A Speaker-Dependent Approach to Separation of Far-Field Multi-Talker Microphone Array Speech for Front-End Processing in the CHiME-5 Challenge

Lei Sun, Jun Du, Tian Gao, Yi Fang, Feng Ma, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—We propose a novel speaker-dependent speech separation framework for the challenging CHiME-5 acoustic environments, exploiting advantages of both deep learning based and conventional preprocessing techniques to prepare data effectively for separating target speech from multi-talker mixed speech collected with multiple microphone arrays. First, a series of multi-channel operations is conducted to reduce existing reverberation and noise, and a single-channel deep learning based speech enhancement model is used to predict speech presence probabilities. Next, a two-stage supervised speech separation approach, using oracle speaker diarization information from CHiME-5, is proposed to separate speech of a target speaker from interference speakers in mixed speech. Given a set of three estimated masks of the background noise, the target speaker and the interference speakers from single-channel speech enhancement and separation models, a complex Gaussian mixture model based generalized eigenvalue beamformer is then used for enhancing the signal at the reference array while avoiding the speaker permutation issue. Furthermore, the proposed front-end can generate a large variety of processed data for an ensemble of speech recognition results. Experiments on the development set have shown that the proposed two-stage approach can yield significant improvements of recognition performance over the official baseline system and achieved top accuracies in all four competing evaluation categories among all systems submitted to the CHiME-5 Challenge.**

***Index Terms*—The CHiME-5 challenge, speech enhancement, speech separation, mask estimation, robust speech recognition.**

I. Introduction

In RECENT decades, automatic speech recognition (ASR) has greatly developed [1] in terms of both tasks and technologies. Various limited tasks were investigated during the 70s and 80s, such as the recognition of connected digit sequences using the TIdigits corpus [2]. A widely used corpus, TIMIT [3] has provided data for acoustic-phonetic studies since the 90s. The emergence of the Wall Street Journal (WSJ) corpus [4] and LibriSpeech [5] corpus have been used for the research of large vocabulary speech recognition. To improve ASR robustness, people have studied the recognition under noisy or reverberant conditions. The AMI corpus [6] was proposed to explore both close-talking and far-field speech recognition under realistic meeting conditions. In addition, several academic challenges have been held to find solutions to different problems. For example, the REVERB challenge [7] provided an opportunity to study reverberant speech, while the CHiME (1-4) [8]–[10] series investigated the effects of background noise. Apart from those environmental interferences, overlapping speech and quick speaker transitions also greatly degrade recognition accuracy in real situations. The 2006 Speech Separation Challenge (SSC) [11] aimed to explore source separation methods, which was helpful to solve speech recognition tasks with two simultaneous speakers. From the perspectives of both academic speech corpora and earlier research challenges, robust ASR research has undergone the following process: from single-channel to multi-channel, from single-array to multi-array, and from simulated data to real-recorded data. These endeavors have been dedicated to promoting the development of more advanced speech recognition systems.

Meanwhile, the front-end and back-end processing methods have also made great progress. Specifically, the goal of a front-end system is to remove or suppress interference factors such as noise and reverberation while maintaining the integrity and intelligibility of the original speech. It can be divided into subdomains with different research priorities, such as speech enhancement, speech separation, speech dereverberation, etc. The most direct and intuitive way to evaluate a front-end processing system is to observe the performance changes of the back-end performance, which is devoted to acoustic modeling and language modeling. Depending on the microphone settings, the front-end approaches can be categorised into single-channel based methods and multi-channel based methods.

Single-channel speech processing is widely needed in many situations where only one microphone is available to capture the signal, such as voice communication systems, personal smart assistants, etc. Traditional single-channel speech enhancement algorithms include spectral subtraction [12], Wiener filtering [13]

Manuscript received November 19, 2018; revised March 3, 2019 and May 7, 2019; accepted May 24, 2019. Date of publication June 3, 2019; date of current version July 25, 2019. This work was supported in part by the National Key R&D Program of China under Contract 2017YFB1002202, in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, in part by the Key Science and Technology Project of Anhui Province under Grant 17030901005, and in part by Tencent. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Michiel Bacchiani. *(Corresponding author: Jun Du.)*

L. Sun and J. Du are with the University of Science and Technology of China, Hefei 230052, China (e-mail: sunlei17@mail.ustc.edu.cn; jundu@ustc.edu.cn).

T. Gao, Y. Fang, and F. Ma are with iFlytek, Hefei 230088, China (e-mail: gtian09@mail.ustc.edu.cn; yifang2@iflytek.com; fengma@iflytek.com).

C.-H. Lee is with the Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: chl@ece.gatech.edu).

Digital Object Identifier 10.1109/JSTSP.2019.2920764

and the minimum mean square error (MMSE) based spectral amplitude estimator [14]. These unsupervised methods are based on many mathematical assumptions about speech and noise, and often introduce artifacts (e.g., musical noise). With the development of machine learning techniques, supervised approaches have shown great abilities to improve the quality of enhanced speech due to large amounts of training data. Nonnegative matrix factorization (NMF) [15] is one typical method where speech and noise bases are obtained separately from training data. However, the relationship between these two classes cannot be effectively learned. Recently, deep neural networks (DNNs) have also demonstrated powerful modeling capability and achieved better performance than traditional methods in the field of speech enhancement [16]–[19]. However, the background noise is not always the exclusive interference factor in some multi-talker situations, where one person is saying when others are speaking at the same time. Such kind of issues were associated with the term "cocktail-party problem" in [20]. Many studies have explored single-channel speech separation methods to segregate the voices of different speakers. For example, deep clustering (DC) [21] and the attractor network (DANet) [22] focused on finding a substantial embedding space of mixture signals, assuming that time-frequency (t-f) units belong to the same speaker can form a cluster. The permutation problem [23], also referred to as the label ambiguity problem in [21], [24], is a difficult problem especially for speaker-independent multi-talker speech separation. Accordingly, permutation invariant training (PIT)-based methods [23], [25] were proposed to address such problems by ignoring the order of mixing sources during the optimization process.

Unlike single-channel data, multi-channel data are acquired by several microphones which are spatially distributed in the physical space. Multi-channel speech processing was initially used to perform sound localization and speech enhancement in chaotic environments [26]. With the emergence of independent component analysis (ICA) [27], which established an instantaneous linear mixing model, blind source separation (BSS) has been proposed to extract individual signals from observed mixed signals [28]. Typical algorithms consist of multi-channel Wiener filtering [29], blind source separation methods [27], [30]–[32], and beamforming methods [33]–[36]. It has been proved that beamforming based methods can both effectively improve the output signal-to-noise ratio (SNR) [37] and enhance the speech recognition performance [38], [39]. Furthermore, some studies have focused on fully utilizing the advantages of deep learning methods and conventional multi-channel methods. In [40], bidirectional long short-term memory (BLSTM) was adopted to estimate signal statistics to steer the beamformer for multi-channel speech enhancement. In [41], a beamforming approach via an iterative mask estimation (IME) that combined CGMM [36] and a deep learning model demonstrated an extremely low word error rate (WER) on CHiME-4 data.

Recently, the latest CHiME-5 challenge provides the first large-scale corpus of real multi-talker conversational speech recorded via commercially available microphone arrays in multiple realistic homes [42]. This corpus essentially congregates a large number of acoustic problems that may exist in real life, which poses a great challenge to existing ASR systems, especially for the front-end processing in the case of noise, reverberation, overlapping speech. New technologies or solutions are needed to simultaneously address these interference factors that were not considered or handled well by previous methods. The WERs of the binaural microphone data and the single-array data in the official baseline report are 47.9% and 81.3% respectively, which fully illustrate the difficulty of the CHiME-5 ASR task.

In this study, we propose a novel speaker-dependent speech separation framework for the challenging CHiME-5 acoustic environments, exploiting advantages of both deep learning based methods and conventional preprocessing techniques to prepare data effectively for separating target speech from multi-talker mixed speech collected with multiple microphone arrays. The contributions are summarized as follows. First, we carefully analyze the CHiME-5 data and find the reason for the poor performance of ASR is that the realistic multi-talker conversations introduce rapid role conversion and overlapping speech. Second, we design a two-stage speaker-dependent speech separation approach to extract the speech of the target speaker in each recording, especially for overlapping data. The method takes the privilege of utilizing oracle speaker diarization information, allowed by the challenge rules. With different learning targets, the sequential two separation stages can overcome the low-resource problem of CHiME-5 data and yield better localization ability of the target speaker. Similarly, a speech enhancement model is used to estimate the speech or noise existing probabilities. Third, given the information derived from deep models, a CGMM based generalized eigenvalue (GEV) beamformer [37] is then used for enhancing the signal at the reference array while avoiding the speaker permutation problem. Finally, the extracted speech of only the target speaker is directly fed into the back-end recognition system. Different front-end models generated with various configurations provide multiple data streams for the ensemble ASR system to improve and stabilize the overall performance.

The remainder of this paper is organized as follows. Section II presents the background information and data analysis of the CHiME-5 challenge. In Section III, we describe the design of our proposed front-end system, including multi-channel speech preprocessing, single-channel speech enhancement, single-channel speaker-dependent speech separation and multi-channel beamforming. Section IV lists the experimental setups. Section V shows the comprehensive results and analyses about how each technique affects the final performance according to ASR results. Section VI discusses the gap between a practical ASR system and the proposed system in the CHiME-5 challenge. In Section VII, we summarize our findings and plan the future work.

II. Analysis of the CHiME-5 Dinner Party Scenario

We initially perform comprehensive analyses of the dinner party scenario in the CHiME-5 challenge. Also, studying the limitations of existing technologies is of great significance to guide our front-end design.

TABLE I
OVERVIEW OF CHIME-5 CORPUS

Data	Training	Development	Evaluation
Sessions	16	2	2
Speakers	32	8	8
Hours	40:33	4:27	5:12
Utterances	79,980	7,440	11,028
Binaural recordings	✓	✓	×
Speaker labels	✓		
Time stamps	✓		

*Note that each 4-speaker group participated in 2 sessions in the training set. There were no duplicated speakers between sessions from both development and evaluation sets.

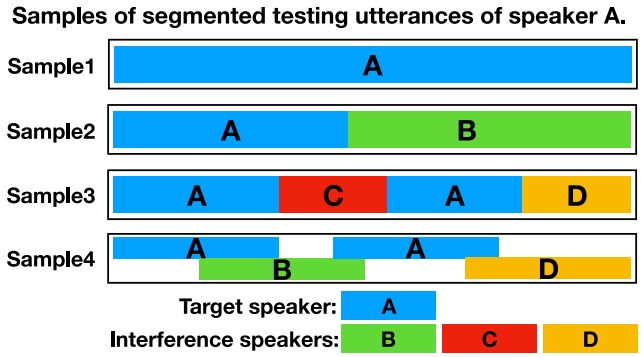


Fig. 1. An illustration of speech samples from one session. Speaker A is set to be the target speaker, while the others are the interference speakers. Those samples are excised from the whole session recording according to human transcriptions. As seen here, other interference speakers appear in a random manner.

A. Background

Speech data in the CHiME-5 were made under a 4-person dinner party scenario and recorded by 6 distant Kinect microphone arrays and 4 binaural microphone pairs in 20 homes [42]. The detailed transcriptions of each speaker were obtained by the annotator through listening to the corresponding binaural recordings, including not only the word sequences but also time stamps of every speaker, which can be seen as the oracle speaker diarization information. There were no restrictions on the content or speaking style for all participants. Each party was held for a minimum of 2 hours and composed of three phases representing different locations, namely 'kitchen', 'dining room' and 'living room'. The binaural data were provided in addition to the far-field array data for both training and development sets. Ultimately, the 20 home sessions were divided into training, development and evaluation sets. According to the challenge rules, we summarize the necessary information in Table I. Specifically, using oracle speaker diarization information was allowed in all sets, though it's not available in a practical speech recognition system.

B. Data Analysis

Through deep analysis of the CHiME-5 speech, the biggest difficulties can be summarized into two aspects. The first one is far-field speech recognition, which is largely affected by background noise, reverberation and distant moving speakers. Unlike the fixed noise categories used in CHiME-3 and CHiME-4, CHiME-5 data contains more noise types that may exist in real life, such as stationary noises from air conditioning, and non-stationary noises from cooking, human movements, etc. In addition, the reverberation also varies among different sessions with changing rooms, speaker locations, and array positions. Moreover, people were allowed to move naturally in these places. The above set of issues posed a big challenge for research into far-field speech processing. The second aspect, which is the most challenging part of CHiME-5, is about the dialogue style. Unlike read speech, the complexity of conversational and spontaneous speech greatly increases the difficulty of a speech recognition system. For instance, casual pronunciation and frequent overlapping speech seriously decrease the discriminating ability of acoustic models. In order to prove the summary of these two aspects, a direct proof is that the WER of relatively clean speech taken from the binaural microphones is already as low as 47.9% in a baseline recognition system [10]. It's largely due to the conversational style. Then, the WER drastically increases up to 81.3% for far-field ASR with single-array data, that can be attributed to the distance between the source and the microphones where noise and reverberation are introduced. In summary, the challenge of CHiME-5 can be simplified to the problem of conversational speech recognition in far-field, multi-talker conditions.

To investigate the speech overlapping problem, we excluded non-speech regions and aligned the time stamps of all speakers to locate the overlapped speech regions. It was observed that approximately **97.3%** of the sentences from the development set contained overlapping speech. Since human labeling is inherently not entirely accurate, the speech segments of the target speaker often contain interference from other speakers. As shown in Fig. 1, we present some examples of how target speaker's speech interacts with the other speakers'. Specifically, Sample 1 is truly a single-talker segment, while the other samples are with multiple talkers. In Samples 2-4, we can observe inserted speech and even overlapped speech. In such cases, speech introduced from the interference speakers can greatly hurt the recognition performance of the target speaker.

C. Speaker-Dependent Separation Scheme

A straightforward way for the front-end processing is to remove the interfering speech before performing speech recognition, which falls into the field of speech separation, especially for deep learning based single-channel approaches [18], [21], [25], [43], [44]. However, these supervised and data-driven methods require pure source data for data simulation in the training stage, that is, speech without interference speakers. The data distributions of non-overlapping speech are listed in Table II. After eliminating regions of overlapping speech, the remaining data for each speaker are extremely limited. Therefore, the speech separation scheme under such low resource condition is another challenging problem.

To solve these problems, we propose a two-stage speaker-dependent speech separation approach for the CHiME-5 challenge, as described in Section III-C. Before simulating paired

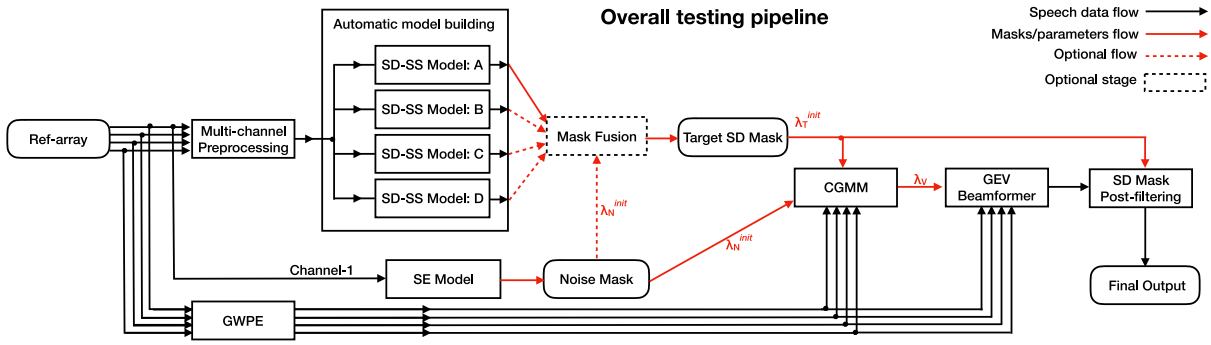


Fig. 2. The overall diagram of our proposed front-end processing system for the CHiME-5 challenge.

TABLE II
DETAILS OF SPEAKER INFORMATION ABOUT THE CHIME-5 DEVELOPMENT SET, ACCORDING TO ORACLE HUMAN TRANSCRIPTIONS

Dev. Set (Session)	Speaker	Gender	Total Duration (minutes)	Non-overlapping (minutes)
S02	P05	Female	66.1	11.0
	P06	Male	70.0	11.9
	P07	Male	47.2	5.6
	P08	Female	59.3	7.3
S09	P25	Female	43.1	9.1
	P26	Female	34.6	6.3
	P27	Female	30.5	6.7
	P28	Female	37.5	9.5

speech mixtures, oracle speaker diarization information is needed to select the non-overlapping data of each speaker as the source materials, while excluding data from other interference speakers. To the best of our knowledge, it is the first time that a supervised speech separation approach is capable of handling such challenging realistic data.

III. System Overview

To fully utilize all information of each speaker, our proposed front-end system is conducted session by session in the testing stage, as illustrated in Fig. 2. It is designed as a tandem system via the integration of single-channel deep learning based and conventional multi-channel techniques to prepare data for effective speech separation (SS). In particular, to extract speech of the target speaker, we propose a two-stage speaker-dependent (SD) model, including SS1 and SS2 to estimate the probability masks of the target speaker at all time-frequency bins. To improve the SD masks, we further use a CGMM to model the spatial information. Finally, a GEV beamformer with single-channel post-filtering is used to generate the front-end output. Detailed descriptions of processing at each stage are given in the following.

A. Multi-Channel Preprocessing

A series of multi-channel operations is first conducted as shown in Fig. 3. At the beginning, we use a generalized weighted prediction error (GWPE) [45] algorithm on the multi-channel signals of the reference array, which is commonly used as a dereverberation preprocessor.

To better estimate and suppress the noise, we design the following two steps. First, we adopt the independent vector analysis (IVA) based blind source separation method [30] to generate independent signals from all channels. Here, the number N of channels and the number M of speakers are both set to 4 when using a single reference array. An auxiliary function [30], [46] is also used for more effective optimization and faster convergence of the IVA method. After estimating the frequency-domain demixing matrix, a back-projection technique [47], [48] is applied to the estimated signals to restore their observed amplitudes. More operational details can be seen in [49].

Second, in the multi-channel post-filtering module, we perform noise suppression on each channel from the IVA output data, and mix all channels together. To start, a minima controlled recursive averaging (MCRA) [50] noise estimator is used to estimate the stationary noise on each channel. For non-stationary noise estimation of a reference channel, the other three channels are averaged after subtracting their own stationary noise. Given both stationary noise and non-stationary noise, the so-called decision directed approach [14] is used to calculate the a priori and a posteriori SNRs, which are used to perform Wiener filtering. Ultimately, four processed channels are mixed together by linear addition to get a single channel signal.

Thus far, we have initially preprocessed the array data, removing the reverberation and improving the signal-to-noise ratio (SNR). Signals with a higher SNR are considered as a prerequisite for the following single-channel supervised learning methods, where large quantities of training data need to be simulated from "theoretically clean data". In simple terms, this section serves as a preprocessing module to relax restrictions, such as no reverberation and no noise. The pipeline of these operations is briefly presented in the time-frequency domain as follows:

$$P(t,f) = [G_1, G_2 ... G_N(t,f)]\boldsymbol{W}_{IVA\_BP}(f)\boldsymbol{X}_{WPE}(t,f) \tag{1}$$

where $\boldsymbol{X}_{WPE}$ represents the multi-channel outputs of GWPE dereverberation, $\boldsymbol{W}_{IVA\_BP}$ represents the time-invariant coefficients achieved by IVA and back-projection, G_N is the weight of each channel in multi-channel noise reduction for which N is known as 4 in advance, and $P(t,f)$ is the final single-channel output of the multi-channel preprocessing stage.

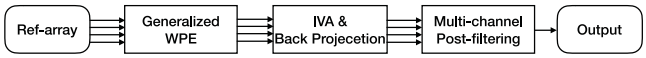


Fig. 3. An illustration of our multi-channel preprocessing procedure to output the single channel enhanced speech.

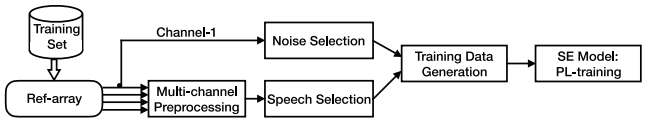


Fig. 4. The flow diagram of building the speech enhancement (SE) model in the training set.

B. Single-Channel Speech Enhancement

Next, we use a deep learning based speech enhancement (SE) model that is robust to the space geometry of the microphone array and non-stationary noise. A strong complementarity was demonstrated between CGMM-based and DNN-based mask estimation for speech enhancement in the best CHiME-4 system [41]. Since the binaural data are not provided in the test set, we decide to consistently use the training data generated by the multi-channel preprocessing stage as our "clean" data to simulate noisy mixtures. A densely connected progressive learning (PL) architecture for LSTM-based speech enhancement [51] is adopted here, where the learning target is replaced by ideal ratio mask (IRM) [52], [53]:

$$z^{\mathrm{IRM}}(t,f) = \frac{S(t,f)}{S(t,f)+N(t,f)} \tag{2}$$

where $S(t,f)$ and $N(t,f)$ represent the power spectra of the speech and noise signals at the time-frequency unit (t,f), respectively. The equation indicates the value of IRM is between 0 and 1, and it is robust as a learning target when the source speech data are not "clean" enough. In the testing phase, the utterances are processed by the speech enhancement model and yield the estimated IRM that represents the speech probability at the time-frequency bin level. When a noise mask needs to be estimated, its value can be represented as $1 - IRM$. To better illustrate the training procedure, we present it in Algorithm 1 and Fig. 4.

C. Single-Channel Speaker-Dependent Speech Separation

As analyzed in Section II-B, how to deal with the overlapping speech is crucial in the CHiME-5 challenge. Previously, much work [21]–[23] focused on single-channel multi-talker speech separation. However, as reported in [54], most of those approaches could not work well on the CHiME-5 data. For example, the environment noise and spontaneous speech severely destroy the embedding process adopted in many approaches such as deep clustering [21] and DANet [22]. Finally, we adopt a speaker-dependent speech separation scheme to effciently use the available source data, which can directly avoid the permutation problem discussed in PIT-based separation methods [23], [25]. Note the proposed method is only conducted within each testing session.

Algorithm 1: Speech Enhancement Model Training.

Step1:**Noise data selection**

1) Use oracle human transcriptions of the training set to select the non-speech segments as noise data from channel-1.
2) To prevent the existence of potential untranscribed speech, use the official baseline speech recognition model to eliminate recognizable segments.

Step2:**Simulation data generation for training SE model**

1) Generate noisy utterances by mixing selected "noise" data and "clean" data processed by a multi-channel preprocessing stage, while retaining the calculated IRM as the learning target.

Step3:**SE model training**

1) Train a deep learning model using the log-power spectral (LPS) features as input features and corresponding IRMs as the output features under the MMSE criterion.

1) First-Stage Speech Separation: As shown in Fig. 5, we first use the non-overlapping part of the multi-channel preprocessed data of each speaker to simulate mixed speech. To build the training set of data pairs of target and mixed speech, the utterances of the target speaker are mixed with speech from interference speakers at several SNR levels. Thus, different speakers can obtain their own training data. In the first-stage, we consider designing an approach to make an aggressive segregation of the target speaker, namely, suppressing the interference speech as much as possible.

Accordingly, we adopt a BLSTM model trained with an intermediate approach named indirect mapping (IM) in [53], which has also been explored in [55], [56]. This method is designed to fully utilize the advantage of both mapping-based and masking-based learning targets [53] of deep models, namely, LPS features [19] and IRM [57]. Specifically, the IM-based approach estimates the IRM by optimizing the BLSTM parameters via the MMSE between the masked and the reference LPS features [55], [56]:

$$E_{\mathrm{IM}} = \sum_{t,f} \left(\log \hat{z}^{\mathrm{IRM}}(t,f) + x^{\mathrm{LPS}}(t,f) - \bar{z}^{\mathrm{LPS}}(t,f)\right)^2 \tag{3}$$

where $\hat{z}^{\mathrm{IRM}}(t,f)$ is the BLSTM estimated IRM that is combined with the logarithm operation and the input noisy LPS features $x^{\mathrm{LPS}}(t,f)$ to generate the masked LPS features. $\bar{z}^{\mathrm{LPS}}(t,f)$ are the reference clean LPS features at the time-frequency unit (t,f). By using E_{IM}, the model can not only suppress the interference speech as much as possible in the manner of mapping-based targets but also yield robust and moderate masks in the manner of masking-based targets. After training, the speaker-dependent speech separation model of every speaker can be generated, which are denoted as **SD-SS1**.

However, useable non-overlapping data of an individual speaker are insufficient, as shown in Table II, especially for speakers such as P07, P08, and P26. Hence, we make full use of the existing SD-SS1 model to expand the useable data size

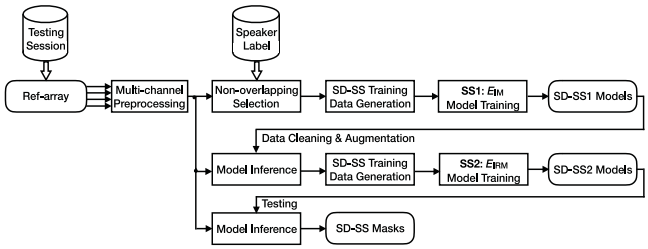


Fig. 5. The flow diagram of building the two-stage speaker-dependent speech separation (SD-SS) models in one test session.

of a specific speaker by suppressing interference speech in all original data without non-overlapping data selection. As a result, both non-overlapping speech and overlapping speech can be seen as containing only target speaker's speech. The speech diversity of each speaker has been greatly enhanced. As seen in Fig. 5, the role of first-stage speech separation is for data cleaning and augmentation.

2) Second-Stage Speech Separation: In the second stage, data separated by the SD-SS1 model are used to simulate new training data with the same number of data pairs as in the first stage. We use **SD-SS2** to denote the separation model of the second stage. In the training phase, we select the original IRM as our training target because it leads to better speech intelligibility and fewer speech distortions. This approach is more appropriate and stable according to the final recognition performance. The optimization function of the SD-SS2 model is defined as:

$$E_{\text{IRM}} = \sum_{t,f} \left(\hat{z}^{\text{IRM}}(t,f) - \bar{z}^{\text{IRM}}(t,f)\right)^2 \tag{4}$$

where $\hat{z}^{\text{IRM}}(t,f)$ and $\bar{z}^{\text{IRM}}(t,f)$ are the BLSTM estimated and the reference IRM, respectively.

In general, the proposed sequential separation stages take full advantage of both learning targets, where E_{IM} aggressively removes the interference speech and augments the data size of usable non-overlapping speech, while E_{IRM} is adopted as a better choice for preserving the speech integrity and intelligibility of the target speaker. Given the oracle speaker diarization information, the proposed speaker-dependent speech separation approach can be conducted in each session. When in the test phase, every utterance will be processed by its corresponding speaker's SD-SS2 model and yield the estimated IRM of the target speaker. To better illustrate the procedure, we present the whole process in Algorithm 2 and Fig. 5.

Algorithm 2: Two-Stage Speaker-Dependent Speech Separation (SD-SS) Model Training.

Step1:**Training the first-stage model: SD-SS1**
1) Use oracle human transcriptions of each session to select the non-overlapping segments of every speaker.
2) Generate mixture utterances by mixing selected segments of the target speaker and interference speakers.
3) Train the BLSTM model with mixture features and target speaker features under the loss function of E_{IM} and obtain the "SD-SS1" model of each speaker in one session.

Step2:**Data cleaning and augmentation**
1) Rather than only using the selected non-overlapping segments, obtain the estimated IRM of all complete sentences by inferring the corresponding "SD-SS1" model.
2) Recover the speech signals from the LPS features masked by the estimated IRM.

Step3:**Training the second-stage model: SD-SS2**
1) Directly generate mixture utterances by mixing complete sentences of the target speaker and interference speakers created from Step2 while retaining the calculated IRM as the learning target.
2) Train the BLSTM model with the mixture features and the IRM under the loss function of E_{IRM} and obtain the "SD-SS2" model of each speaker in one session.

D. Multi-Channel Beamforming

Beamforming has been the most effective method in multi-channel speech enhancement [33]–[35] and distant ASR [38], [39]. In this section, we first briefly introduce our adopted GEV beamformer which aims to maximize the signal-to-noise power ratio in the output [37]. Using the information provided by single-channel deep models, a CGMM is adopted to better estimate the cross-power density matrices in the GEV beamformer, while avoiding the speaker permutation problem.

1) GEV Beamformer: In the short-time Fourier transform (STFT) domain [58], the signal model can be simply expressed as a vector notation:

$$\boldsymbol{y}(t,f) = \boldsymbol{g}(f)s(t,f) + \boldsymbol{n}(t,f) = \boldsymbol{x}(t,f) + \boldsymbol{n}(t,f) \tag{5}$$

where f is the frequency bin index and t is the frame index; $\boldsymbol{x}(t,f)$ and $\boldsymbol{n}(t,f)$ are D-dimensional complex vectors that consist of the STFT-domain representations of convolved speech signals and noises, respectively; $\boldsymbol{y}(t,f)$ is the observed signal from N microphones of the reference array, $s(t,f)$ is the STFT of the source speech signal; and $\boldsymbol{g}(f)$ is the signal steering vector. We assume that the analysis window is longer than all the channel impulse responses and $\boldsymbol{n}(t,f)$ is relatively stationary.

The goal of a GEV beamformer is to find a linear vector of filter coefficients $\boldsymbol{W}(f)$ to maximize the signal-to-noise power ratio in each frequency bin [37]:

$$\boldsymbol{W}_{\text{GEV}}(f) = \arg\max_{\boldsymbol{W}} \frac{\boldsymbol{W}^H(f)\boldsymbol{R}_{\boldsymbol{xx}}(f)\boldsymbol{W}(f)}{\boldsymbol{W}^H(f)\boldsymbol{R}_{\boldsymbol{nn}}(f)\boldsymbol{W}(f)} \tag{6}$$

where $\boldsymbol{R}_{\boldsymbol{xx}}(f)$ and $\boldsymbol{R}_{\boldsymbol{nn}}(f)$ are the cross-power density matrices of the speech and noise terms, respectively. The above cost function has the same form as the Rayleigh coefficient. With $EV\{\}$ denoting the eigenvector corresponding to the largest eigenvalue, we can have a closed-form solution:

$$\boldsymbol{W}_{\text{GEV}}(f) = EV\{\boldsymbol{R}_{\boldsymbol{nn}}^{-1}(f)\boldsymbol{R}_{\boldsymbol{xx}}(f)\} \tag{7}$$

The cross-power density matrices can be defined as:

$$\boldsymbol{R}_{\boldsymbol{vv}}(f) = \sum_{t=1}^{T} \lambda_v(t,f)\boldsymbol{y}(t,f)\boldsymbol{y}^H(t,f) \tag{8}$$

where v can represent the speech or noise class, and $\lambda_v(t,f)$ denotes the probabilities of v in the time-frequency bin (t,f). To reduce the speech distortions introduced by the GEV beamformer, a single-channel post-filtering is adopted as a blind analytic normalization (BAN):

$$W_{\text{BAN}}(f) = \frac{\sqrt{\boldsymbol{W}_{\text{GEV}}^H(f)\boldsymbol{R}_{\boldsymbol{nn}}(f)\boldsymbol{R}_{\boldsymbol{nn}}(f)\boldsymbol{W}_{\text{GEV}}(f)/N}}{\boldsymbol{W}_{\text{GEV}}^H(f)\boldsymbol{R}_{\boldsymbol{nn}}(f)\boldsymbol{W}_{\text{GEV}}(f)} \tag{9}$$

Finally, the estimate for the source signal is achieved as:

$$\widetilde{s}(t,f) = W_{\text{BAN}}(f)\boldsymbol{W}_{\text{GEV}}^H(f)\boldsymbol{y}(t,f) \tag{10}$$

Obviously, the key of the GEV beamformer is the estimation of time-frequency masks $\lambda_v(t,f)$, which are used to calculate the spatial correlation matrices.

2) CGMM-Based Mask Estimation With Deeply Learned Masks: For time-frequency mask estimation, an approach using a speech spectral model based on a CGMM has been proved to be beneficial to an ASR system in [38]. Under the assumption of the sparseness of the speech in the time-frequency domain [36], the observed signals can be clustered into $K+1$ classes with each representing a noisy speech signal or only noise, where K denotes the number of sources (or speakers). In practice, the number of speakers where overlapping speech occurs is uncertain in the CHiME-5 data. Since we only need to focus on the target speaker, in this study, we simply set the clusters of observed signals as three kinds, corresponding to the observed target speaker (T), observed interference speakers (I) and noises (N) only. Thus, the generative model of the observed signal $\boldsymbol{y}(t,f)$ is modeled by a CGMM with mixture weights $\alpha_m(f)$ satisfying $\sum_m \alpha_m(f) = 1$, as follows:

$$\boldsymbol{y}(t,f) \sim \sum_m \alpha_m(f)\mathcal{N}_c\Big(0, \Theta_m(t,f), \boldsymbol{R}_m(f)\Big) \tag{11}$$

where symbol m represents different categories, T, I and N, while $\Theta_m(t,f)$ and $\boldsymbol{R}_m(f)$ indicate the signal variance and the covariance matrices of category m, respectively.

Conventionally, the CGMM parameters are estimated iteratively by the expectation-maximization (EM) algorithm with poor initial values, for example, by simply using the covariance matrix of the observed signals with a short duration or an identity matrix to initialize the $\boldsymbol{R}_m(f)$. However, the approach leads to a severe problem that the permutation of the output signals cannot be determined, especially when the number of speaker sources is undetermined. To avoid it, we use our single-channel deep learning based models to control the initialization process of CGMM parameters. As shown in Fig. 2, speaker-dependent speech separation models (SS models) yield the time-frequency masks of the target speaker and interference speakers. An initialization of noise masks is also provided by deep learning based speech enhancement models (SE models), which can additionally handle non-stationary noises. Thus, the covariance matrices of three categories in the CGMM are all initialized from our deep learning based masks in Sections III-B and III-C.

Algorithm 3: Mask Fusion of Four SD-SS Models in One Session.

Input: Testing utterance of the target speaker A, and SD-SS models of all four speakers A, B, C, D

Output: The improved mask estimation of the target speaker A for CGMM initialization: λ_T

1: Given an individual testing utterance of speaker A, define the estimated masks as: λ_{TA}, λ_{TB}, λ_{TC}, and λ_{TD}, where λ_{TN} indicates the output of the SD-SS model from speaker index N.
2: **for** each time-frequency bin (t,f) **do**
3: initialize $\lambda_T(t,f)$ with the mask value of the target speaker $\lambda_{TA}(t,f)$
4: find the max value max and its corresponding speaker index M among four λ_{TN} values
5: **if** $max < 0.3$ **then**
6: $\lambda_T(t,f) = 0$
7: **end if**
8: **if** index $M = A$ and $max \geq 0.5$ **then**
9: $\lambda_T(t,f) = 1$
10: **end if**
11: **if** index $M \neq A$ and $max \geq 0.8$ **then**
12: $\lambda_T(t,f) = 0$
13: **end if**
14: **end for**

After the convergence of the EM algorithm, the posterior probabilities of three classes are merged into two categories indicating the target speaker speech and other sounds. The time-frequency masks are denoted as λ_{Tc} and $1-\lambda_{Tc}$, where c indicates the convergence of the CGMM. Finally, the λ_v in the GEV beamformer is determined.

3) Mask Fusion: As mentioned above, good parameter initialization is quite important for the CGMM method. To attain better estimation, we further utilize a mask fusion approach to take advantage of the speaker-specific information of all 4 speakers in one session. The details are presented in Algorithm 3. After combining all the information from every speaker-dependent model, the confidence of the probability mask in each time-frequency bin is much enhanced. This approach is equivalent to getting global information by combining the knowledge of each model, and achieves more reliable mask estimation than using only one model.

4) SD Mask Post-Filtering: Unlike the conventional output of a GEV beamformer in Eq. (10), we add another operation, namely SD mask post-filtering, by multiplying the masks of the target speaker that are attained after the convergence of the CGMM. The process provides strong nonlinear suppression ability in the single-channel time-frequency domain. In our experiments, this step was important to improve the final speech recognition results. The final output is written as:

$$\widetilde{s_p}(t,f) = W_{\text{BAN}}(f)\boldsymbol{W}_{\text{GEV}}^H(f)\boldsymbol{y}(t,f)\lambda_{Tc}(t,f) \tag{12}$$

IV. Experimental Setup

The CHiME-5 challenge contains two tracks, a single array containing only the reference array data and a multiple array containing data from all six arrays. placed in different positions of the home. Here, we focus on the single-array track where only one reference array can be used to recognize a given test utterance. In this section, we introduce our experimental configuration.

A. Official ASR Baseline in CHiME-5

To better illustrate the effectiveness of our proposed front-end, we used the official time delay neural network (TDNN) model [59] with lattice-free maximum mutual information (LF-MMI) training via the KALDI toolkit [60]. Mel-frequency cepstral coefficients (MFCCs) and i-vectors were adopted as input features. The data used in the acoustic model training were only from the official training set, including both close-talking data from binaural microphones and far-field data from the reference microphone array. Three different levels of speed perturbation were conducted to augment the data size, which are 0.9, 1.0 and 1.1. In the front-end, a weighted delay-and-sum beamformer [61] was used as the default multi-channel speech enhancement approach. The baseline WER of 81.3% was officially reported on the development test set with the officially provided 3-gram language model. Following the official approach as in [42], our reproduction of the baseline achieved the WER of 80.6%, which was slightly better. In the rest of the paper, the results of the baseline system refer to our reproduced version.

B. Single-Channel Model Training

As illustrated in Fig. 2, multi-channel preprocessing did not rely on model training, and the generated data remained unchanged which was approximately 40 hours. While for single-channel deep learning based speech enhancement and separation, the necessary training steps were required. However, there were two big differences between them. First, the speech enhancement approach was speaker-independent while the separation approach was speaker-dependent. Second, the enhancement model was built on the whole training set, while the separation model was built on each testing session using the oracle speaker diarization information.

1) Speech Enhancement Model Training: The data multi-channel preprocessing approach was considered as "clean" source data. Then, unlabeled segments derived from human transcriptions were extracted from the channel-1 data, as the noise corpus. Utterances from "clean" speech were corrupted with the abovementioned noise segments at three SNR levels (-5 dB, 0 dB and 5 dB) to build a 120-hour training set, consisting of pairs of clean and noisy utterances. For model architecture, we adopted the densely connected progressive learning based speech enhancement model in [51]. All signals were sampled at the 16 kHz rate. The frame length and shift were 512 and 256 samples, respectively. The 257-dimensional feature vector was used for both LPS and IRM targets. The computational network toolkit (CNTK) [62] was used for training. More training details can be found in the best configuration in [51].

When in the testing phase, only channel-1 signals of the reference array were adopted as the input to the speech enhancement model to maintain consistency with the noise types in the simulated training mixtures. The generated mask estimations represented the speech presence probability in each time-frequency bin.

2) Speech Separation Model Training: As described in Section III-C, in the first stage, we used selected non-overlapping segments to simulate 50,000 mixture utterances of the target speaker, with the corruption from other interference speakers at five SNR levels (-5 dB, 0 dB, 5 dB, 10 dB and 15 dB), to train the SD-SS1 model. Then, data cleaning and augmentation were accomplished after utilizing the estimated IRM taken from the SD-SS1 model. Next in the second stage, the separated sentences were used to simulate another 50,000 utterances for training the SD-SS2 models.

The input features were the same as the speech enhancement model. However, to better utilize the sequential information, we used a two-layer BLSTM for both SD-SS1 and SD-SS2 model, each direction with 512 cells. Additionally, 7-frame expansion was used for the input. In testing, a given utterance was directly sent to the corresponding SD-SS2 model. The generated masks represented the speech presence probability of the target speaker in each time-frequency bin.

C. Multi-Channel Beamforming

For the GEV beamformer, the multi-channel STFT coefficients were extracted from the test speech at a 16 kHz sampling frequency using a Hanning window of length 512 and shift of 128, resulting in 257 frequency bins. Given the estimated time-frequency masks of the target speaker, interference speakers and noise, the CGMM-based mask estimator was initialized. After convergence of the EM algorithm, the GEV beamformer generated the final speech by using the final time-frequency mask of the target speaker.

V. Results and Analysis

Because the oracle text labels of the evaluation set were not provided, we explored the front-end methods on the development set, which contained two separate sessions, namely, Session 02 (S02) and Session 09 (S09). Typically, S02 was used for tuning parameters, and S09 was used for evaluation.

A. Single-Channel Speaker-Dependent Speech Separation

In Table III, WER comparisons are listed at different stages of our proposed speech separation system on S02 from the development set. We fixed the back-end acoustic model and evaluated different versions of processed data. Moreover, individual results of all 4 speakers, including P05, P06, P07 and P08, are also presented to show the effectiveness of the speaker-dependent strategy.

Several observations could be made. First, our multi-channel preprocessing approach yielded slightly better results than the

TABLE III
WER COMPARISON OF THE OFFICIAL BASELINE METHOD USING BEAMFORMIT AND OUR METHODS IN SESSION 02

S02 WER (%)	Official BeamformIt	Multi-channel Preprocessing	SD-SS1	SD-SS2
P05	83.2	82.0	79.3	**78.6**
P06	77.1	75.7	72.5	**70.8**
P07	79.7	79.9	81.4	**76.0**
P08	88.0	87.5	83.0	**75.8**
Ave.	81.2	80.3	78.0	**74.8**

official BeamformIt method [61], which was attributed to the dereverberation using GWPE [45]. Although this gain was not very significant, multi-channel preprocessing was still needed for the following speech separations to work well, as it was not necessary to explicitly consider the effects of strong noise and reverberation. Next, the separated speech achieved by the SD-SS1 model reduced the WERs of all speakers except P07. There were two main problems of the first-stage speech separation. One problem was that the training target used for SD-SS1 might lead to large speech distortions. The other problem was that each target speaker had little usable data quantity. For instance, P07 only had approximately 5.6 minutes of non-overlapping data, which was insufficient for training a speaker-dependent model. After data cleaning and augmentation, we attained more and purer speech for every speaker, which enabled the second stage to yield better performance. In the last column, SD-SS2 reduced the average WER to 74.8%, an absolute WER reduction of 6.4% in comparison to the official baseline method. By comparing the results of SD-SS1 and SD-SS2, the proposed approach greatly improved the performance for speakers whose useable data were limited, such as P07 and P08. These results demonstrated that the data cleaning and augmentation operations between the two stages were indeed effective.

To better illustrate the effectiveness of our speech separation system, an utterance of P05 selected from Session 02 was presented, as shown in Fig. 6. In the uppermost part, the distribution of speech from different speakers was drawn after manual audiometry, where red bars indicated the target speaker P05 and blue bars denoted the corresponding interference speaker P06. Compared with the spectrogram after multi-channel preprocessing, speech processed by SD-SS1 models removed most of the interference parts both on overlapping regions and non-overlapping regions. Though it also introduced some speech distortions to the target speech, the spectrogram indicated that the SD-SS1 model greatly met our expectation of data cleaning. By using these different learning targets introduced in Section III-C, the final processed speech made a trade-off between speech distortion and speech intelligibility, yielding better recognition performance. As seen in the bottom of Fig. 6, the power and strength of interference speech were largely impaired.

The overall results of the development set were listed in Table IV. Compared with S02, the performance gains were relatively small in S09. On the one hand, the recording quality in S09 was worse than that in S02. Speech sounds were often imperceptible, even by human auditory sensation. On the other hand, 4 speakers in S09 are all female, which made it quite

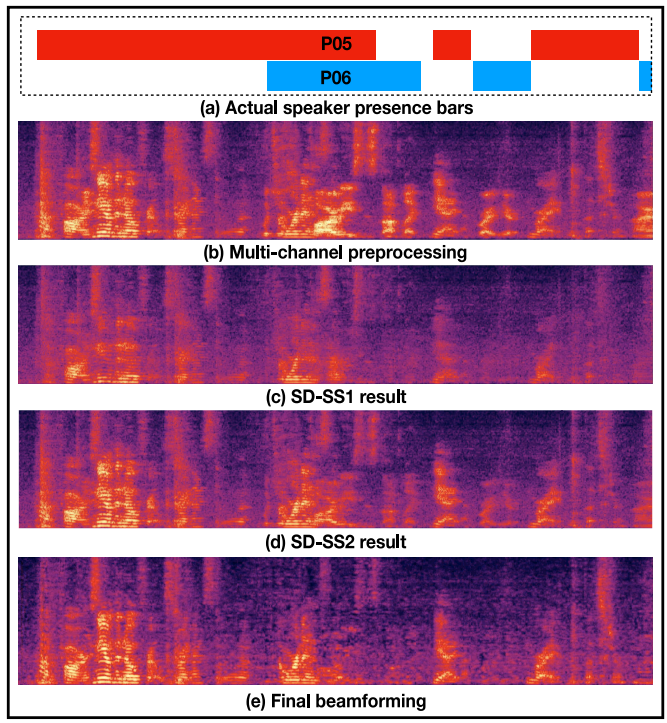


Fig. 6. An utterance of speaker P05 from Session 02. In (a), the red bar represents speech regions of the target speaker P05, while the blue bar represents the interference speaker P06. (b)-(d) are the spectrograms from multi-channel preprocessing, SD-SS1 model and SD-SS2 model, respectively. The spectrogram after our final multi-channel beamforming is listed in (e).

TABLE IV
OVERALL WER COMPARISON ON THE DEVELOPMENT SET

Dev. Set WER (%)	Session	Dining	Kitchen	Living	Ave.	Overall
Official BeamformIt	S02	78.2	86.4	78.3	81.2	80.6
	S09	81.1	81.2	77.2	79.8	
Multi-channel Preprocessing	S02	79.4	86.9	75.5	80.3	80.2
	S09	82.6	81.1	77.2	80.1	
SD-SS1	S02	77.8	83.7	73.3	78.0	78.5
	S09	83.6	78.9	76.6	79.3	
SD-SS2	S02	**74.4**	**81.8**	**69.2**	**74.8**	**76.0**
	S09	**81.6**	**77.3**	**76.4**	**78.1**	

challenging to separate them from each other [63]. Compared with multi-channel preprocessing, the recognition performance was progressively improved by two separation stages. According to different scenarios, the results in the living room were better than those in the dining room and kitchen, partially due to fewer environmental noises. Overall, compared with the WER of 81.3% reported in the official baseline [42], our best result for the single-channel setting yielded an absolute WER reduction of 5.3%.

Before moving to the next stage of multi-channel beamforming, we further explored the effectiveness of our single-channel estimated masks in a mismatch manner with official BeamformIt data. Besides the abovementioned separation masks estimated by SD-SS2 models, enhancement masks were also generated as the description in Section IV-B1. Those two kinds of masks were separately applied to the BeamformIt data, then spectral features of masked speech data were recognized again by the official

TABLE V
OVERALL WER COMPARISON ON THE DEVELOPMENT SET TO SHOW THE EFFECTIVENESS OF ESTIMATED TIME-FREQUENCY MASKS FROM OUR DEEP LEARNING BASED MODELS

Dev. Set WER (%)	Dining	Kitchen	Living	Ave.
Official BeamformIt	79.4	84.2	77.9	80.6
+ Our SE masks	79.4	84.4	77.6	80.5
+ Our SD-SS masks	76.1	80.4	73.3	76.6

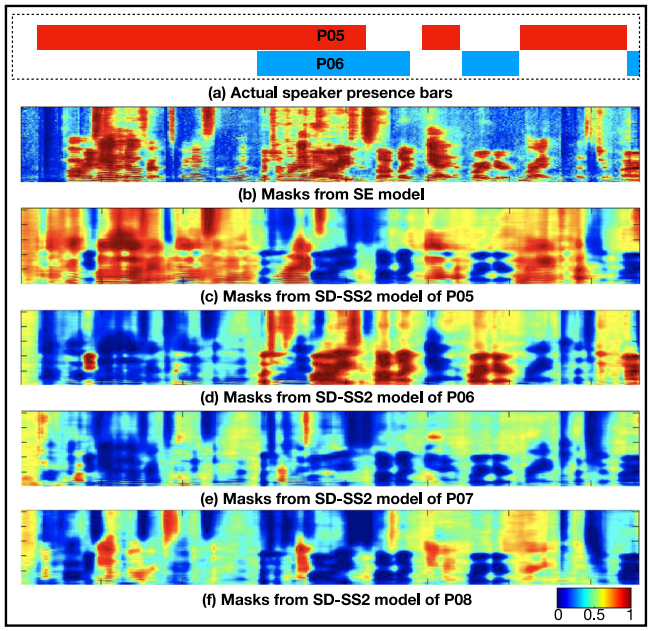


Fig. 7. Illustration of masks from deep learning based models, including the speech enhancement model and speaker-dependent speech separation model from every speaker.

back-end model. The detailed WER comparison was presented in Table V. After masking speaker-dependent speech separation masks, the overall WER was significantly reduced from 80.6% to 76.6%. Though one enhancement mask illustrated in Fig. 7 seemed acceptable, our speech enhancement model was of little help to improve the recognition performance according to the WERs.

B. Multi-Channel Beamforming With Deeply Learned Masks

1) The CGMM Settings: Unlike former CHiME challenges, the recognition performance of CHiME-5 is severely degraded, mainly due to the conversational speaking style. To quantitatively explore the difference, we first used a CGMM that only considered speech and noise, denoted as CGMM1. The number of categories in CGMM was set to 2, while the deep learning based speech enhancement model provided the noise masks for parameter initialization. Second, another 2-class CGMM, denoted as CGMM2, was adopted with the initialization from SD-SS models, which aimed to separate only the target speaker and other interferences. Furthermore, we extended the CGMM with three classes, consisting of the target speaker, interference speakers and the noise, namely CGMM3. The comparison of CGMM settings is summarized as follows:

1) CGMM1: **2 classes**, modeling noise and mixture speech, with the **initialization by the SE model**
2) CGMM2: **2 classes**, modeling the target speaker and other sounds (including noise and potential interference speakers), with the **initialization by SD-SS models**
3) CGMM3: **3 classes**, modeling the target speaker, interference speakers and noise, with the **initialization by both the SE model and SD-SS models**

TABLE VI
THE WER COMPARISON ON SESSION 02 FOR MULTI-CHANNEL BEAMFORMING BASED ON DIFFERENT CGMM SETTINGS

S02 WER (%)		CGMM1		CGMM2		CGMM3		CGMM3 Mask Fusion	
P05		79.6		71.4		67.8		68.4	
P06		70.4		64.5		61.6		61.1	
P07		77.7		68.1		67.1		64.6	
P08		93.5		70.4		70.1		65.5	
Ave.		78.4		68.1		65.9		64.5	
Single	Multi	66.0	78.6	63.5	68.2	64.7	65.9	64.6	64.5

The recognition performance of different CGMM settings was shown in Table VI. Compared with our baseline system, CGMM1 yielded a relatively small reduction in average WER in Session 02, from 81.2% to 78.4%. However, CGMM1 underperformed our proposed single-channel speech separation approach presented in Table III (with a WER of 74.8%), which could be attributed to two possible reasons. One reason was due to the complex environmental factors in CHiME-5, while the other reason implied that speech separation masks were more effective than speech enhancement masks in terms of recognition accuracy, as also stated in Table V. Next, CGMM2 used target speaker probability masks initialized by our SD-SS models instead of noise masks in CGMM1. The WER of CGMM2 was significantly reduced to 68.1%, indicating that the speaker-dependent design was quite important for the CHiME-5 challenge. Finally, CGMM3 modeled all three categories and achieved a better WER (65.9%) result than both CGMM1 and CGMM2.

As mentioned in Section III-D2, the questions about whether there were interference speakers and the number of them were both uncertain. It was worth exploring the relationship between the number of categories set in the CGMM and the final performance on different classes of data. The bottom row in Table VI presents the average WER results of single-speaker segments and multi-talker segments, denoted as Single and Multi. These segments were classified according to official time annotations. From CGMM2 to CGMM3, where the class number changed from 2 to 3, the WER of single-speaker segments suffered a slight increase from 63.5% to 64.7%. However, the WER of multi-talker part obtained an obvious decline, from 68.2% to 65.9%. In fact, there was a big difference between the amounts of these two kinds of segments. In S02, the single-speaker part contained 87 segments, while the multi-talker part had 3735 segments. Thus, the overall performance largely depended on the performance of the multi-talker segments, which also made it reasonable to set the number of CGMM classes to 3.

TABLE VII
ANALYSIS OF DIFFERENT COMPONENTS IN CGMM2

S02	Components		WER(%)				
	CGMM Optimization	SD Mask Post-filtering	P05	P06	P07	P08	Ave.
CGMM2	×	×	78.7	70.3	77.4	84.5	76.5
	✓	×	75.6	68.6	76.0	83.6	74.6
	×	✓	73.0	67.5	74.4	77.6	72.2
	✓	✓	71.4	64.5	68.1	70.4	68.1

TABLE VIII
OVERALL WER COMPARISON ON THE DEVELOPMENT TEST SET

Dev. Set WER (%)	Session	Dining	Kitchen	Living	Ave.	Overall
Official BeamformIt	S02	78.2	86.4	78.3	81.2	80.6
	S09	81.1	81.2	77.2	79.8	
CGMM3 Mask Fusion	S02	64.7	72.9	57.5	64.5	68.8
	S09	75.5	74.3	77.9	75.8	
Back-ends Fusion	S02	63.1	71.9	57.0	63.6	67.6
	S09	74.3	72.7	75.9	74.2	

2) Important Components in Multi-Channel Beamforming: In order to better evaluate the effectiveness of many components of our proposed approach, we exhibited some results of discarding some steps in CGMM2 in Table VII. The experiments revolved around the adoption of two components, one was about the CGMM optimization process for mask refinement, and the other one was about the SD mask post-filtering proposed in Section III-D4. Obviously, Both methods contributed greatly to the final performance, that made them indispensable in our final approach as written in Eq. (12).

3) Mask Fusion: In addition to using only the target speaker's model, the optional stage named "Mask Fusion" in Fig. 2 could be activated by using all four SD-SS models in one session. From the rightmost column in Table VI, remarkable improvements were obtained, especially for P07 and P08 whose usable non-overlapping data were very limited. It demonstrated that using the information from speaker-dependent models of other speakers could help these weakly-trained target speaker models which were short of source data. Masks generated by deep learning models were also illustrated in Fig. 7. Speaker-dependent speech separation models basically captured the corresponding speaker's presence, even in the overlapping part between P05 and P06. For those mismatched models, such as P07 and P08, the false alarm was also acceptable. Only when all separation models performed their tasks could the mask fusion strategy maximized its abilities, which was also the reason why the mask fusion achieved the best performance in Table VI.

C. Effects of Front-End for Back-End Fusion

Up to now, the effectiveness of our proposed speaker-dependent front-end has been verified, especially regarding the superiority of the deep learning based two-stage speaker-dependent approach. One more advantage of the supervised method was that it can provided multiple separation models using several different training configurations. For example, the original non-overlapping segments were mostly too short, the simulated mixture data could not be effectively utilized by BLSTM networks to capture long-term sequential information. Accordingly, we concatenated short segments to form long segments and trained a new batch of SD-SS models with them. To improve the performance stability, different versions of segregated speech from the short-segment models and long-segment models were combined via lattice combination with equal weights, while keeping the back-end system unchanged. The overall ASR results on the development set were shown in Table VIII. Compared with the baseline, our proposed front-end was able to reduce the average WER by absolute 13.0%.

VI. DISCUSSION

A. System Settings

In this work, the proposed overall system contains a number of experimental settings or parameters. Those settings can be roughly divided into two categories. One refers to the most crucial settings related to system design, which were explored by a series of experiments, including: the design of learning targets in the two-stage speech separation, the process of data selection, the CGMM settings, and the final expression of the output signal. For those settings, we've presented comprehensive results to support the effectiveness of them in Section V.

The other one category consists of parameters which were not verified as the optimal schemes for our proposed system. For example, we used a progressively learning based uni-LSTM network for speech enhancement tasks and a BLSTM network for speech separation tasks. They were directly used in according to our experiences of both tasks, no detailed comparison was conducted about the model architectures. On the base of our proposed framework, we believe that using more powerful or well-tuned models is possible to get better results.

B. System Complexity

As seen in Fig. 2, the overall diagram is too complicated for a practical ASR system, let alone using extra oracle diarization information. Here we discuss several future directions to simplify the system complexity. First, speaker-independent speech separation of multi-talker speech under noisy conditions is the most basic solution to the CHiME-5 challenge.

Second, multi-channel preprocessing was theoretically needed as it was expected to prepare enhanced speech for the deep-learning based mask estimation methods. However, it's unable to perfectly remove the existing noise and reverberation, and the recognition results of it showed little improvements than of official BeamformIt, as listed in Table IV. We can consider performing both speech separation and speech enhancement training directly on official BeamformIt data, so the multi-channel preprocessing and channel-1 data selection can be omitted. It's also helpful to evaluate the generalization ability of the proposed deep-learning based mask estimation methods.

C. Evaluation Metric of Multi-Talker Data

The reported WERs of CHiME-5 data are far above common human performance level, especially for far-field array data. Even for binaural data, the official baseline only yields a WER of 47.9%. We think it's still an open question about finding a

good evaluation method of such multi-talker data, such as setting multiple labels of a given overlapping sentence. For overlapping data, currently a pure recognition model doesn't know how to discriminate different persons. However, it's also difficult for humans to recognize every word of each speaker within one chance. More robust and felexible front-end processing is needed to separate signals of different talkers.

VII. Conclusion

In this study, we have presented a novel speaker-dependent approach which can effectively handle far-field multi-talker speech in the CHiME-5 challenge. Unlike previous solutions, the proposed method jointly addresses multiple environmental factors and conversational speaking styles in the CHiME-5 data. Specifically, a two-stage single-channel speaker-dependent speech separation framework is designed to extract speech of the target speaker in each session, based on the given oracle speaker diarization information, allowed by the challenge rules. In addition, the estimated probability masks of the target speaker finely avoid the permutation problem in the CGMM-based mask estimator with three classes. Then, a GEV beamformer with CGMM-based mask estimation and SD mask post-filtering is adopted for enhancing the signal. Compared with the officially reported results, our proposed approach achieved a significant average WER reduction which declined to 67.6% from 81.3%. Finally, by integrating the proposed front-end, our final system ranked the first place of all four evaluation categories among all participating systems in the CHiME-5 challenge. In future studies, we plan to relax the assumptions involving the oracle speaker diarization information, and simplify our front-end system.

References

[1] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8599–8603.

[2] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1984, vol. 9, pp. 328–331.

[3] J. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," 1988.

[4] D. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang*, 1992, pp. 357–362.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[6] J. Carletta, "Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus," *Lang. Resour. Eval.*, vol. 41, no. 2, pp. 181–190, 2007.

[7] K. Kinoshita *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, 2013, pp. 1–4.

[8] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Comput. Speech Lang.*, vol. 27, no. 3, pp. 621–633, 2013.

[9] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 126–130.

[10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.

[11] M. Cooke and T. Lee, "Speech separation challenge," 2006. [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm

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

[13] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[15] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[16] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[17] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7092–7096.

[18] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[19] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[20] E. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953.

[21] J. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 31–35.

[22] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 246–250.

[23] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 241–245.

[24] C. Weng, D. Yu, M. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1670–1679, Oct. 2015.

[25] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017, pp. 1901–1913.

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

[27] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[28] Y. Liang, "Enhanced independent vector analysis for audio separation in a room environment," Ph.D. dissertation, Loughborough Univ., School Elect., Electr. Systems Eng., Loughborough, U.K., 2013.

[29] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1368–1381, Jul. 2011.

[30] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.

[31] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[32] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 549–557, Mar. 2011.

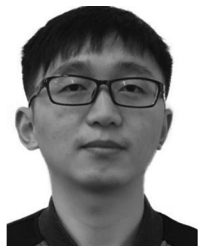
[33] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

[34] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, Sep. 2009.
[35] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.
[36] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.
[37] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.
[38] T. Yoshioka *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 436–443.
[39] T. Hori *et al.*, "The MERL/SRI system for the 3rd CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 475–481.
[40] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 444–451.
[41] Y. Tu, J. Du, L. Sun, F. Ma, and C. Lee, "On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones," *Proc. INTERSPEECH*, 2017, pp. 394–398.
[42] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. INTERSPEECH*, 2018, pp. 1561–1565.
[43] J. Du, Y. Tu, Y. Xu, L. Dai, and C. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. 12th Int. Conf. Signal Process.*, 2014, pp. 473–477.
[44] T. Gao, J. Du, L. Dai, and C. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," *Speech Commun.*, vol. 95, pp. 28–39, 2017.
[45] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.
[46] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2010, pp. 165–172.
[47] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.
[48] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proc. 41st SICE Annu. Conf.*, 2002, vol. 4, pp. 2138–2143.
[49] D. Kitamura, "Algorithms for independent low-rank matrix analysis," 2018. [Online]. Available: http://d-kitamura.net/pdf/misc/AlgorithmsForIndependentLowRankMatrixAnalysis.pdf
[50] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.
[51] T. Gao, J. Du, L. Dai, and C. Lee, "Densely connected progressive learning for LSTM-based speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5054–5058.
[52] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends Amplification*, vol. 12, no. 4, pp. 332–353, 2008.
[53] L. Sun, J. Du, L. Dai, and C. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. Hands-Free Speech Commun. Microphone Arrays*, 2017, pp. 136–140.
[54] I. Medennikov *et al.*, "The STC system for the CHiME 2018 challenge," in *Proc. HiME Workshop Speech Process. Everyday Environ.*, 2018, pp. 1–5. [Online]. Available: spandh.dcs.shef.ac.uk/chime_workshop/programme.html
[55] F. Weninger, J. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2014, pp. 577–581.
[56] F. Weninger *et al.*, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2015, pp. 91–99.
[57] 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.
[58] R. J. Mcaulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.
[59] D. Povey *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. INTERSPEECH*, 2016, pp. 2751–2755.
[60] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, IEEE Signal Processing Society, 2011.
[61] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.
[62] F. Seide and A. Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 2135–2135.
[63] Y. Wang, J. Du, L. Dai, and C. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1535–1546, Jul. 2017.

Lei Sun received the B.S. degree from the Northeastern University at Qinhuangdao, Qinhuangdao, China. He is currently working toward the Ph.D. degree with the University of Science and Technology of China, Hefei, China. His research interests include speech enhancement, speaker diarization, and robust speech recognition.

Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2004 to 2009, he was with iFlytek Speech Lab, USTC, and also worked as an Intern twice for nine months with Microsoft Research Asia (MSRA), Beijing, China. In 2007, he also worked as a Research Assistant for six months with the Department of Computer Science, The University of Hong Kong. From July 2009 to June 2010, he was with iFlytek Research on speech recognition. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing, USTC.

Tian Gao received the B.S. degree from the Department of Communication Engineering, Hefei University of Technology, Hefei, China, in 2013, the D.E. degree from the University of Science and Technology of China, Hefei, China, in 2018. He is currently with iFlytek Research, Hefei, China. His research interests include speech enhancement and robust speech recognition.



Yi Fang received the D.E. degree from the Institute of Acoustics, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Hefei, China, in 2018. He is currently with iFlytek Research, Hefei, China. His current research interests include speech separation, acoustic echo suppression, and microphone arrays.

Feng Ma received the B.Eng. and M.S. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2009 and 2012, respectively. He is currently with iFlytek Research, Hefei, China. His current research interests include acoustic echo cancellation, microphone arrays, and robust speech recognition.

Chin-Hui Lee is a Professor with the School of Electrical, and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Before joining academia in 2001, he had 20 years of industrial experience ending in Bell Laboratories, Murray Hill, NJ, USA, as a Distinguished Member of Technical Staff, and the Director of the Dialogue Systems Research Department. He has authored/coauthored more than 400 papers, and 30 patents, and was highly cited for his original contributions with an h-index of 66. He is a Fellow of ISCA. He received numerous awards, including the Bell Labs President's Gold Award in 1998. He won the SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition." In 2012, he was invited by ICASSP to give a plenary talk on the future of speech recognition, and the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition.