simulate essential room acoustic characteristics, such as reflections and reverberation. This approach also provides acoustics control over variables like speaker and noise positioning, making it suitable for our controlled experimental setup under noisy underdetermined conditions. Speakers were randomly selected from the WSJ0 folders si_dt_05 and si_et_05, and noise data came from another two types of DEMAND noise, excluding training data. We tested under SNR conditions of -10 dB, 10 dB, and 30 dB, with 30 tests per condition. Each test utterance averaged 5–8 s. Performance was evaluated using the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), sources-to-artifacts ratio (SAR), and SNR, with higher values indicating higher performance.

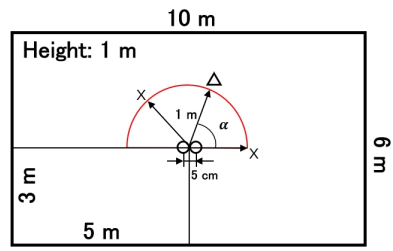

Figure 7: Configurations of evaluation, where △ and × denote the target and two interferences, and α is the DOA of the target.

We selected several related methods as our baselines, including GCIVA [24], MVAE [18], and IntCVAE [37]. For our ablation study on different components of the proposed source models incorporating GSTs, we trained a joint network of GST and TarCVAE without the neural postfilter. We evaluated two systems: TarCVAE + GIntCVAE and GTarCVAE + GIntCVAE, both using an IRM as the postfilter. To assess the new neural postfilter, we trained an independent CNNBLSTM-based neural postfilter without a GST and a joint network of GST and CNNBLSTM without TarCVAE, called the GST-Neural postfilter. Additionally, we evaluated GTarCVAE + GIntCVAE + Neural postfilter, GTarCVAE + GIntCVAE + GST-Neural postfilter, and the joint network of GTarCVAE-Neural postfilter + GIntCVAE. For GC-based methods, the

extracted target was the output from the corresponding channel. For the MVAE without target selection, we calculated the metrics of all separated signals with the ground truth of the target and chose the one with the best result as the extracted target. Categorization of each baseline and proposed method is summarized in Table 2.

Table 2: Comparison between baselines and proposed methods.

Method	Application scenario	Source model	Postfilter
GCIVA [24]	Determined	Laplace	N/A
MVAE [18]	Determined	TarCVAE	N/A
Previous method 1 [37]	Clean underdetermined	TarCVAE +IntCVAE	IRM
Previous method 2 [38]	Noisy underdetermined	TarCVAE +GIntCVAE	IRM
Proposed method 1	Noisy underdetermined	GTarCVAE +GIntCVAE	IRM
Proposed method 2	Noisy underdetermined	GTarCVAE +GIntCVAE	Neural postfilter
Proposed method 3	Noisy underdetermined	GTarCVAE +GIntCVAE	GST-Neural postfilter
Proposed method 4	Noisy underdetermined	Jointly trained GTarCVAE +GIntCVAE	Jointly trained GST-Neural postfilter

5.3.2 Evaluation results

Tables 3, 4, and 5 present a summary of the evaluation results obtained under different noisy conditions. The average SDR, SIR, SAR, and SNR indicate that our proposed methods consistently achieves improvements over baselines across various noisy conditions, particularly in terms of SIR and SDR, with statistical differences observed based on a paired one-sided t-test ($p < 0.05$). For example, at a low SNR of -10 dB, Previous method 2 achieves a 4.31 dB improvement in SIR over Previous method 1 ($p < 0.05$). Proposed method 4 achieves a further 1.57 dB improvement in SIR over Previous method 2 ($p < 0.05$). These results underscore the advantage of introducing GSTs into the source model for noisy interference mixtures, suggesting that, in noisy underdetermined conditions with a dual-channel system, refining the source model is important for TSE.

Furthermore, the introduction of the CNNBLSTM-based neural postfilter has also shown clear improvement in the TSE performance, with the cIRM estimated by the neural postfilter surpassing the traditional IRM. For instance, at a low SNR of -10 dB, Proposed method 4 achieves an average SDR of 3.28 dB, which is an improvement of 1.73 dB over Previous method 2 and 1.45 dB over Proposed method 1 ($p < 0.05$). The experimental results show that the cIRM estimated by neural postfilter suppresses residual noise better than

Table 3: Average SDR, SIR, SAR, and SNR [dB] of the extracted target under noisy underdetermined environment of SNR = -10 dB.

Method	SIR	SDR	SAR	SNR
GCIVA [24]	-3.68	-8.25	-6.39	-5.31
MVAE [18]	-3.12	-7.12	-2.44	-4.78
Previous method 1 [37]	1.03	-3.24	1.25	-2.23
Previous method 2 [38]	5.34	1.55	3.79	2.03
Proposed method 1	5.63	1.83	3.92	2.34
Proposed method 2	6.34	2.61	4.11	3.03
Proposed method 3	6.87	3.22	4.33	3.25
Proposed method 4	**6.91**	**3.28**	**4.36**	**3.32**

Table 4: Average SDR, SIR, SAR, and SNR [dB] of the extracted target under noisy underdetermined environment of SNR = 10 dB.

Method	SIR	SDR	SAR	SNR
GCIVA [24]	2.41	1.78	5.06	1.08
MVAE [18]	4.68	2.14	6.59	2.13
Previous method 1 [37]	8.76	4.43	6.15	4.37
Previous method 2 [38]	13.32	9.53	10.15	7.27
Proposed method 1	13.61	9.88	10.56	7.81
Proposed method 2	14.21	10.73	11.16	8.51
Proposed method 3	14.54	11.31	**11.55**	9.02
Proposed method 4	**14.62**	**11.40**	11.52	**9.13**

Table 5: Average SDR, SIR, SAR, and SNR [dB] of the extracted target under noisy underdetermined environment of SNR = 30 dB.

Method	SIR	SDR	SAR	SNR
GCIVA [24]	6.53	4.38	9.38	4.53
MVAE [18]	9.48	7.18	10.47	5.82
Previous method 1 [37]	15.11	9.61	11.19	8.89
Previous method 2 [38]	21.12	14.27	12.56	11.81
Proposed method 1	21.45	14.65	12.97	12.21
Proposed method 2	21.74	15.16	13.42	12.76
Proposed method 3	21.98	15.53	**13.54**	13.02
Proposed method 4	**22.03**	**15.61**	13.53	**13.10**

IRM, and the joint network approach is effective under noisy underdetermined conditions.

6 Conclusion

In this paper, we propose a dual-channel TSE method for noisy underdetermined conditions. Using a GSS framework, we integrate the GST module into the IntCVAE source model to develop GIntCVAE, a new model for noisy mixed

speech. To better model the target signal with noise, we also incorporate the GST module into the TarCVAE source model, creating GTarCVAE. Additionally, we analyze the embedding space of the GST in the trained GIntCVAE by visualizing the GST output embedding features to understand the information learned by the GST. We introduce a CNNBLSTM-based neural postfilter to address residual diffuse noise in the extracted target signal. We train a joint network of GTarCVAE and the neural postfilter with a shared GST module.

The experimental results highlight several points: (1) Introducing a GST into the CVAE source model enhances the GSS framework-based TSE method under noisy undetermined conditions. (2) The GST in GIntCVAE effectively learns the mixing conditions and spatial acoustic information of the interference mixture in noisy mixed speech, especially the noise level. (3) The CNNBLSTM-based neural postfilter more effectively enhances the extracted target signal with residual noise than the traditional IRM, and the joint network of GTarCVAE and the neural postfilter performs well under noisy undetermined conditions. Note that all current works are based on simulated mixed signals, and the proposed method is very time-consuming. In the future, we will further investigate its application to real recorded signals and develop an online algorithm.

References

[1] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions", *arXiv preprint arXiv:1907.04975*, 2019.

[2] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement", *arXiv preprint arXiv:1804. 04121*, 2018.

[3] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", *The Journal of the Acoustical Society of America*, 65(4), 1979, 943–50.

[4] H. Barfuss, K. Reindl, and W. Kellermann, "Informed Spatial Filtering Based on Constrained Independent Component Analysis", *Audio Source Separation*, 2018, 237–78.

[5] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis", *IEEE Transactions on Signal Processing*, 68, 2020, 3545–58.

[6] K. Buckley and L. Griffiths, "An adaptive generalized sidelobe canceller with derivative constraints", *IEEE Transactions on Antennas and Propagation*, 34(3), 1986, 311–9.

[7] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 5554–8.

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

[9] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation", *arXiv preprint arXiv:1804.03619*, 2018.

[10] H. Erdogan and T. Yoshioka, "Investigations on Data Augmentation and Loss Functions for Deep Learning Based Speech Background Separation", *Proc. of Interspeech*, 2018, 3499–503.

[11] T. Fujimura and T. Toda, "Analysis of Noisy-target Training for DNN-based speech enhancement", in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, 1–5.

[12] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement", *arXiv preprint arXiv:1711.08789*, 2017.

[13] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech", *IEEE Transactions on Signal Processing*, 49(8), 2001, 1614–26.

[14] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network", *arXiv prepr-int arXiv:2005.04686*, 2020.

[15] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming", *IEEE Transactions on Antennas and Propagation*, 30(1), 1982, 27–34.

[16] R. Gu and Y. Zou, "Temporal-spatial neural filter: Direction informed end-to-end multi-channel target speech separation", *arXiv preprint arXiv:2001.00391*, 2020.

[17] 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 Transactions on Emerging Topics in Computational Intelligence*, 2(2), 2018, 117–28.

[18] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder", *Neural computation*, 31(9), 2019, 1891–914.

[19] M. Kawanaka, Y. Koizumi, R. Miyazaki, and K. Yatabe, "Stable training of DNN for speech enhancement based on perceptually-motivated black-box cost function", in *ICASSP 2020-2020 IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, 7524–8.

[20] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components", in *International conference on independent component analysis and signal separation*, Springer, 2006, 165–72.

[21] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models", *Advances in neural information processing systems*, 27, 2014.

[22] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9), 2016, 1626–41.

[23] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis", *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2), 2007, 715–26.

[24] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement", in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, 846–50.

[25] P. C. Loizou, *Speech enhancement: theory and practice*, CRC Press, 2013.

[26] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10), 2019, 1601–15.

[27] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 746–50.

[28] H. R. Muckenhirn, I. L. Moreno, J. Hershey, K. Wilson, P. Sridhar, Q. Wang, R. A. Saurous, R. Weiss, Y. Jia, and Z. Wu, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking", in *Conference of the International Speech Communication Association*, 2019.

[29] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9), 2016, 1652–64.

[30] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming", *IEEE Transactions on Speech and Audio Processing*, 10(6), 2002, 352–62.

[31] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming", *IEEE Transactions on Speech and Audio Processing*, 10(6), 2002, 352–62.

[32] D. B. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus", in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[33] 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*, 22(6), 2014, 1096–108.

[34] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings", *Proceedings of Meetings on Acoustics*, 19(1), May 2013, 035081.

[35] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.", *Journal of Machine Learning Research*, 9(11), 2008.

[36] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview", *IEEE/ACM Transactions on Audio, Apeech, and Language Processing*, 26(10), 2018, 1702–26.

[37] R. Wang, L. Li, and T. Toda, "Direction-aware target speaker extraction with a dual-channel system based on conditional variational autoencoders under underdetermined conditions", in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, 347–54.

[38] R. Wang and T. Toda, "Directional Target Speaker Extraction under Noisy Underdetermined Conditions through Conditional Variational Autoencoder with Global Style Tokens", in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2023, 1–5.

[39] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis", in *International Conference on Machine Learning*, PMLR, 2018, 5180–9.

[40] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3), 2015, 483–92.

[41] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation", in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, 667–73.

[42] 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*, IEEE, 2009, 3065–8.

[43] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ICA-based blocking matrix for improved noise estimation", *EURASIP Journal on Advances in Signal Processing*, 2014, 2014, 1–24.

[44] Y. Zou, Z. Liu, and C. H. Ritz, "Enhancing target speech based on nonlinear soft masking using a single acoustic vector sensor", *Applied Sciences*, 8(9), 2018, 1436.