Doctoral Dissertation

Deep Source Modeling for Direction-aware Dual-channel Target Speaker Extraction in Noisy Underdetermined Conditions

Wang Rui

Supervisor: Tomoki Toda

Graduate School of Informatics, Nagoya University

January, 2025

Abstract

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. In social communication, this ability plays an important role in dealing with various voices and extracting useful information from them. However, machines have yet to achieve such an auditory attention capability as humans.

Many efforts have been spent on its engineering solution, which yields the research on target speaker extraction (TSE). With the development of speech processing technology, TSE has become an attractive research topic in recent years. TSE aims to extract a target speaker's voice from mixed signals, which is desired for numerous applications like speech recognition systems and hearing aids. TSE can be implemented by blind source separation (BSS) methods for a multi-channel system using a microphone array, where the BSS methods aim to separate all sources in the mixed signal without using prior information. However, in many realistic conditions, recovering all mixed sources is often unnecessary, as only one or a few specific sources are needed. Besides, to achieve the target selection, prior knowledge or additional information of the target speaker is necessary. One of the most effective pieces of information is a spatial cue of the target speaker in the spatial sound field. One effective implementation of utilizing spatial cues is the geometric source separation (GSS), which uses geometric constraints (GCs) in BSS frameworks to separate the target speaker.

Conventional GSS methods are usually designed for determined conditions, where the number of sound sources equals the number of microphones. However, in many realistic applications, hardware limitations often lead to underdetermined conditions, where the number of sound sources is larger than the number of available microphones. Although the GSS method can use GCs based on the spatial information of the target signal to achieve target selection, its essence relies on the BSS framework that achieves separation by maximizing the statistical independence among signals. At the same time, the traditional source models used in this process are mostly designed for dealing with a single source in clean determined conditions. In underdetermined conditions, a powerful source model is necessary for modeling the mixture of multi-speakers. Moreover, when diffuse noise is present, the source model needs to further represent the noisy signal to handle noisy underdetermined conditions.

This dissertation aims to propose a direction-aware TSE method in noisy underdetermined conditions. The main topic of the research is developing a source model for a dual-channel TSE to deal with such conditions. To achieve the research motivation, this study was conducted to solve the noisy underdetermined problem step by step. The first step is to achieve the dual-channel TSE in underdetermined conditions without background noise. On this basis, the second step further investigates the impact of diffusion noise on the proposed TSE method in underdetermined conditions and proposes solutions for noisy underdetermined environments.

In step 1, this study focuses on the TSE in underdetermined conditions. To achieve the research goal, a dual-channel framework with the combined

capabilities of target selection based on GCs, a more powerful deep source model, and nonlinear postprocessing is proposed. A linear 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, a time-frequency (T-F) mask estimated from the separated interference mixture speech is used to extract the target speaker's speech. Additionally, to mitigate the effect of DOA estimation errors, an improved method based on enhancing the objective function is proposed to allow the modification of the target DOA information. The experimental results demonstrate the effectiveness of the proposed method. All the processes are described in Chapter 3.

In step 2, considering the presence of noise in real-world environments, this study extends the proposed TSE method to address noisy underdetermined conditions. To improve the limitations of the source model, a new source model incorporating global style tokens (GST) within a CVAE is introduced to handle noisy multi-speaker mixed speech in Chapter 4. The GST is jointly trained with the CVAE as an embedding layer to learn latent representations, which serve as the conditional variable for the CVAE. While this new model demonstrates improved performance in noisy underdetermined conditions, residual noise remains in the extracted target signal in Chapter 5. To address this, Chapter 5 introduces a conditional neural postfilter with GST to estimate a complex T-F mask for denoising. Furthermore, a joint network is developed, where the conditional neural postfilter is trained alongside the CVAE, sharing the GST module. Experimental results demonstrate that the proposed source models and neural postfilter effectively improve

performance in noisy underdetermined conditions.

Contents

List of Figures

List of Tables

List of Abbreviations

TSE Target speaker extraction

BSS Blind source separation

ICA Independent component analysis

FDICA Frequency-domain independent component analysis

IVA Independent component analysis

AuxIVA Auxiliary function-based independent vector analysis

STFT Short-time Fourier transform

DOA Direction of arrival

GSS Geometric source separation

GC Geometric constraint

GCIVA Geometrically constrained independent vector analysis

GSC Generalized sidelobe canceller

ILRMA Independent low-rank matrix analysis

NMF Nonnegative matrix factorization

IVE Independent vector extraction

SOI Source of interest

DNN Deep neural network

MVAE Multichannel variational autoencoder

CVAE Conditional variational autoencoder

IRM Ideal ratio mask

cIRM Complex ideal ratio mask

LCMV Linearly constrained minimum variance

DS Delay-and-sum

BM Blocking matrix

MAP Maximum a posteriori

SIR Signal-to-interference ratio

SDR Signal-to-distortion ratio

SAR Signal-to-artifacts ratio

SNR Signal-to-noise ratio

TarCVAE Target CVAE

IntCVAE Interference CVAE

GD Gradient descent

BP Backpropagation

RIR Room impulse responses

ISM Image source method

GST Global style token

GTarCVAE GST-TarCVAE

GIntCVAE GST-IntCVAE

SLT Style token layer

CNNBLSTM Convolutional neural network-bidirectional long short-term memory

MSE Mean-squared-error

List of Symbols

$\mathbf{s}(f, n)$ The source signals at time-frequency domain

$\mathbf{x}(f, n)$ The set of microphone signals at time-frequency domain

f, n The frequency and time indices

$\mathbf{W}(f)$ Demixing matrix

$\hat{s}_{tar}(f, n)$ The extracted target

$\mathbf{V}(f, n)$ Variance matrix of source distribution

α The target DOA

α_0 The estimated target DOA

$\boldsymbol{d}(f, \alpha)$ Steering vector toward target DOA α

$\mathcal{Q}$ Auxiliary variable of AuxIVA

$Q_j(f)$ Weighted covariances of source model in AuxIVA method

$\boldsymbol{S}$ Input of the encoder

$\boldsymbol{z}$ Latent space variable of the CVAE

$\boldsymbol{c}$ Conditional variable of the CVAE

ϕ Parameters of the encoder network

θ Parameters of the decoder network

$q_\phi(\boldsymbol{z}|\boldsymbol{S}, \boldsymbol{c})$ Conditional distribution of $\boldsymbol{z}$ given $\boldsymbol{S}$ in the encoder

$p_\theta(\boldsymbol{S}|\boldsymbol{z}, \boldsymbol{c})$ Conditional distribution of $\boldsymbol{S}$ given $\boldsymbol{z}$ in the decoder

$\mu_\phi(\boldsymbol{S}, \boldsymbol{c})$ The mean vector of decoder distribution

$\sigma^2_\phi(\boldsymbol{S}, \boldsymbol{c})$ The variance vector of decoder distribution

$\sigma^2_\theta(\boldsymbol{z}, \boldsymbol{c})$ The output of decoder

g Global-scale parameter

β The set of parameters of neural postfilter

Chapter 1

Introduction

1.1 Research background

In human daily life, speech is the most effective form of communication, serving as a key medium that carries rich information and enables us to perceive the world. However, in everyday speech communication, various factors such as interfering speakers and background noise often hinder the clarity of the desired sound. Enhancing a target speaker from a mixture of signals is critical for numerous applications, including meeting recognition systems and smart home devices. For example, during a conference, many individuals may speak simultaneously, causing the desired speaker's voice to overlap with other sounds. Although the human brain has the remarkable ability to focus on a specific sound—such as in a cocktail party with multiple speakers —by filtering out other interferences through selective auditory attention, machines have yet to reach the same level of capability in isolating and enhancing the target speaker in such complex auditory environments [1] [2].

Many efforts have been spent on the engineering solution of selective auditory attention, which yields research on target speaker extraction (TSE). TSE aims to extract a particular speaker from a mixture of audio signals. It has become a desired front-end processing in the research of speech signal processing like automatic speech recognition (ASR) and has many practical applications such as speech enhancement, speech recognition and hearing aids [3].

An important technique related to TSE is blind source separation (BSS) methods [4–9], which forms the theoretical foundation of the TSE problem. BSS aims to recover all original sources from a mixed signal, with TSE being a specialized task within this domain. BSS approaches the separation process as the inverse of the source mixing process, seeking to estimate a demixing matrix that can separate all sources. Over the past several decades, numerous BSS methods have been developed. The first major class of BSS methods is independent component analysis (ICA), which is based on linear mixing and demixing processes and the assumption of independence in the source model [10]. ICA has been extensively studied in statistics and information theory. Frequency-domain ICA (FDICA) offers faster convergence compared to time-domain deconvolution methods [11–14], but it faces a notable challenge known as the permutation problem, which refers to inconsistencies in output channels across frequency bins. Independent vector analysis (IVA), a multivariate extension of ICA, addresses convoluted BSS problems by using a multivariate source prior to short-time Fourier transform (STFT) components, thus mitigating the permutation issue [15], [16].

In practical applications, recovering all mixed sources is often unneces-

sary. Unlike traditional BSS methods, TSE focuses solely on extracting the target speaker from the observed mixture. This is especially critical in complex multi-speaker environments where the clarity of a specific speaker's voice is essential, despite interference from other speakers, background noise, or various sound sources. The TSE process typically involves multiple stages, starting with the identification of the target speaker, often using prior knowledge or auxiliary information, followed by feature extraction and separation. Among the various types of auxiliary information, extracting speaker information from audio samples has proven to be effective. Frequency-domain approaches like SpeakerBeam [17] and VoiceFilter [18] isolate the target speaker using an adaptation utterance or reference signal. Meanwhile, time-domain methods such as SpEx+ [19] have gained attention by incorporating a speaker encoder. In addition, visual features such as lip movements [20] [21] [22] [23] and facial frames [24] [25] are also widely used to enhance speaker isolation.

In a spatial sound field with a multi-channel system, spatial information, such as the direction of arrival (DOA) of sound sources, serves as a distinctive feature that helps differentiate between sources. Several studies have demonstrated the potential of incorporating spatial information into traditional BSS frameworks. Geometric source separation (GSS) [26] [27] [28] [29] [30] is one such approach that utilizes geometric constraints (GCs) within BSS frameworks to separate a target source from a mixture. Classical methods like geometrically constrained independent vector analysis (GCIVA) [31]. GCIVA employs a generalized sidelobe canceller (GSC) [32] [33] structure, utilizing a 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 the spatial information to generate GCs to achieve the target speaker selection in the inference stage.

1.2 Research issues

Although the GSS method is an effective way to implement TSE by using spatial information, most existing GSS methods are developed for clean and determined conditions, where the number of microphones equals the number of sources and the effects of diffuse noise are not taken into account. Practical scenarios often involve noisy underdetermined conditions due to hardware limitations and environmental noise, posing challenges for traditional GSS methods. 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 clean single source. In cases of 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. Furthermore, when background noise is present, it becomes crucial to model the more complex signal, which includes both speech and noise components.

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 [34]. Furthermore, a Bayesian framework-based method has been proposed to introduce a background source (BG) model derived by independent vector extraction (IVE) [35] that allows for underdetermined cases to extract the source of interest (SOI) [36]. In recent years, deep neural networks (DNNs) have been leveraged to model spectral features owing to their robust capabilities [37] [38]. The multichannel variational autoencoder (MVAE) [39] method utilizes the conditional variational autoencoder (CVAE) [40] as the generative source model in an IVA framework, which was proposed for determined conditions. The 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 its effectiveness in determined conditions owing to its representation power. However, the CVAE used in MVAE trained by clean speech can only deal with determined conditions. Furthermore, without additional information on the target speaker, MVAE can not achieve the target selection.

Another key issue in noisy underdetermined conditions is diffuse noise. Within the GSS framework, target selection is based on the generated GC. When the number of microphones is limited, diffuse noise originating from the target speaker's direction, or from within a certain vicinity, inevitably remains when using the generated GC. This is a common problem in research on directional speech extraction, separation, and enhancement [42] [43]. Similar to underdetermined conditions, the challenge of diffuse noise in traditional GSS methods stems from the limitations of the source model, as these methods typically assume a clean, single-source model. In noisy underdetermined

conditions, environmental noise results in diffuse noise mixing with the target and interference signals, posing significant challenges for traditional GSS methods.

For the issue of diffuse noise, another limitation lies in the postfiltering process. Typically, diffuse noise can be addressed using a postfilter; however, traditional postfilters, such as real-value time-frequency (T-F) masks, are not very effective against diffuse noise. Research has demonstrated that a complex mask generated by a neural postfilter significantly outperforms traditional T-F masks [44]. More recently, a DNN-based speech enhancement method employing a convolutional neural network-bidirectional long short-term memory (CNNBLSTM) network has been shown to effectively generate a complex T-F mask [45]. In [45], a trained CNNBLSTM network is used to generate a complex ideal ratio mask (cIRM) [46] for speech enhancement, which enhances both magnitude and phase responses of noisy speech simultaneously, offering better performance over traditional ideal ratio mask (IRM) approaches.

1.3 Research objectives

This study aims to propose a dual-channel TSE method for noisy underdetermined conditions. Deep source models based on CVAEs have demonstrated strong capabilities in modeling speech signals and show significant potential for overcoming the limitations of GSS methods in noisy underdetermined conditions. Therefore, the primary focus of this study is to develop a deep source model capable of handling complex noisy mixed speech under

these conditions. Leveraging deep source modeling and GCs within the GSS framework, this study seeks to achieve directional TSE in noisy underdetermined conditions.

This study follows a two-step approach to achieve the research goal. In the first step, this study focuses on underdetermined conditions. A dual-channel TSE method is proposed, integrating the GCs-based TSE algorithm, deep source model, and T-F mask-based postfilter within a GSS framework. The limitation of the traditional source model in underdetermined conditions is addressed by developing a new deep source model based on CVAEs tailored to these conditions. Following this, an improved algorithm is proposed for addressing the impact of DOA estimation errors. In the second step, the study is extended to noisy underdetermined conditions, introducing refined deep source models and a new neural postfilter to further enhance performance in these challenging environments.

There are several novelties in this study. First, this study applies DNN to learn the source model of complex signals in noisy underdetermined conditions instead of using an assumed statistical source model. Second, the combination of the deep source model and GSS framework achieves the directional TSE in challenging environments. These novelties provide potential applications for dual-channel systems in speech processing in real-world sound fields. This dissertation makes the following key contributions:

1. Advanced Source Modeling with CVAE:

·A novel IntCVAE-based source model is proposed, which is incorporated with the GSS framework to enhance the capabilities of directional TSE

methods, enabling TSE in underdetermined conditions.

·The GIntCVAE model incorporates global style tokens (GST) to provide continuous conditioning representations, improving robustness in noisy environments.

2. DOA Modification for Robust Spatial Filtering:

·A novel DOA modification strategy is proposed to address inaccuracies in DOA estimation, improving the system's robustness to DOA estimation errors under real-world conditions with imperfect spatial information.

3. Integrated Postfiltering:

·The GTarCVAE model and a CNNBLSTM-based neural postfilter are jointly trained to suppress residual noise. This joint network leverages advanced source modeling and postfiltering techniques, resulting in higher-quality extracted speech.

4. Applications and Broader Impact:

·The proposed methods have potential applications in the front-end processing of speech recognition systems, assistive hearing devices, and telecommunication tools, enhancing their functionality in noisy, multi-speaker environments.

·The integration of spatial cues with neural source models may provide insights applicable to other fields, such as visual-audio signal processing and environmental monitoring.

1.4 Structure of this dissertation

In this dissertation, the structure of the content is shown in Fig. 1.1. At great length, Chapter 1 describes the background and research issues TSE in noisy underdetermined conditions. Also, the objectives, originality, and significance of this research are presented here.

In Chapter 2, there is a literature review of the target speaker extraction. First, the problem formulation is addressed at the beginning of this chapter. Then, the review of traditional methods is elaborated. They both included descriptions of GSS framework-based method and its limitations. Finally, the recently proposed DNN-based method, CVAE, is described in detail.

After that, in Chapter 3, the research of TSE in underdetermined conditions is described in detail here. This chapter presents a GSS-based framework, comprising GCs based on the target DOA, a deep source model, and a T-F mask-based postfilter. Within this framework, a directional TSE algorithm is proposed, and a novel deep source model for mixed speech signals is proposed and described in detail here. To evaluate the proposed source model, experiments of modeling power are carried out. Then, the issues made by DOA estimation errors are elaborated. To address this problem, an improved method with a DOA modification process is proposed. Finally, several experimental evaluations are made.

To extend this method to noisy underdetermined conditions, Chapter 4 introduces a refined deep source model by incorporating a deep embedding layer. The previous deep source model's conditional variable is found to have limitations in representing complex noisy mixed signals. The newly

introduced embedding layer learns latent representations of these signals, improving the model's ability to handle such complex signals. After that, several experimental evaluations are performed.

To address the residual diffuse noise in the extracted signal, Chapter 5 introduces a neural postfilter for enhancing the signal quality. The postfilter using the T-F mask designed in a manner based on signal processing has limited performance in handling diffuse noise. In this chapter, a DNN-based neural postfilter is proposed to estimate a complex T-F mask for denoising. Furthermore, to better model the initially extracted target signal with noise, a conditional neural postfilter is jointly trained with a new source model tailored for noisy target speech. Experimental evaluations are conducted to assess the effectiveness of these approaches.

Finally, Chapter 6 is a summary of this dissertation with respect to the research question, the proposed method, and the results of this research. In addition, the future works of this study are discussed.

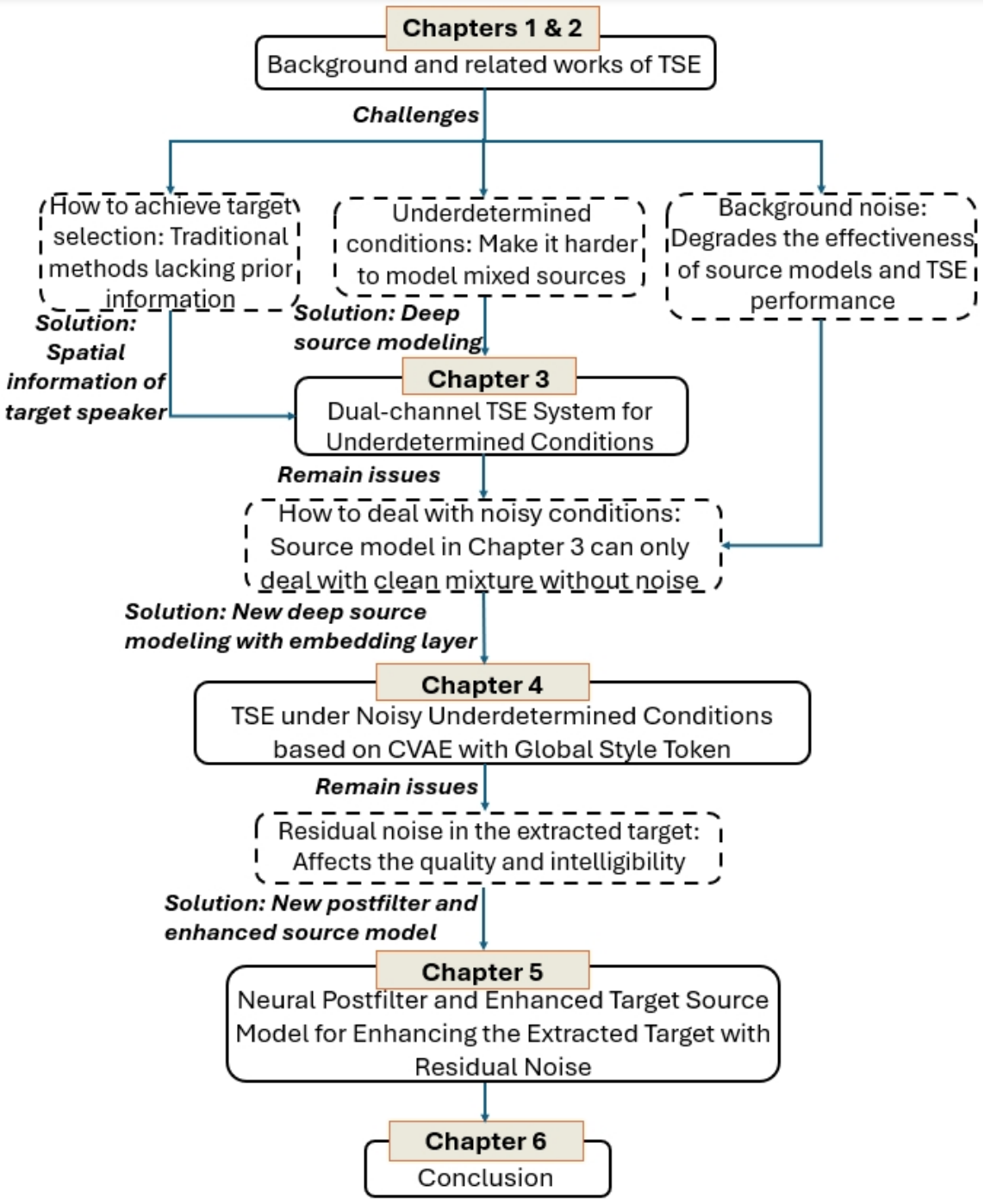


Figure 1.1: Scheme of this dissertation.

Chapter 2

Target Speaker Extraction

2.1 Overview

Generally speaking, TSE is a specific mission of sound source separation. Various separation methods have been proposed that form the theoretical basis for the implementation of TSE. This chapter gives a brief review of related BSS methods and describes the fundamental algorithm of these methods, which will be helpful for later discussions. Firstly, the formulation of the dual-channel TSE problem in underdetermined conditions is described in Sect. 2.2. By assuming the interference mixture except for the target speaker as one source, such a formulation can be interpreted as a special form of speech separation system based on separation matrix in underdetermined conditions. After that, this chapter provides a brief review of related separation methods based on signal independence. After Sect. 2.2, a spatial information-based method developed on the aforementioned basic separation methods is described in Sect. 2.3, which provides an effective means for im-

plementing TSE. Then, the next section of this chapter reviews the recently proposed new method based on deep source modeling. The final section summarizes this chapter.

2.2 Problem formulation of a dual-channel TSE problem in underdetermined conditions

Let us 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. These signals are expressed as

$$\mathbf{s}(f,n) = [s_1(f,n), s_2(f,n)]^T, \tag{2.1}$$

$$\mathbf{x}(f,n) = [x_1(f,n), x_2(f,n)]^T, \tag{2.2}$$

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. A separation system based on the demixing matrix $\mathbf{W}(f)$ is expressed as

$$\mathbf{s}(f,n) = \mathbf{W}^{\mathrm{H}}(f)\mathbf{x}(f,n), \tag{2.3}$$

$$\mathbf{W}(f) = [\boldsymbol{w}_1(f), \boldsymbol{w}_2(f)], \tag{2.4}$$

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 not easy. 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 $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$, where $j = 1, 2$ that denotes the index of each source. Based on the further assumption that $s_1(f,n)$ and $s_2(f,n)$ are independent of each other, $\mathbf{s}(f,n)$ then follows

$$\mathbf{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|\mathbf{0}, \mathbf{V}(f,n)), \tag{2.5}$$

where $\mathbf{V}(f,n) = \mathrm{diag}[v_1(f,n), v_2(f,n)]$. On the basis of Eqs. (2.3) and (2.5), $\mathbf{x}(f,n)$ follows

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, (\mathbf{W}^{\mathrm{H}}(f))^{-1}\mathbf{V}(f,n)\mathbf{W}(f)^{-1}). \tag{2.6}$$

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

$$\begin{aligned}\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) &\stackrel{c}{=} 2N\sum_f \log|\det\mathbf{W}(f)| \\ &\quad - \sum_{f,n,j}(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)|^2}{v_j(f,n)}),\end{aligned} \tag{2.7}$$

where $\stackrel{c}{=}$ denotes equality up to constant terms and source model parameters are presented as $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$. It means the equation holds except for an irrelevant constant, which does not affect the outcome of the optimization.

2.3 Related separation methods based on signal independence

The goal of BSS is to estimate the demixing matrix from the observed signals without any prior knowledge about the sources or the mixing process. Early foundational methods, such as ICA, were developed under the assumption of determined conditions, where the number of sources equals the number of microphones. These methods form the basis of BSS, leveraging the independence of source signals as a key criterion for separation. Deter-

mined conditions represent the fundamental scenario for BSS, providing a theoretical foundation that later methods have expanded upon to address more complex and realistic environments.

2.3.1 Basic theory of ICA and FDICA

ICA is a statistical method for blind source separation, designed to decompose observed mixed signals into statistically independent source signals, effectively serving as an inverse process of source mixing. It was originally developed for determined instantaneous mixtures in the time domain, under the assumption of no delays or reverberation. The core assumption of ICA is that the source signals are non-Gaussian and mutually independent. This assumption enables an effective solution to BSS problems in determined conditions, as the mixture of sources tends to follow a Gaussian distribution, even when the original sources are non-Gaussian, consistent with the central limit theorem.

However, ICA is limited to instantaneous mixtures and struggles to handle convolutive mixtures, where signals experience delays and reverberation during propagation in many realistic acoustic fields. To address these limitations, frequency-domain independent component analysis (FDICA) [11–14] was proposed. FDICA extends ICA by transforming the convolutive mixing problem in the time domain into a simpler instantaneous mixing problem in the time-frequency domain using the STFT. By applying ICA independently at each frequency bin, FDICA enables the separation of convolutive mixtures $\mathbf{x}(f, n)$ in the time-frequency domain, where demixing matrix $\mathbf{W}(f)$ is esti-

mated for the separation. Under such assumption, a simplistic probabilistic model for the sources can be formulated as

$$p(\mathbf{x}(f,n)|\mathbf{W}(f)) = |\mathbf{W}^{\mathrm{H}}(f)|^2 p(\mathbf{s}(f,n)) \tag{2.8}$$

$$= |\mathbf{W}^{\mathrm{H}}(f)|^2 \prod_j p(s_j(f,n)), \tag{2.9}$$

To estimate the demixing matrix with given source signals, the negative log-likelihood of $\mathcal{X}$ given $\mathcal{W}$ is applied as the objective function to be minimized is expressed as

$$J_{\mathrm{FDICA}}(\mathcal{X}|\mathcal{W}) = -\log p(\mathcal{X}|\mathcal{W}) \tag{2.10}$$

$$\stackrel{c}{=} -N \sum_f \log|\det \mathbf{W}(f)| - \sum_{f,n,j} (\log p(\boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)), \tag{2.11}$$

In many ICA applications under determined conditions, the source model of $p(s_j(f,n)) = p(\boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n))$ is empirically assumed to the conventional ones like Laplace distribution to describe a single source signal.

2.3.2 Basic theory of IVA

Although ICA and FDICA have been widely used for BSS, they face some limitations, particularly the permutation ambiguity. In ICA, it cannot inherently determine which component corresponds to which original source. This ambiguity arises because the separation is based solely on statistical independence, without any additional information or constraints linking the

output to the original sources. For FDICA, the ambiguity occurs because the separation at each frequency bin is independent so that the order of separated components (sources) can vary across different frequency bins.

IVA is a multivariate extension of FDICA and is used to avoid the permutation problem in BSS by making appropriate assumptions of the distribution of signals $p(\boldsymbol{s}_j(f,n))$. Unlike FDICA, which independently models each of the frequency components resulting in the permutation ambiguity problem, IVA models all the sources and separated signals as frequency vector variables. The source model in IVA treats all frequency bins as a single variable $\boldsymbol{s}_j(n) = [s_j(1;n), ..., s_j(F;n)]^{\mathrm{T}}$, which is assumed to follow a spherically symmetric multivariate distribution. This assumption allows for higher-order correlations between frequency components to be captured. The spherically symmetric property implies that the distribution depends solely on the norm of the multivariate vector, i.e., $p(\boldsymbol{s}_j(n)) = f(||\boldsymbol{s}_j(n)||)$. The objective function of IVA to be minimized that follows the negative log-likelihood is expressed as

$$J_{\mathrm{IVA}}(\mathcal{W}) = \sum_j \mathbb{E}[G(\boldsymbol{s}_j(n))] - 2\sum_f \log|\det \mathbf{W}(f)|, \tag{2.12}$$

where $G(\boldsymbol{s}_j(n)) = -\log p(\boldsymbol{s}_j(n))$ is called the contrast function. $p(\boldsymbol{s}_j(n))$ is the probability density function of the j-th separated signal $\boldsymbol{s}_j(n)$ known as the source model. In FDICA, the non-Gaussian source distribution such as the Laplace distribution of $p(\boldsymbol{s}_j(n))$ is assumed for each frequency component, whereas IVA assumes that the source model follows a non-Gaussian

spherically symmetric source distribution $p(\boldsymbol{s}_j(n))$ for the frequency vector variables like the spherically symmetric multivariate Laplace distribution. One typical choice of IVA source model is using a spherically symmetric multivariate Laplace distribution as a super-Gaussian distribution for modeling sources [15] [16] [47]. Under such an assumption of distribution, the contrast function is expressed as

$$G(\boldsymbol{s}_j(n)) = G_R(r_j(n)), \tag{2.13}$$

$$r_j(n) = ||\boldsymbol{s}_j(n)||_2 = \sqrt{\sum_f |s_j(f,n)|^2}, \tag{2.14}$$

where, $G_R(r_j(n))$ is a continuous and differentiable function of the real variable $r_j(n)$, ensuring that $G'_R(r_j(n))/r_j(n)$ remains continuous everywhere and monotonically decreases for $r_j(n) \geq 0$. Most of the IVA contrast functions employed in studies like [15] [16] satisfy the conditions of $G_R(r)$, including

$$G_R(r_j(n)) = Kr_j(n), \tag{2.15}$$

where K is a positive constant.

2.4 Geometric source separation method for target selection

In the traditional source separation method, the demixing matrix-based source separation processing aims to just separate the observed mixture signals into individual source signals. However, real-world applications often require additional information to select target speech post-separation, addressing output-channel permutation issues. As this dissertation mentioned in Chpater 1, there are several frameworks for utilizing spatial information to implement TSE, in which BSS has shown its own merit in incorporating signal independence and spatial information. This dissertation focuses on the BSS framework-based method.

2.4.1 Formulation of GCs-based TSE

In underdetermined conditions, the GSS framework applies GCs based on spatial information to implement TSE. Now, let us consider that GCs [26] restrict the far-field response of the jth demixing filter in the target DOA α, which is described as

$$J_{gc}(\mathcal{W}) = \sum_{j} \lambda_j \sum_{f} |\boldsymbol{w}_j^{\mathrm{H}}(f)\boldsymbol{d}(f,\alpha) - b_j|^2, \tag{2.16}$$

$$\boldsymbol{d}(f,\alpha) = \exp[-j(\mathbf{p}/c)f\cos(\alpha)], \tag{2.17}$$

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 [48]. If $b_j = 1$, the corresponding $\boldsymbol{w}_j(f)$ is estimated to form a delay-and-sum (DS) beamformer [49] toward α to preserve the target. In contrast, a small b_j value creates a null beamformer towards α, acting as a blocking matrix (BM) [50] that suppresses the target while estimating a mixture of all interferences. Following the formulation of Eq. (2.7) in Sect. 2.2, the overall objective function to be minimized is

$$J(\mathcal{W}, \mathcal{V}) = -\log p(\mathcal{X}|\mathcal{W}, \mathcal{V}) + J_{gc}(\mathcal{W}). \tag{2.18}$$

2.4.2 Geometric Constraint-Based IVA

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 [36], 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 geometrically constrained independent vector analysis (GCIVA), which merges a linear GC with the IVA framework.

The fundamental framework in GCIVA is IVA, which has been described

in Sect. 2.3.2. By applying the linear GC in IVA, the objective function of the GCIVA method is

$$J_{\mathrm{GCIVA}}(\mathcal{W}) = J_{\mathrm{IVA}}(\mathcal{W}) + J_{gc}(\mathcal{W}), \tag{2.19}$$

where the GC is given by Eq. (2.16) to restrict the far-field response of the jth estimated demixing filter by using the target DOA.

In the estimation processing of the demixing matrix in GCIVA, the auxiliary function approach [51] has been utilized, as demonstrated in the study [31]. This approach has led to the development of auxiliary function-based independent vector analysis (AuxIVA) [52], a fast and stable IVA algorithm developed. Rather than directly optimizing the original objective function in Eq. (2.12), which is challenging to solve analytically, the auxiliary function $J_{\mathrm{AuxIVA}}(\mathcal{W}, \mathcal{Q})$ is minimized in terms of $\mathcal{W}$ and the auxiliary variable $\mathcal{Q}$, which is expressed as

$$J_{\mathrm{AuxIVA}}(\mathcal{W}, \mathcal{Q}) = \sum_{j} \sum_{f} \{\frac{1}{2} \boldsymbol{w}_j^{\mathrm{H}}(f) Q_j(f) \boldsymbol{w}_j(f) - \log|\det \mathbf{W}(f)|\}, \tag{2.20}$$

where $\mathcal{Q} = \{Q_j(f)\}_{j,f}$, and $Q_j(f)$ is the weighted covariances expressed as

$$Q_j(f) = \mathbb{E}[\frac{G'_R(r_j(n))}{r_j(n)} \mathbf{x}(f) \mathbf{x}^{\mathrm{H}}(f)], \tag{2.21}$$

Therefore, the optimized objective function of GCIVA by using AuxIVA

is

$$J^{+}_{\mathrm{GCIVA}}(\mathcal{W},\mathcal{Q}) = J_{\mathrm{AuxIVA}}(\mathcal{W},\mathcal{Q}) + J_{gc}(\mathcal{W}), \tag{2.22}$$

The update rule for demixing matrix $\mathbf{W}(f)$ is derived based on the idea adopted in vectorwise coordinate descent (VCD) [53], which is noteworthy for its fast convergence, low computational cost, and nonrequirement of the step-size parameter. This allows the demixing matrix to be refined iteratively through the objective function Eq. (2.22). The derived update rules are summarized as

$$\boldsymbol{u}_j = \boldsymbol{D}_j^{-1}\mathbf{W}(f)^{-1}\boldsymbol{e}_j, \tag{2.23}$$

$$\hat{\boldsymbol{u}}_j = \lambda_j b_j \boldsymbol{D}_j^{-1}\boldsymbol{d}_j, \tag{2.24}$$

$$h_j = \boldsymbol{u}_j^H \boldsymbol{D}_j \boldsymbol{u}_j, \tag{2.25}$$

$$\hat{h}_j = \boldsymbol{u}_j^H \boldsymbol{D}_j \boldsymbol{u}_j, \tag{2.26}$$

$$\boldsymbol{w}_j(f) = \begin{cases} \frac{1}{\sqrt{h_j}}\boldsymbol{u}_j + \hat{\boldsymbol{u}}_j & (\text{if } \hat{h}_j = 0), \\ \frac{\hat{h}_j}{2h_j}[-1+\sqrt{1+\frac{4h_j}{|\hat{h}_j|^2}}]\boldsymbol{u}_j + \hat{\boldsymbol{u}}_j & (\text{o.w.}). \end{cases} \tag{2.27}$$

where $\boldsymbol{D}_j = Q_j(f) + \lambda_j \boldsymbol{d}_j \boldsymbol{d}_j^{\mathrm{H}}$ and $\boldsymbol{e}_j$ is the jth column of the identity matrix. The algorithm of GCIVA in study [31] is summarized as follows

Algorithm 1 GCIVA Algorithm [31]

Require: Observed mixture signal $\mathbf{x}(f,n)$, iteration number L

Initialize $\mathcal{W}$ with identity matrix
for $l = 1$ to L **do**
  for $j = 1$ of J **do**
    $s_j(f,n) = \boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)$
    (updating auxiliary variables)
    initialize $Q_j(f)$ using Eq. (2.21)
    (updating demixing matrices)
    update $\boldsymbol{w}_j(f)$ using Eqs. (2.23) to (2.27)
  end for
end for

2.5 Limitations of conventional methods in implementing TSE in underdetmined conditions

Traditional separation methods based on signal independence, such as ICA, FDICA, and IVA, face several limitations in underdetermined conditions, as they primarily focus on separating observed mixture signals into individual sources. More recently, GC-based methods like GCIVA have demonstrated potential in achieving TSE. However, GCIVA also encounters significant challenges when applied to underdetermined conditions. One major issue is the use of a spherically symmetric multivariate Laplace distribution

as the source model, which is not well-suited for underdetermined scenarios where the number of sources exceeds the number of microphones. The source model in GCIVA assumes uniform variance across frequency bins and models all frequency bins as multivariate variables, which limits its adaptability in complex environments with diffuse noise and multi-speaker mixture. These assumptions restrict the effectiveness of GCIVA in accurately modeling and enhancing target speech in underdetermined conditions, where more robust and flexible source models are required.

2.6 Deep neural network method based on new source model

2.6.1 VAE and CVAE source method

Methods like ICA, IVA, and GCIVA, despite integrating spatial information, face limitations in 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 variational autoencoders (VAEs) and generative adversarial networks (GANs), as highlighted in recent studies [54–56], 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. [57] and others [58–60], include the application of VAEs for enhanced noise modeling

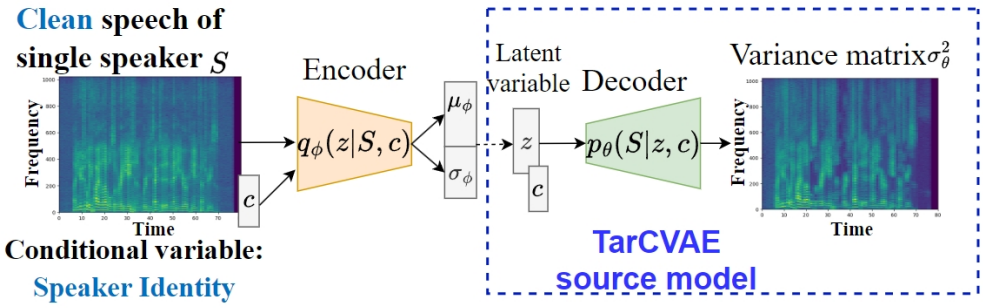


Figure 2.1: Illustration of TarCVAE.

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 [39] first introduced the conditional VAE (CVAE) [40] model in multi-channel speech separation. Figure 2.1 shows illustrations of CVAE. In this dissertation, the CVAE in MVAE is called the target CVAE (TarCVAE). 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. 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.

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

$$\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},\boldsymbol{c})\sim p_D(\boldsymbol{S},\boldsymbol{c})}[\mathbb{E}_{z\sim q_\phi(z|\boldsymbol{S},\boldsymbol{c})}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})] - \mathrm{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})||p(\boldsymbol{z})]], \tag{2.28}$$

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 Gaussian distributions:

$$q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c}) = \prod_k \mathcal{N}(\boldsymbol{z}(k)|\mu_\phi(k;\boldsymbol{S},\boldsymbol{c}),\sigma^2_\phi(k;\boldsymbol{S},\boldsymbol{c})), \tag{2.29}$$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I}), \tag{2.30}$$

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

$$p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c},g)=\prod_{f,n}\mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|0,v(f,n)), \tag{2.31}$$

$$v(f,n)=g\cdot\sigma^2_\theta(f,n;\boldsymbol{z},\boldsymbol{c}), \tag{2.32}$$

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 of the generated spectrogram. In the separation process of MVAE, only the decoder is used as the source model, and the latent variable $\boldsymbol{z}$ and the conditioning variable $\boldsymbol{c}$ are updated using the back propagation based on the IVA objective function in the demixing matrix estimation.

2.6.2 Limitations of MVAE method based on CVAE source model in underdetmined conditions

An MVAE trains a TarCVAE using power spectrograms of clean speech samples along with the corresponding speaker ID as auxiliary label inputs, enabling the trained decoder output distribution to serve as a universal generative model for source signals. While MVAE has demonstrated impressive performance in determined cases, its effectiveness remains limited in underdetermined scenarios. Moreover, as a BSS-based approach, MVAE lacks prior information guidance, making it incapable of effectively selecting the target speaker.

2.7 Evaluation metric

An important aspect of developing speech separation algorithms is evaluating the quality of separated signals by comparing them to reference signals. Metrics such as SIR signal-to-interference ratio (SIR), signal-to-distortion ratio (SDR), and signal-to-artifacts ratio (SAR), are widely used evaluation metrics, as they offer a comprehensive assessment of separation quality [41].

In this dissertation, several metrics are selected to make evaluations. SIR measures the suppression of interference from other sources, reflecting the algorithm's effectiveness in isolating the target signal from other competing signals. SDR evaluates the overall quality of the separated signal by considering all components of distortion in the estimated signal. SAR shows the amount of the true source in relation to unwanted artifacts in the estimation. In speech separation, a given estimate $\hat{s}(t)$ of a source $s_i(t)$ is decomposed as a sum

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t), \tag{2.33}$$

where, $s_{target}(t)$ represents a permissible deformation of the target source $s_i(t)$, $e_{interf}(t)$ accounts for permissible deformations of interferences, $e_{noise}(t)$ represents a permissible deformation of perturbing noise, and $e_{artif}(t)$ denotes artifacts introduced by the separation process, such as musical noise or artificial spectral structures. The metrics are defined as

$$\text{SDR} = 10\log_{10}\frac{||s_{target}(t)||^2}{||e_{interf}(t)+e_{noise}(t)+e_{artif}(t)||^2}, \tag{2.34}$$

$$\text{SIR} = 10\log_{10}\frac{||s_{target}(t)||^2}{||e_{interf}(t)||^2}, \tag{2.35}$$

$$\text{SAR} = 10\log_{10}\frac{||e_{interf}(t)+e_{noise}(t)+e_{artif}(t)||^2}{||e_{artif}(t)||^2}, \tag{2.36}$$

In noisy conditions, the signal-to-noise ratio (SNR) serves as a metric to evaluate the effectiveness of noise suppression of an estimation source. It is defined as

$$\text{SNR} = 10\log_{10}\frac{||\hat{s}(t)||^2}{||s_i^2(t)||^2}, \tag{2.37}$$

2.8 Summary of Chapter 2

This chapter provided an overview of Target speaker extraction (TSE) and reviews related separation methods, forming the theoretical foundation for the proposed approach. The problem formulation of dual-channel TSE in underdetermined conditions was introduced, highlighting the challenges posed by multiple interfering sources and limited microphones. Conven-

tional blind source separation (BSS) methods, including ICA, FDICA, and IVA, were discussed, followed by the introduction of the geometric source separation (GSS) framework and geometric constraint-based IVA (GCIVA) method, which utilizes spatial information as the cue to achieve target selection. The limitations of these methods in underdetermined conditions, particularly in source modeling, were addressed. To address these limitations, deep generative models, specifically Conditional Variational Autoencoders were introduced as a more powerful approach for modeling source signals. The multichannel variational autoencoder (MVAE) method was also reviewed, highlighting its strengths in determined conditions and its limitations in underdetermined scenarios. Finally, this chapter presented the evaluation metrics used throughout the study, including SIR, SDR, SAR, and SNR, which assess separation quality and robustness. These discussions laid the groundwork for the development of the proposed TSE framework in subsequent chapters.

Chapter 3

Dual-channel TSE System for Underdetermined Conditions

3.1 Overview

This chapter details the proposed direction-aware TSE approach for underdetermined conditions, addressing two main challenges. Firstly, the proposed framework incorporates 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, this chapter introduces a novel CVAE, named Interference CVAE (IntCVAE). IntCVAE is designed to effectively model mixed speech signals, particularly in situations involving varying numbers of speakers.

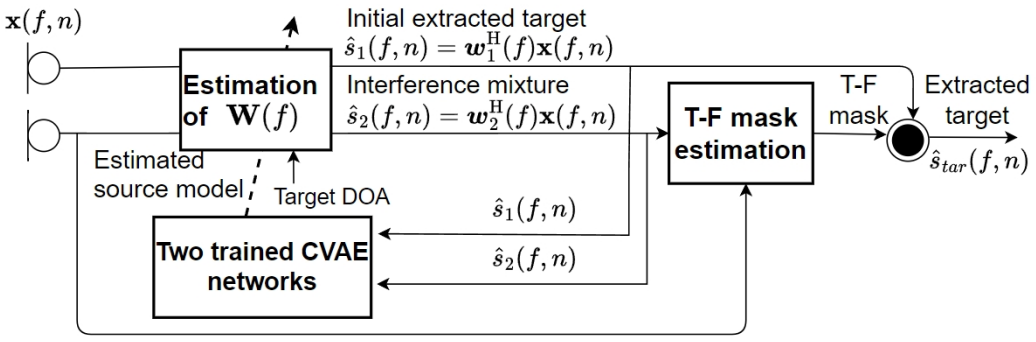


Figure 3.1: Proposed framework of directional target speaker extraction based on dual-channel system.

3.2 Direction-aware TSE method in underdetermined conditions

3.2.1 Proposed framework

Figure 3.1 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 Eq. (2.16) is set to 1 to create a delay-and-sum (DS) beamformer, which yields a spatial beamformer towards the direction of the target. So the GC on the target channel is shown as

$$J_{gc}^{\mathrm{one}}(\mathcal{W}) = \lambda_1 \sum_f |\boldsymbol{w}_1^{\mathrm{H}}(f)\boldsymbol{d}(f,\alpha) - 1|^2. \tag{3.1}$$

A preliminary estimation of the target can be obtained by calculating

$$s_1(f,n) = \boldsymbol{w}_1^{\mathrm{H}}(f)\mathbf{x}(f,n). \tag{3.2}$$

On channel 2 on the other hand, also called the interference channel, $b_2 = 0$ is set in Eq. (2.16) 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

$$J_{gc}^{\mathrm{null}}(\mathcal{W}) = \lambda_2 \sum_f |\boldsymbol{w}_2^{\mathrm{H}}(f)\boldsymbol{d}(f,\alpha)|^2. \tag{3.3}$$

The interference mixture can be obtained from the output of the interference channel as

$$s_2(f,n) = \boldsymbol{w}_2^{\mathrm{H}}(f)\mathbf{x}(f,n). \tag{3.4}$$

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

3.2.2 CVAE-Based target and interference source models

To extract the target speaker in the underdetermined condition of multiple interfering speakers, it is desired to accurately model the single target speaker's speech and interference mixture speech. Two CVAEs are used to model these two components. The first is TarCVAE from MVAE, which has been introduced in Sect. 2.6.1. The second is the proposed interference CVAE (IntCVAE), which will be elaborated in the following parts of this section.

Figure 3.2 illustrates the structure of IntCVAE. TarCVAE, originally introduced in MVAE [39], is utilized on the target channel to model the single target signal. On the interference channel, IntCVAE is proposed as the source model. Unlike TarCVAE, which models a single speaker, IntCVAE takes a mixture of multiple speakers as input $\boldsymbol{S}$, while $\boldsymbol{c}$ is a one-hot vector representing the number of speakers present in the mixture. 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 process of the demixing matrix. By incorporating TarCVAE and IntCVAE on the two respective channels, the proposed framework enables effective modeling of the target speaker and the interference mixture, both of which are separated via the GCs-based approach. This design establishes a robust and adaptable source modeling strategy for dual-channel systems operating in underdetermined conditions.

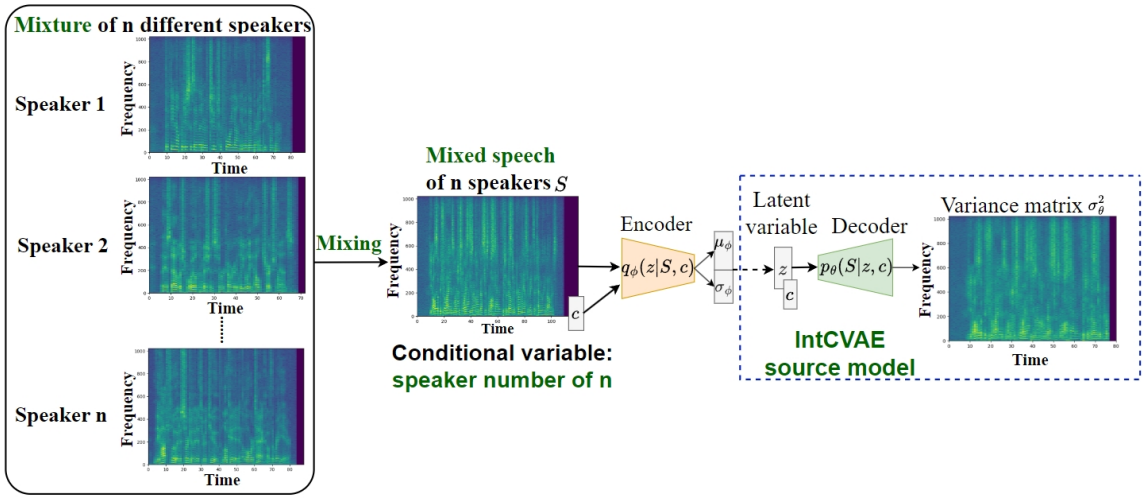


Figure 3.2: Illustration of IntCVAE.

3.2.3 TSE algorithm

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 Eq. (2.7).

The update rule for $\mathbf{W}(f)$ is the VCD method, which is derived from the GCIVA method in Sect. 2.4.2. This enables the iterative refinement of the demixing matrix using the objective function defined in Eq. (2.18). Assuming only a dual-channel case as in Sect. 2.2, the derived update rules are summarized as

$$\boldsymbol{u}_j = \boldsymbol{D}_j^{-1}\mathbf{W}(f)^{-1}\boldsymbol{e}_j \qquad (j=1,2), \tag{3.5}$$

$$\hat{\boldsymbol{u}}_1 = \lambda_1 \boldsymbol{D}_1^{-1}\boldsymbol{d}, \tag{3.6}$$

$$h_j = \boldsymbol{u}_j^H \boldsymbol{D}_j \boldsymbol{u}_j \qquad (j=1,2), \tag{3.7}$$

$$\hat{h}_1 = \boldsymbol{u}_1^H \boldsymbol{D}_1 \hat{\boldsymbol{u}}_1, \tag{3.8}$$

$$\boldsymbol{w}_j(f) = \begin{cases} \frac{\hat{h}_1}{2h_1}[-1+\sqrt{1+\frac{4h_1}{|\hat{h}_1|^2}}]\boldsymbol{u}_1 + \hat{\boldsymbol{u}}_1 & (j=1), \\ \frac{1}{\sqrt{h_2}}\boldsymbol{u}_2 & (j=2), \end{cases} \tag{3.9}$$

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 ($j=1,2$). 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

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j^{\mathrm{H}}(f)\mathbf{x}(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z},\boldsymbol{c})}. \tag{3.10}$$

where F and N refers to the number of frequency indices f and time indices n. The proposed algorithm is thus summarized as follows

Algorithm 2 CVAE-based TSE

Require: Network parameters θ and ϕ of two CVAEs trained using Eq. (2.28), 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 Eq. (3.10)
    for $k = 1$ to 100 **do**
      update Ψ using BP with $\log p_\theta(\boldsymbol{S}|\boldsymbol{z}, \boldsymbol{c})$
      while keeping θ fixed
    end for
    compute $v(f,n)$ using Eq. (2.32)
    (updating demixing matrices)
    update $\boldsymbol{w}_j(f)$ using Eqs. (3.5) to (3.9)
  end for
end for

Unlike the GCIVA algorithm in Sect. 2.4.2, where the demixing matrix is updated using auxiliary variables $Q_j(f)$ derived from the assumed GCIVA source model based on spherically symmetric multivariate Laplace distribution for single source modeling, the CVAE-based TSE method updates the source model parameters $v_j(f,n)$ iteratively through the trained CVAE in

each iteration. The trained IntCVAE offers greater flexibility in modeling the interference mixture, enabling more accurate estimation of source model parameters. This adaptability makes it better suited for the estimation algorithm of the demixing matrix in underdetermined conditions.

3.2.4 Postprocessing based on T-F Mask

In underdetermined conditions with multiple interferences, our GC-based method generates 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 is developed for postprocessing. 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

$$\hat{s}_{tar}(f,n) = \hat{s}_1(f,n)(1 - \frac{|\hat{s}_2(f,n)|^2}{|\mathbf{x}(f,n)|^2}). \tag{3.11}$$

3.3 Improved TSE method against DOA errors

3.3.1 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 in DOA information will bring significant errors to the steering vector, which is the main reason for the degradation of the performance in many systems [61] [62] [63]. Especially in 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 [64]. Among them, the imposition of multiple linear constraints along with minimum variance beamformer has been considered a useful method [65–68]. 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 steering vector.

Therefore, to address this problem, this section proposes a robust TSE algorithm against DOA estimation errors based on the former framework. The objective function of the estimation of the demixing matrix is refined to enable the given DOA can be updated in this processing.

3.3.2 Improved method with DOA modification

In estimating the demixing matrix, the objective function is shown by Eq. (2.18). 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, this section improves 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

$$J_c(\alpha|\alpha_0) = \lambda_\alpha ||\alpha - \alpha_0||_2^2. \tag{3.12}$$

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

$$\begin{aligned} &\mathcal{L}(\mathcal{W}, \mathcal{V}, \alpha) \\ &= -\log p(\mathcal{X}|\mathcal{W}, \mathcal{V}) + J_{gc}(\mathcal{W}, \alpha) + J_c(\alpha|\alpha_0). \end{aligned} \tag{3.13}$$

To obtain the optimal DOA, a GD (gradient descent)-based algorithm is adopted 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 Sect. (3.2.3), the improved algorithms with variable DOAs are summarized as follows

Algorithm 3 CVAE-based TSE with DOA modification

Require: Network parameters θ and ϕ of two CVAEs trained using Eq. (2.28), 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 Eq. (3.10)
    for $k = 1$ to 100 **do**
      update Ψ using BP with $\log p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})$
      while keeping θ fixed
    end for
    compute $v(f,n)$ using Eq. (2.32)
    (updating demixing matrices)
    update $\boldsymbol{w}_j(f)$ using Eqs. (22) to (26)
    (updating DOA)
    for $h = 1$ to 100 **do**
      update DOA α using GD with Eq. (3.13), (2.16), and (2.17)
    end for
  end for
end for

3.4 Experimental evaluations

3.4.1 Training condition

The training data was from the Wall Street Journal (WSJ0) corpus [69]. The WSJ0 folder si_tr_s (around 25 h) was used 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. Nine groups of a mixture of speeches of 2 to 10 speakers with 200 utterances per group (around 9 h) were used. 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 can maintain consistent energy levels across all sources, resulting in an evenly balanced audio mix where no single speaker's voice dominates the composite signal.

The architectures of networks are described as follows. The CVAE in both TarCVAE and IntCVAE was the same architecture as in [39], designed with an encoder, a latent space, and a decoder. The encoder consists of three convolutional layers: two two-dimensional gated CNN layers followed by a regular two-dimensional CNN layer. These layers can incrementally encode the input spectrogram with a conditional variable, converting it into the latent space. The decoder mirrors the encoder with two two-dimensional gated CNN layers and a final two-dimensional convolutional transpose layer, enabling it to reconstruct the input spectrogram. The CVAEs were trained

Table 3.1: Average SDRs [dB] of clean signal and mixed signal outputs obtained by different CVAEs.

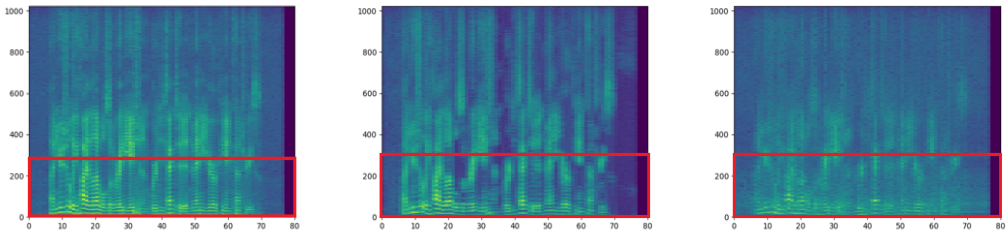
	single speech	mixed speech
TarCVAE	18.25	13.65
IntCVAE	15.57	17.74

using the Adam optimizer, with learning rates of 0.0001 for the CVAEs, and were trained 1000 epochs. All implementations were based on PyTorch 1.8.1, with hardware conditions of a computer with Intel(R) Xeon(R) Gold 6248 CPU@ 2.50GHz, 32GB RAM, and one NVIDIA RTX 3090 GPU.

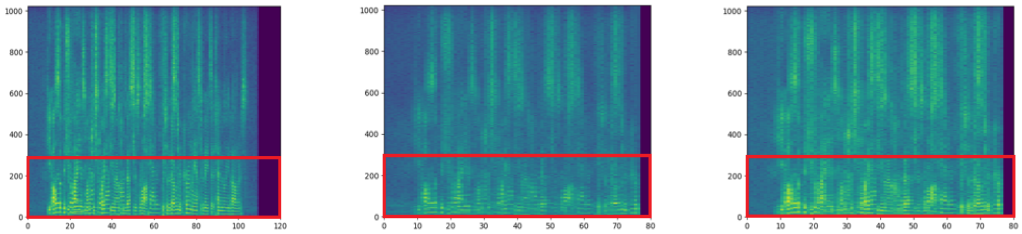
3.4.2 Evaluation of the reconstruction power

Evaluation condition

To evaluate the reconstruction ability of our trained CVAEs on single speech and mixed speech signals, this section took the clean signals of one speaker and the mixed signals of two speakers as the inputs of TarCVAE and IntCVAE and calculated the 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, 50 utterances were randomly selected 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.



(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.

Figure 3.3: Magnitude spectrograms of reference sources and sources reconstructed by CVAEs.

Evaluation results

Table 3.1 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. 3.3 shows examples of the CVAE source model fitted to the spectrogram of the original clean and mixed speech. As shown by Figs. 3.3.(a), 3.3.(b), and 3.3.(c), it can be observed 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 Figs.

3.3.(c), 3.3.(d), and 3.3.(e), it can be observed 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 in underdetermined conditions.

3.4.3 TSE performance in underdetermined conditions

Evaluation condition

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) [70]. 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 the position of speakers. Figure 3.4 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 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. The images of three speakers were mixed using SIR uniformly. 60 tests under each reverberation condition were conducted. The average length of the test utterance was 10 seconds.

Table 3.2: 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	✓

The evaluation selected GCIVA and MVAE as the baseline methods, and to conduct an ablation study on different components of our proposed methods, our designed T-F mask was incorporated 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 3.2 presents a comparison between the baseline and proposed methods.

We computed the SDR, SIR, and 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, the signal in the direction α was set 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, all separated signals were evaluated and the one with the best evaluation result was selected as the

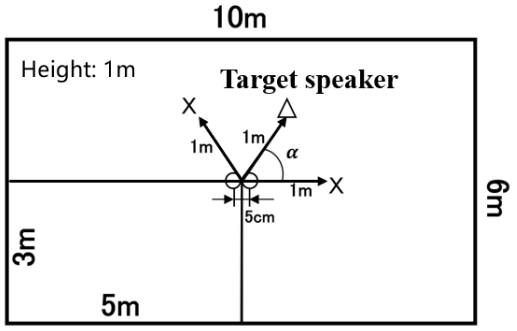


Figure 3.4: Configuration of evaluation, where $\triangle$ and $\times$ denote the target and interferences, respectively, and α is the DOA of the target.

extracted target.

Evaluation results

Table 3.3 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, it can be seen 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, it can be observed that a more powerful source model can bring improvement to 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, especially in terms of SIR and SDR ($p < 0.05$ in paired samples t-test).

Table 3.3: 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**

3.4.4 Evaluation of the impact of the angle between sources and distance between sources and microphones on the performance of the proposed method

In the previous section, the effectiveness of the proposed method under the underdetermined conditions was confirmed. Considering that the selection of targets depends on 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 [71] [72]. 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, the impact of the angle between the desired target and the nearest interference was evaluated, which can be considered as the spatial resolution of the GC-based TSE method.

Evaluation condition

For evaluation, RIRs by ISM for the same room shown in Fig. 3.4 with a reverberation time of $RT_{60} = 150$ ms were simulated. Three speakers were randomly located in the range of 0°–180°. The test dataset was the same as

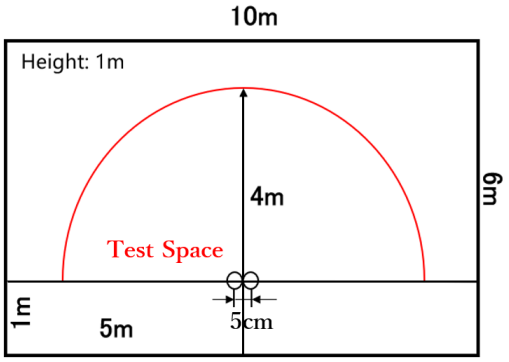


Figure 3.5: Configurations of the test space.

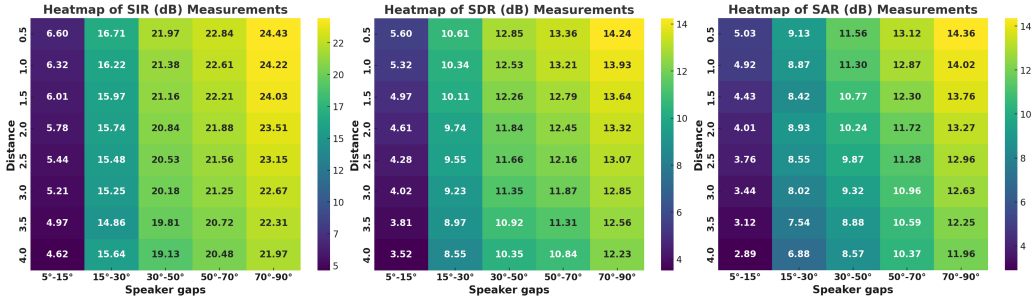


Figure 3.6: Average SDR, SIR, and SAR of proposed method in 3-speaker case.

those in Sect. 3.4.3. 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. 3.5.

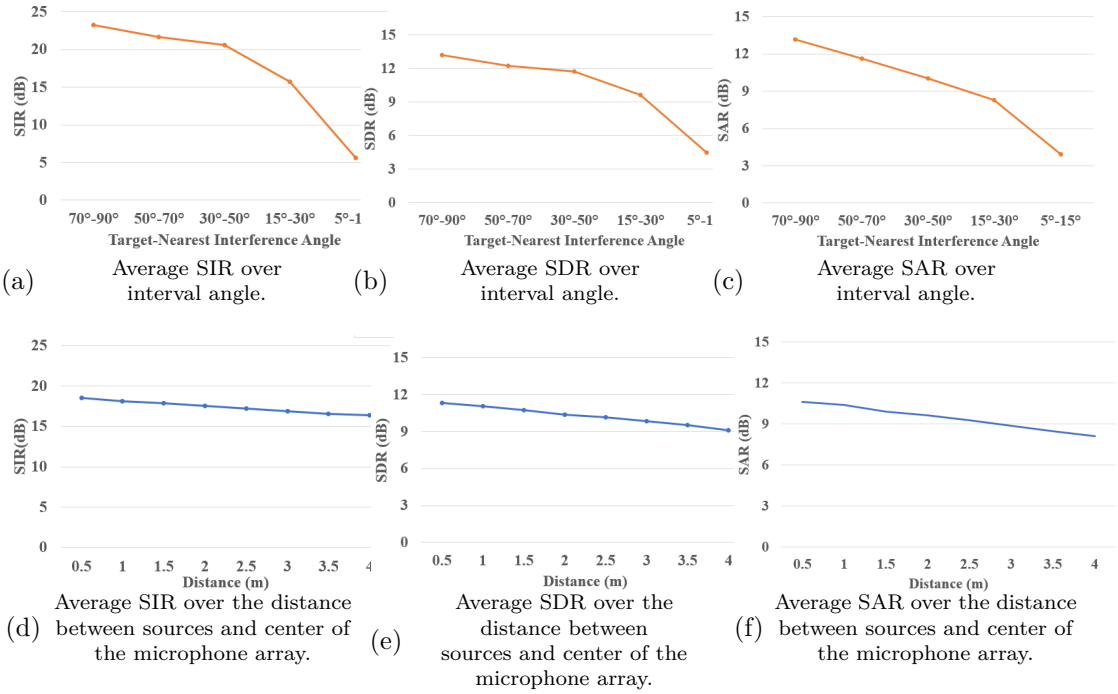
Evaluation results

Figure 3.6 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 Figs. 3.7(a)-3.7(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. It can also be observed that when the angle is less than 15 degrees, such a decline trend becomes more significant.

3.4.5 Evaluation of DOA modification

Evaluation condition

In this section, the impact of DOA errors on our proposed method and the robustness of the proposed method against DOA errors were evaluated. 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



(a) Average SIR over interval angle.

(b) Average SDR over interval angle.

(c) Average SAR over interval angle.

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

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

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

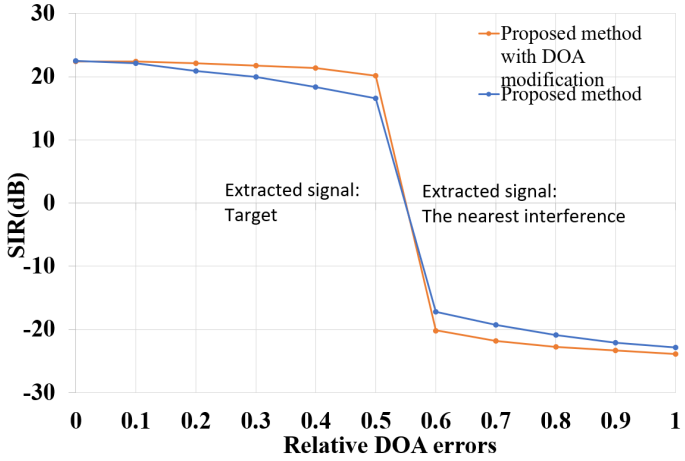
Figure 3.7: Average performance of the proposed method with different interval angles and distances between sources and center of the microphone array.

tend to extract the target signal from the mixed signal. Therefore, instead of using the absolute error for the evaluation, this evaluation 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. 3.4. 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

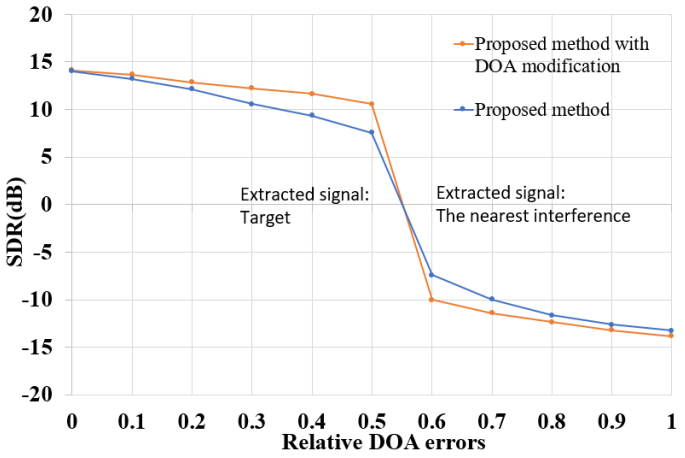
interference source nearest to the target. The evaluation was carried out in the simulated room shown in Fig. 3.5 with $RT_{60} = 150$ ms. Three sources were located randomly with an interval angle between the target and the nearest interference of 30°–90°, and the distance of each source from the center of the microphone array was 1 meter. This evaluation focused on SIR and SDR as they better reflect separation performance in directional TSE systems. Changes in SIR and SDR indicate the directional bias of the extracted signal. For instance, significant drops or negative values in SIR and SDR suggest that the extracted signal is shifting toward interference sources. SAR primarily assesses artifacts introduced during processing and its interpretation is less direct compared to SIR and SDR, as it reflects processing artifacts that may not directly correlate with spatial separation performance. SDR already encompasses the trade-off between distortion and artifacts. For simplicity and clarity, this section limited the analysis to SIR and SDR.

Evaluation results

Figure 3.8 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 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



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



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

Figure 3.8: Average SIR and SDR with different relative DOA errors of two proposed methods.

it is more robust to DOA errors than the previously proposed method without DOA modification. Additionally, it can be observed from Fig. 3.8 that the SIR and SDR of these two methods become negative when the relative error exceeds 0.5 times the interval angle. This relative error of 0.5 corresponds to the case where the DOA provided to the system lies equidistant between

the target speaker and the nearest interfering speaker in terms of angular separation. For example, if the angular separation between the target and the nearest interferer is 60°, a relative error of 0.5 indicates that the given DOA deviates by 30° toward the interfering speaker. In such cases, the extracted signal transitions from the target speaker to the nearest interferer due to the skewed input DOA, which highlights the importance of the proposed DOA modification method robust against DOA errors.

3.5 Summary of Chapter 3

This chapter 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, this chapter first utilizes CVAE to model the mixed speech signal, which overcomes the limitations of the source model in the traditional BSS algorithm in underdetermined conditions. As another contribution, the TSE algorithm with DOA modification is proposed to overcome the negative impact of DOA estimation errors.

The 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 better performance in underdetermined conditions. (3) Owing to the 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 reduced the negative impact of DOA estimation errors. It is worth noting that this chapter only focuses on the processing of clean environments without background noise.

Chapter 4

TSE in Noisy Underdetermined Conditions based on CVAE with Global Style Token

4.1 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 Chapter 3 has shown its effectiveness in underdetermined conditions, it still has limitations in modeling noisy mixed speech when environmental noise exists. One 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.

To address this issue, this section proposed a new source model called GST-IntCVAE (GIntCVAE for short) for modeling noisy interference mixture signals. GIntCVAE introduces GSTs [73] 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.

4.2 Source model for noisy underdetermined conditions based on CVAE with GST

Figure 4.1 illustrates the proposed GIntCVAE. TarCVAE is the same as Fig. 2.1 in Sect. 2.6.1, which 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

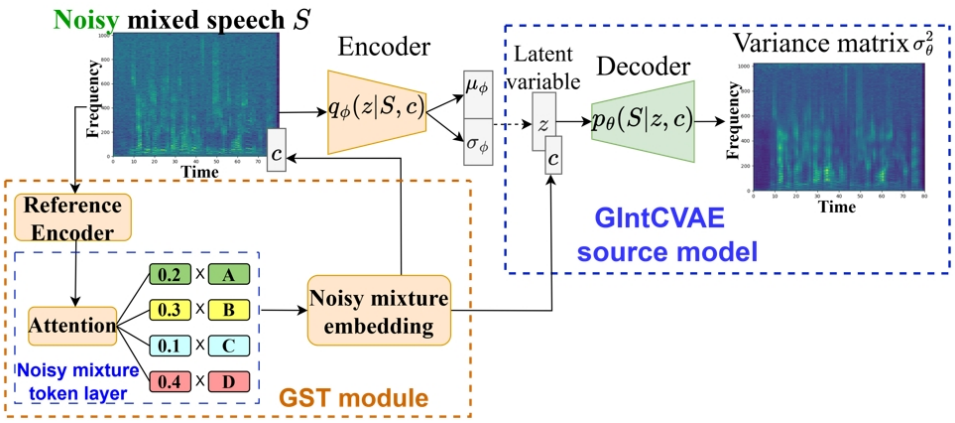


Figure 4.1: Illustration of and GIntCVAE.

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.

4.3 Inference processing of the new proposed method

In the inference stage, only the decoder is used to model the source and output distribution parameters, with GST weights being updated using BP

while fixing the noisy mixture tokens. The demixing matrix is updated as iteratively as the algorithm in Sect. 3.2.3. The algorithm of demixing matrix estimation in this section 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. 3.5 to 3.9.

(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. 3.10.

(d) Update v by calculating Eq. 2.32.

4.4 Experimental evaluations

4.4.1 Training conditions

The TarCVAE is the same as Sect. 3.4.1. For GIntCVAE, it 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. The noisy training dataset was generated by mixing the clean dataset with 4 types of diffuse noise at varying SNR levels from the DEMAND dataset [74], which contains 6 types of diffuse noise. There were 4 SNR conditions for training of GIntCVAE: 0, 10, 20, and 30 dB.

The architectures of our networks are described as follows. The CVAEs in TarCVAE and GIntCVAE is the same as Sect. 3.4.1. For the GST mod-

ule, the architecture follows [73], containing a reference encoder and a style token layer (SLT). 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 SLT 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 other training settings like the learning rate are the same as Sect. 3.4.1.

4.4.2 Investigation of embedding space of GIntCVAE

This evaluation investigated the embedding space of the trained GIntCVAE by visualizing the latent representations produced by the trained GST, which aims to analyze what the trained GST learned 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, the evaluation wants to determine the impact of the number of speakers and SNR on GST's capability to discriminate noisy mixed speech.

Evaluation condition

In the evaluation, different SNR conditions of noisy mixed signals from different numbers of speakers were used 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. The t-distributed stochastic neighbor embedding (t-SNE) [75] was used to compress all the 128-dimensional GST embedding outputs to 2D representations.

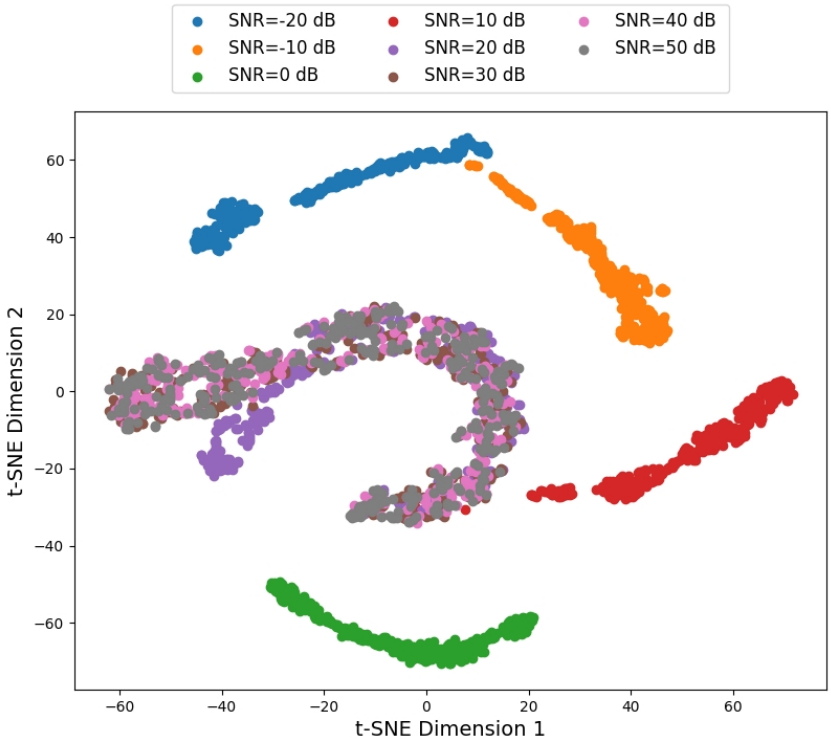


Figure 4.2: t-SNE visualization of GST by SNR conditions.

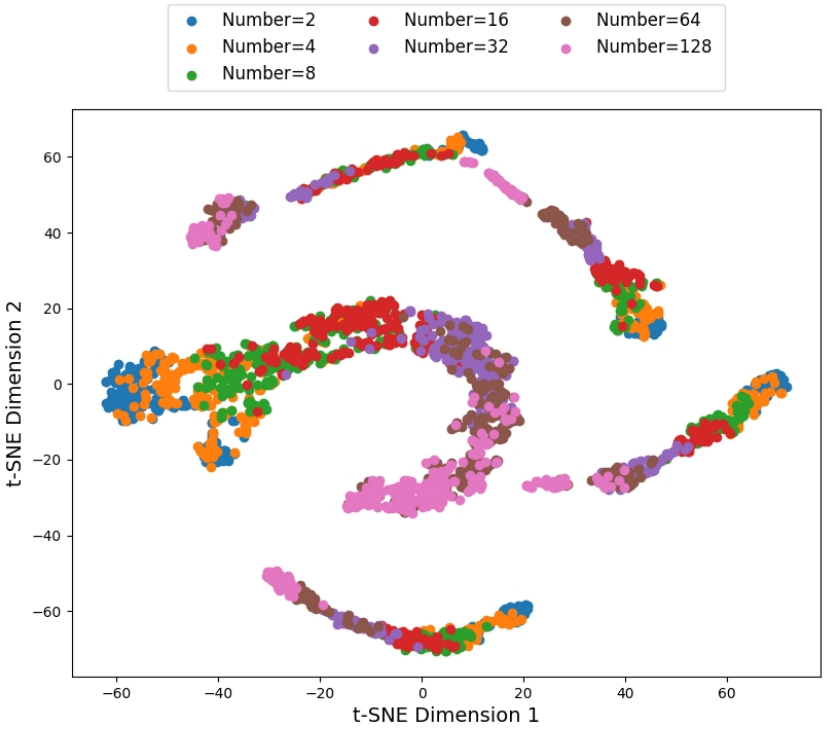


Figure 4.3: t-SNE visualization of GST by numbers of speakers.

Evaluation results

We visualized the features compressed by t-SNE with the same input 8×7×50 samples, using eight colors to represent different SNR conditions and seven colors to represent varying numbers of speakers. Figures 4.2 and 4.3 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.

4.4.3 Experimental evaluations of TSE in noisy underdetermined conditions

Evaluation condition

This evaluation tested the proposed method and baselines on three-source mixtures with a fixed reverberation time of 150 ms. Two-channel recordings synthesized by the ISM were applied to create mixtures of three speakers with noise. The simulated room is shown as in Fig. 3.4 in Chapter 3. The speakers were randomly located at angles from 0 degrees to 180 degrees in different directions, with a minimum angle between speakers set to 10 degrees. The three 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. The methods were evaluated under 3 different SNR

conditions of -10, 10, and 30 dB, using the SDR, SIR, SAR, and SNR as performance metrics.

We compared our proposed methods to several baselines including GCIVA [31] and MVAE [39]. To evaluate the effect of GIntCVAE in TSE performance, which introduces GST in IntCVAE, this evaluation also used a method from Chapter 3 as a baseline, which applied TarCVAE+IntCVAE.

Table 4.1: Average SDR, SIR, SAR, and SNR [dB] of the extracted target in noisy underdetermined environment of SNR = −10 dB.

Method	SIR	SDR	SAR	SNR
GCIVA	−3.68	−8.25	−6.39	−5.31
MVAE	−3.12	−7.12	−2.44	−4.78
TarCVAE + IntCVAE	1.03	−3.24	1.25	−2.23
TarCVAE + GIntCVAE	**5.34**	**1.55**	**3.79**	**2.03**

Table 4.2: Average SDR, SIR, SAR, and SNR [dB] of the extracted target in noisy underdetermined environment of SNR = 10 dB.

Method	SIR	SDR	SAR	SNR
GCIVA	2.41	1.78	5.06	1.08
MVAE	4.68	2.14	6.59	2.13
TarCVAE + IntCVAE	8.76	4.43	6.15	4.37
TarCVAE + GIntCVAE	**13.32**	**9.53**	**10.15**	**7.27**

Table 4.3: Average SDR, SIR, SAR, and SNR [dB] of the extracted target in noisy underdetermined environment of SNR = 30 dB.

Method	SIR	SDR	SAR	SNR
GCIVA	6.53	4.38	9.38	4.53
MVAE	9.48	7.18	10.47	5.82
TarCVAE + IntCVAE	15.11	9.61	11.19	8.89
TarCVAE + GIntCVAE	**21.12**	**14.27**	**12.56**	**11.81**

Evaluation results

The evaluation results under different noisy conditions are summarized in Tables 4.1, 4.2, and 4.3. These tables show that the new proposed method in this chapter consistently achieves enhanced performance across various noisy conditions, with notable improvements in SIR and SDR ($p < 0.05$). For example, at an SNR of -10 dB, the method using TarCVAE and GIntCVAE achieves an improvement in SIR of 4.31 dB and in SDR of 4.79 dB over the method using TarCVAE and IntCVAE ($p < 0.05$). These results demonstrate the advantage of introducing GST into the source model for noisy interference mixtures in GIntCVAE and its effectiveness in noisy underdetermined conditions with a dual-channel TSE system.

4.5 Summary of Chapter 4

This chapter proposed a method for modeling noisy mixed speech by integrating GST into CVAE to enhance TSE performance in noisy underdetermined conditions. The GST is jointly trained with CVAE to form a new source model, GIntCVAE, in which GST generates a latent representation of the noisy mixed speech to serve as a conditional variable for the CVAE. The key features are as follows:(1) Introducing a GST into the CVAE source model enhances the GSS framework-based TSE method in noisy undetermined conditions. (2) The GST in GIntCVAE effectively learns the mixing conditions of the interference mixture in noisy mixed speech, especially the noise level. Experimental results revealed that the proposed method achieved better performance than the baseline methods in noisy underdetermined con-

ditions.

Chapter 5

Neural Postfilter and Enhanced New Target Source Model for Enhancing the Extracted Target with Residual Noise

5.1 Overview

For a TSE problem in 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 Chapter 3 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, in Chapter 4, 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 our problem formulation, 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 diffuse 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 it is 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 diffuse noise on channel 1. In this chapter, a complex T-F mask estimation network was adopted as the neural postfilter, and a new GTarCVAE source model with the neural postfilter was jointly trained.

5.2 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 [44]. CNNBLSTM is a widely used architecture for complex T-F mask estimation in DNN-based speech enhancement tasks [76]. Recent studies, such as [45], have demonstrated its effectiveness in generating cIRM for speech enhancement [46]. 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 section, this CNNBLSTM-based T-F mask estimation network is adopted as the neural postfilter.

Here is a brief review of the neural postfilter based on our problem formulation in Sect. 3.2.1. $\hat{s}_1(f,n)$ is the initial extracted target signal on channel 1, which contains residual interferences in channel 1. Then, the neural postfilter can be represented as

$$\hat{s}_{tar}(f,n) = \mathcal{M}(\hat{s}_1(f,n);\beta)\hat{s}_1(f,n), \tag{5.1}$$

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.

5.3 Joint network of neural postfilter and TarCVAE

For the source model on channel 1, previous chapters 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, this section proposes a joint trained network of GTarCVAE and the neural postfilter with

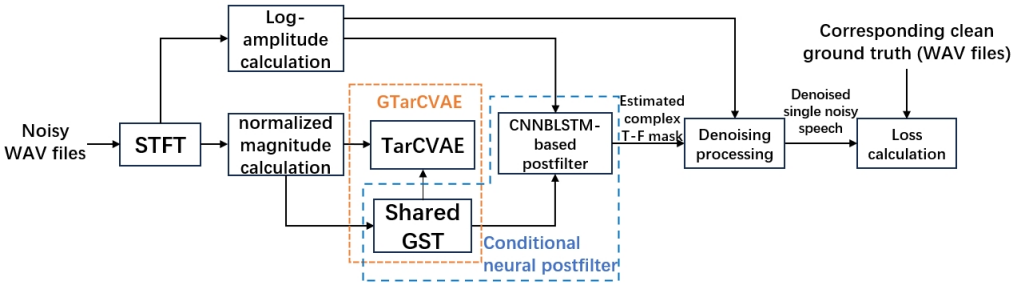


Figure 5.1: Illustration of joint training of GTarCVAE and postfilter.

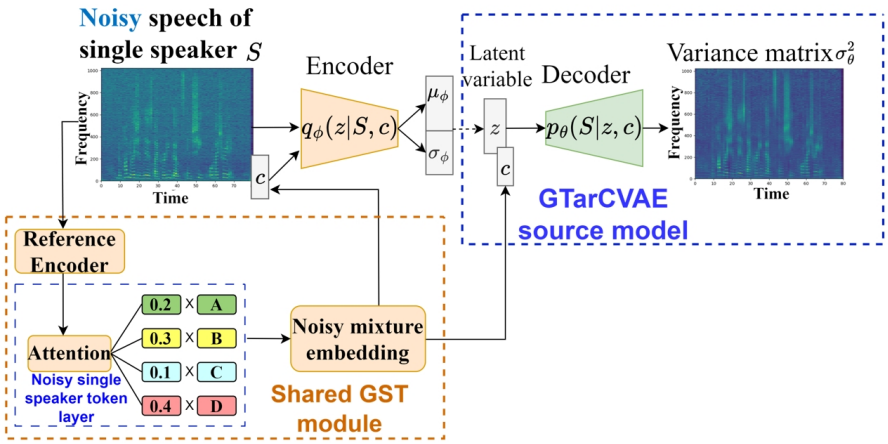


Figure 5.2: Illustration of GTarCVAE.

a shared GST module.

Figure 5.1 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 5.2. Similar to GIntCVAE in Sect. 4.2, 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 [44], and it will be described in detail later.

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 [44], the format is the log-amplitude spectrogram. Therefore, two different calculation processes were added to obtain two types of training dataset from the same data source. In addition, for the neural postfilter, the clean-target training strategy was adopted, where the clean ground truth is required to calculate the loss in the training. Following the loss function Eq. 2.28 of the CVAE, it is assumed 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

$$\begin{aligned}\mathcal{J}(\phi,\theta,\beta) = \mathbb{E}_{(\boldsymbol{S},\boldsymbol{c})\sim p_D(\boldsymbol{S},\boldsymbol{c})}[\mathbb{E}_{z\sim q_\phi(z|\boldsymbol{S},\boldsymbol{c})}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},\boldsymbol{c})] \\ - \mathrm{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{S},\boldsymbol{c})||p(\boldsymbol{z})]] - \mathcal{D}[\mathcal{M}(\boldsymbol{S}|\beta)\boldsymbol{S},\hat{\boldsymbol{S}}],\end{aligned} \tag{5.2}$$

where $\mathcal{D}$ is the function that measures the difference between the denoised signal and the clean ground truth. Here, this study followed [45] to set $\mathcal{D}$ as the mean-squared-error (MSE).

5.4 Inference processing of the new proposed method

In the inference stage, the decoder of the trained GTarCVAE and GIntC-VAE serves as the source model on channels 1 and 2, where the source model parameters of the noisy target speaker and noisy interference mixed speak-

ers, the weight sum of GSTs tokens, and the demixing matrix are updated as iteratively as the algorithm described in Sect. 4.3. Then, the trained neural postfilter is used to denoise the initially extracted target speaker $\hat{s}_1(f,n)$ on channel 1. 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 $\hat{s}_1(f,n)$.

5.5 Experimental evaluations

5.5.1 Training setting of CVAEs and neural postfilter

In the experiments, the joint network of GTarCVAE and the neural postfilter were trained on 25 hrs of noisy audio data. The clean source data was obtained from the si_tr_s folder of the WSJ0 corpus [69], which includes recordings from 101 speakers (50 male and 51 female), each contributing 141 sentences. The clean speech was mixed with four types of diffuse noise at varying SNR levels from the DEMAND dataset [74], which contains six types of diffuse noise. GIntCVAE was the same as Sect. 4.4.1, which 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 [39],

which has been described in detail in Sect. 3.4.1. For the GST module, the architecture follows [73], which has been also described in detail in Sect. 4.4.1. The neural postfilter employs the CNNBLSTM architecture as described in [44]. 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 the 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 complex-valued T-F mask.

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

5.5.2 Evaluation of TSE with neural postfilter in noisy underdetermined conditions

Evaluation conditions

This evaluation assessed the proposed method and baselines on three-source mixtures with a fixed reverberation time of 150 ms. Using the ISM, two-channel recordings of three speakers with noise in a simulated room were

synthesized, as shown in Fig. 3.4 in Sect. 3.4.3. 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. 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. Evaluations were carried out 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 SDR, SIR, SAR, and SNR.

This evaluation selected several related methods as the baselines, including GCIVA [31], MVAE [39], and IntCVAE. For the ablation study on different components of the proposed source models incorporating GSTs, a joint network of GST and TarCVAE was trained without the neural postfilter. Two systems were evaluated: TarCVAE + GIntCVAE and GTarCVAE + GIntCVAE, both using an IRM as the postfilter. To assess the new neural postfilter, an independent CNNBLSTM-based neural postfilter without a GST and a joint network of GST and CNNBLSTM without TarCVAE were trained, called the GST-Neural postfilter. Additionally, this section evaluated GTarCVAE + GIntCVAE + Neural postfilter, GTarCVAE + GIntCVAE + GST-Neural postfilter, and the joint network of GTarCVAE-Neural postfilter + GIntCVAE. The categorization of each baseline and the proposed method is summarized in Table 5.1.

Table 5.1: Comparison between baselines and proposed methods.

Method	Application scenario	Source model	Postfilter
GCIVA [31]	Determined	Laplace	N/A
MVAE [39]	Determined	TarCVAE	N/A
Previous method 1 in Chapter 3	Clean underdetermined	TarCVAE +IntCVAE	IRM
Previous method 2 in Chapter 4	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

Evaluation results

Tables 5.2, 5.3, and 5.4 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 achieve 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 IRM, and the joint network approach is effective in noisy underdetermined conditions.

Table 5.2: Average SDR, SIR, SAR, and SNR [dB] of the extracted target in noisy underdetermined environment of SNR = −10 dB.

Method	SIR	SDR	SAR	SNR
GCIVA [31]	−3.68	−8.25	−6.39	−5.31
MVAE [39]	−3.12	−7.12	−2.44	−4.78
Previous method 1 in Chapter 3	1.03	−3.24	1.25	−2.23
Previous method 2 in Chapter 4	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 5.3: Average SDR, SIR, SAR, and SNR [dB] of the extracted target in noisy underdetermined environment of SNR = 10 dB.

Method	SIR	SDR	SAR	SNR
GCIVA [31]	2.41	1.78	5.06	1.08
MVAE [39]	4.68	2.14	6.59	2.13
Previous method 1 in Chapter 3	8.76	4.43	6.15	4.37
Previous method 2 in Chapter 4	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 5.4: Average SDR, SIR, SAR, and SNR [dB] of the extracted target in noisy underdetermined environment of SNR = 30 dB.

Method	SIR	SDR	SAR	SNR
GCIVA [31]	6.53	4.38	9.38	4.53
MVAE [39]	9.48	7.18	10.47	5.82
Previous method 1 in Chapter 3	15.11	9.61	11.19	8.89
Previous method 2 in Chapter 4	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**

5.6 Summary of Chapter 5

This chapter proposed an improved TSE method for enhancing the extracted target signal in noisy underdetermined conditions and focuses still on the dual-channel GSS framework using the minimum number of microphones to leverage spatial information. Although the GIntCVAE has shown its power in modeling the noisy mixed speech signal, the residual noise in the extracted target signal is a remaining issue in noisy underdetermined conditions. To better model the target signal with noise, this chapter incorporates the GST module into the TarCVAE source model, creating GTarCVAE. To solve the residual noise, this chapter introduces a CNNBLSTM-based neural postfilter to address residual diffuse noise in the extracted target signal. A joint network of GTarCVAE and the neural postfilter with a shared GST module was trained.

The experimental results highlight several points: (1) Introducing a GST into the CVAE source model enhances the GSS framework-based TSE method in noisy undetermined conditions. (2) The CNNBLSTM-based neural postfilter more effectively enhances the extracted target signal with residual noise than the traditional IRM. (3) The joint network of GTarCVAE and the neural postfilter performs well in 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, this research will further investigate its application to real recorded signals and develop an online algorithm.

Chapter 6

Conclusion

6.1 Summary

Target speaker extraction (TSE) in noisy underdetermined environments poses significant challenges due to multi interference speakers and background noise with the limited number of microphones. This dissertation aims to address these challenges by developing a framework that leverages spatial information through geometric source separation (GSS) to achieve dual-channel TSE in noisy underdetermined conditions. By progressively improving source models and postfiltering techniques, this work demonstrates an effective approach to TSE under complex conditions.

The core philosophy of this research lies in combining spatial acoustic knowledge with deep source models to leverage their complementary strengths. Spatial information, specifically direction of arrival (DOA), provided important cues for target extraction. However, conventional methods struggle with robustness in underdetermined conditions or DOA errors. In

Chapter 3, to address these limitations, conditional variational autoencoder (CVAE)-based source models were introduced. The rationale for selecting the CVAE source model stems from its ability to handle the intricate variations in complex acoustic environments. The interference CVAE (IntCVAE) was proposed. Unlike traditional approaches that rely on fixed or limited assumptions about source characteristics, target CVAE (TarCVAE) from the multichannel variational autoencoder (MVAE) method and the proposed IntCVAE provided a flexible and probabilistic framework for modeling both target and interference mixture. This flexibility is particularly necessary in underdetermined conditions. Similarly, DOA modification was introduced as a necessary enhancement to mitigate the sensitivity of spatial filtering to DOA estimation errors. In real-world applications, precise DOA estimation was often challenging due to noise, reverberation, or dynamic speaker movements. By modifying the given DOA to reduce biases toward interference sources, the proposed method achieved greater robustness, ensuring that the extracted signal remains closer to the desired target.

In Chapter 4, to extend this approach to noisy underdetermined environments, the limitations of the discrete conditioning method in IntCVAE were addressed. Global style token (GST) was incorporated into the source model, resulting in the GST-IntCVAE (GIntCVAE) model. The use of GST is based on its ability to provide continuous conditioning representations, which better reflect the variability of acoustic conditions, like noise levels and number of speakers. This flexibility enables the model to adapt more effectively to different noise levels in mixed speech, addressing the rigidity of discrete conditioning. Experimental results confirmed that GIntCVAE im-

proves robustness in noisy underdetermined conditions, establishing a better foundation for TSE under realistic acoustic challenges.

Despite these improvements, residual noise remains a challenge due to the limitations of the TarCVAE source model and the traditional ideal ratio mask (IRM)-based postfilter. To address this, Chapter 5 introduced two key advancements: GTarCVAE and a convolutional neural network-bidirectional long short-term memory (CNNBLSTM)-based neural postfilter. The CNNBLSTM-based postfilter was chosen for its superior capability to capture both temporal and spatial dependencies, enabling more precise suppression of residual noise. By jointly training GST-TarCVAE (GTarCVAE) with the neural postfilter to estimate a complex ideal ratio mask (cIRM), this approach ensures not only a reduction in noise but also the preservation of speech quality, making it a robust solution for extracting target speech in noisy and complex environments.

Overall, this dissertation presented a dual-channel TSE framework addressing the challenges of underdetermined and noisy environments. The philosophy guiding this work is the integration of advanced source modeling and spatial information through GSS, coupled with the introduction of robust postfiltering techniques. This dissertation provided a robust framework for extracting target speech from complex acoustic mixtures. Experimental results throughout the dissertation validated the effectiveness of these methods, highlighting their potential for applications in speech processing systems.

6.2 Future works

While the methods proposed in this dissertation represent a dual-channel TSE method in underdetermined and noisy conditions, several areas for further research and enhancement remain.

Firstly, all current experiments were conducted using simulated signals, which may differ in acoustic characteristics and spatial information from real-world environments. Future research can try to adapt and optimize the proposed models for real-world acoustic scenarios to ensure practical applicability.

Additionally, the proposed method is currently computationally intensive. Efforts to improve processing efficiency, potentially through streamlined architectures or advanced optimization techniques, would be beneficial for enabling faster running time.

Another area of interest is to further test and strengthen the method under more challenging conditions with both high noise levels and strong reverberation, which are common in many real-world situations. Enhancing the model's robustness under these extreme conditions would extend its applicability and reliability in diverse settings.

The current approach assumes the location of the target speaker is known in advance. Future research could explore the integration of sound source localization techniques, allowing the TSE framework to operate effectively without prior knowledge of the target speaker's position. Furthermore, existing methods assume the target speaker remains in a fixed position, but real-world scenarios often involve dynamic speakers. Extending the frame-

work to track and extract the target speaker in real-time as their position changes is another interesting research direction.

In addition, the current cue for target selection relies on spatial information. Other potential cues, such as temporal information (e.g., voice activity detection), spectral characteristics (e.g., voice quality), visual cues (e.g., lip movements), and others, could also be explored for their applicability.

Extending the proposed methods to scenarios involving more than three speakers would also be highly valuable. As the number of speakers increases, the complexity of source separation intensifies. Developing solutions that maintain performance in densely populated acoustic environments would expand the scope and utility of this work.

Bibliography

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

[2] 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.

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

[4] 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.

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

[6] 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.

[7] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, Vol. 2003, No. 11, pp. 1135–1146, Nov. 2003.

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

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

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

[11] 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, Oct. 2001.

[12] 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.

[13] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combin-

ing ICA and beamforming," *IEEE Trans. Audio, Speech, Lang. Process*, Vol. 14, No. 2, pp. 666–678, Mar. 2006.

[14] 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*, pp. 3247–3250, May. 2007.

[15] 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*, pp. 165–172, 2006.

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

[17] 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*, pp. 5554–5558, Apr. 2018.

[18] Q. Wang et al., "Voicefilter: Targeted voice separation by speakerconditioned spectrogram masking," in *Proc. INTERSPEECH*, pp. 2728–2732, Sep. 2019.

[19] 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*, pp. 1406–1410, Oct. 2020.

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

[21] 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.

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

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

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

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

[26] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 6, pp. 352–362, 2002.

[27] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, mo. 2, pp. 715–726, 2007.

[28] 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 (ICASSP)*, IEEE, pp. 3065–3068, Apr. 2009.

[29] K. Reindl, S. Meier, H. Barfuss, and W. Kellermann, "Minimum mutual information-based linearly constrained broadband signal extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, No. 6, pp. 1096–1108, Juen. 2014.

[30] H. Barfuss, K. Reindl, and W. Kellermann, "Informed Spatial Filtering Based on Constrained Independent Component Analysis," *Audio Source Separation*, pp. 237–278, 2018.

[31] 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. pp. 846–850, May. 2020.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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, June. 2020.

[37] 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*, Vol. 24, No. 9, pp. 1652–1664, Sep. 2016.

[38] 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, Vol. 27, No. 10, pp. 1601–1615, Oct. 2019.

[39] 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.

[40] 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*, pp. 3581–3589, 2014.

[41] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, Vol. 14, No. 4, pp. 1462-1469, 2006.

[42] Y. Zou, Z. Liu, and C. H. Ritz, "Enhancing target speech based on nonlinear soft masking using a single acoustic vector sensor," *Applied Sciences*, Vol. 8, No. 9, pp. 1436–1452, 2018.

[43] 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.

[44] M. Kawanaka, Y. Koizumi, R. Miyazaki, and K. Yatabe, "Stable training of DNN for speech enhancement based on perceptuallymotivated black-box cost function," In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7524–7528, May. 2020.

[45] T. Fujimura and T. Toda, "Analysis of Noisy-target Training for DNN-based speech enhancement," In: *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5 pages, June. 2023.

[46] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol 24, No. 3, pp. 483–492, 2015.

[47] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE transactions on audio, speech, and language processing*, Vol. 15, No. 1, pp. 70–79, 2006.

[48] J. Bourgeois and W.Minker (Eds.), *Linearly Constrained Minimum Variance Beamforming*, Springer, Boston, pp. 27–38, 2009.

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

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

[51] D. R. Hunter and K. Lange, "A tutorial on mm algorithms," *The American Statistician*, Vol. 58, No. 1, pp. 30–37, 2004.

[52] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *2011 IEEE Workshop on*

Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, pp. 189–192, 2011.

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

[54] 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.

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

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

[57] 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. ICASSP*, pp. 716–720, Apr. 2018.

[58] 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. for Signal Process. (MLSP), 6 pages, Sep. 2018.

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

[60] E. Karamatli, A. T. Cemgil, and S. Kırbız, "Audio source separation using variational autoencoders and weak class supervision," *IEEE Signal Processing Letters*, Vol. 26, No. 9, pp. 1349–1353, 2019.

[61] 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.

[62] S. Sivasankaran, E. Vincent, and D. Fohr, "Analyzing the impact of speaker localization errors on speech separation for automatic speech recognition," *arXiv preprint*, arXiv:1910.11114, 2019.

[63] 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, Language Process*, Vol. 26, No. 1, pp. 145–160, 2018.

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

[65] 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.

[66] 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.

[67] 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.

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

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

[70] 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.

[71] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. Interspeech*, Graz, Austria, pp. 4290–4294, Sep. 2019.

[72] L. Chen, M. Yu, D. Su, and D. Yu, "Multi-band pit and model integration for improved multi-channel speech separation," in in Proc. *ICASSP*, pp. 705–709, 2019.

[73] 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*, pp. 5180–5189, 2018.

[74] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," Supported by Inria under the Associate Team Program VERSAMUS, June. 2013.

[75] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, Vol. 9, No. 11, pp. 2579–2605, 2008.

[76] H. Erdogan and T. Yoshioka, "Investigations on data augmentation and loss functions for deep learning based speech background separation," *Proc. of Interspeech*, pp. 3499–3503, Sep. 2018.

Acknowledgement

Time flies. In the twinkling of an eye, I have spent almost four years in Nagoya. In April 2021, I had the honor of becoming a doctoral student at Toda laboratory of Nagoya University. As an excellent university, Nagoya University has a beautiful environment and strong scientific research strength. Looking back on the previous four years, there are many good memories. As graduation approaches, there are too many emotions. Here, I have too much to thank for.

First, I would like to thank my dissertation advisor: Prof. Toda for his patient guidance and help, which enabled me to obtain valuable knowledge and wonderful scientific research experience. I am also very grateful to Prof. Toda for guiding me in the right research direction and providing me with tremendous academic support, allowing me to have many opportunities for international exchange. I also want to thank Prof Toda's concern for me, as an international student, and his outstanding knowledge and personality have greatly benefited me, making him a role model for my lifelong learning.

At the same time, I would like to thank Dr. Li li and Fujimura san for their generous help. Especially for Dr. Li li, she provided me with tremendous help during my early exploration of the research topic and was

a leader on my research path. I would like to express my gratitude to my colleagues in the laboratory for their help.

Besides, I would also like to express my heartfelt gratitude to Ms. Noro, the lab secretary, for her constant help and care in my daily life. Mrs. Noro has been so kind and considerate, like a gentle older sister, assisting me with many university-related matters and allowing me to focus on my research.

In addition, I would like to express my sincere gratitude to my closest friends, including, but not limited to, Shaowen, Shuming, Jiajun, Martin, Xiaohan, and Brother Li from our lab, as well as Dr. Hu, who has been a steadfast companion since my master's days. During these years away from home, I feel truly fortunate to have had you by my side. We have shared many memorable moments—cycling, traveling, and creating lasting memories. You have been like family to me here in Japan, offering mutual support and companionship as we traveled together. I hope we can continue this friendship for years to come.

Finally, I want to thank the most important person in my life—my mother. Six and a half years ago, when I expressed my desire to study abroad, it was my mother who unwaveringly supported me. Throughout these years overseas, on every lonely and helpless night, no matter how challenging things became, hearing my mother's voice over the phone gave me the strength to keep going; she has been my sole source of emotional support during my doctoral journey. Growing up in a single-parent family, I carry a deep sense of guilt for the hardships my mother has endured. The challenges and struggles I've faced in my academic path pale in comparison to all that she has borne over the past three decades. My mother's love and sacrifices

are boundless, and I don't know if I will ever be able to repay her kindness in this lifetime. Now, as I prepare to return home, I finally have the chance to be by her side. I hope that in the years to come, my mother remains healthy, happy, and at peace.

Publications

Journal papers

1. R. Wang, L. Li, T. Toda, "Dual-channel target speaker extraction based on conditional variational autoencoder and directional information," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 32, pp. 1968–1979, Mar. 2024.

2. R. Wang, T. Fujimura, T. Toda, "Target speaker extraction under noisy underdetermined conditions using conditional variational autoencoder, global style token, and neural postfilter," *APSIPA Transactions on Signal and Information Processing*, 2024 (Accepted).

3. R. Wang, B. N. Khanh, D. Morikawa, and M. Unoki, "Method of estimating three dimensional direction-of-arrival based on monaural modulation spectrum," *Applied Acoustics*, 203, 109215, 9 pages, Feb. 2023.

International conferences

1. 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, pp. 347–354, 2022.

2. 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, 5 pages, Oct. 2023.

3. N. Li, L. Wang, M. Unoki, S. Li, R. Wang, Me. Ge, J. Dang, "Robust Voice Activity Detection Using a Masked Auditory Encoder Based Convolutional Neural Network," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6828 - 6832, Jun. 2021.

4. R. Wang, B. N. Khanh, D. Morikawa, and M. Unoki, "Method of Estimating 3D DOA based on Monaural Modulation Spectrum," In: *2021 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP)*, pp. 137–140, Mar. 2021.

Technical report

R. Wang, L. Li, and T. Toda, “Target speaker extraction based on conditional variational autoencoder and directional information in underdetermined condition,” *Technical Report of IEICE*, Vol. 121, No. 383, EA. 2021-76, pp. 76–81, Mar. 2022.

Domestic conference

1. R. Wang, B. N. Khanh, D. Morikawa, and M. Unoki, “Method of estimating DOA based on monaural modulation spectrum,” 日本音響学会春季研究発表会講演論文集, 3-1-21, pp. 321-324, Mar. 2021.

2. R. Wang, Li Li, and T. Toda, “Direction-aware target speaker extraction with conditional variational autoencoders and its sensitivity to direction-of-arrival error,” 日本音響学会春季研究発表会講演論文集, 2-2-6.

Awards

1. 2021 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP), Student paper award.

2. 日本音響学会北陸支部・2021 優秀学生賞.

3. 日本音響学会第 25 回学生優秀発表賞.

4. IEEE WASPAA 2023 Travel Grants.