



# Improved Speaker-Dependent Separation for CHiME-5 Challenge

Jian Wu<sup>1,2\*</sup>, Yong Xu<sup>3</sup>, Shi-Xiong Zhang<sup>3</sup>, Lian-Wu Chen<sup>2</sup>, Meng Yu<sup>3</sup>, Lei Xie<sup>1†</sup>, Dong Yu<sup>3</sup>

<sup>1</sup>School of Computer Science, Northwestern Polytechnical University, Xi'an, China

<sup>2</sup>Tencent AI Lab, Shenzhen, China

<sup>3</sup>Tencent AI Lab, Bellevue, USA

{jianwu, lxie}@nwpu-aslp.org, {lucayongxu, auszhang, lianwuchen, raymondmyu, dyu}@tencent.com

## Abstract

This paper summarizes several contributions for improving the speaker-dependent separation system for CHiME-5 challenge, which aims to solve the problem of multi-channel, highly-overlapped conversational speech recognition in a dinner party scenario with reverberations and non-stationary noises. Specifically, we adopt a speaker-aware training method by using i-vector as the target speaker information for multi-talker speech separation. With only one unified separation model for all speakers, we achieve a 10% absolute improvement in terms of word error rate (WER) over the previous baseline of 80.28% on the development set by leveraging our newly proposed data processing techniques and beamforming approach. With our improved back-end acoustic model, we further reduce WER to 60.15% which surpasses the result of our submitted CHiME-5 challenge system without applying any fusion techniques.

**Index Terms:** CHiME-5 challenge, speaker-dependent speech separation, robust speech recognition, speech enhancement, beamforming

## 1. Introduction

With the recent progress in front-end audio processing, acoustic modeling and language modeling, automatic speech recognition (ASR) techniques are widely deployed in our daily life. However, the performance of ASR will severely degrade in challenging acoustic environments (e.g., overlapping, noisy and reverberated speech), mainly due to the unseen complicated acoustic conditions in the training. Many previous work on acoustic robustness focused on one aspect, e.g., speech separation [1, 2, 3, 4], enhancement [5, 6, 7, 8, 9], dereverberation [10, 11, 12], and etc. Those experiments were conducted on simulated data, which is not realistic in real applications. Recently released CHiME-5 challenge [13] provided a large-scale multi-speaker conversational corpus recorded via Microsoft Kinect in real home environments and targeted at the problem of distant multi-microphone conversational speech recognition. As the recordings are extremely overlapped among multiple speakers and corrupted by the reverberation and background noises, WERs reported on the dataset are fairly high. In this paper, we make several efforts based on our previously submitted speaker-dependent system [14] which ranked 3rd under unconstrained language model (LM) and 5th under constrained LM for the single device track, respectively.

The difficulties of CHiME-5 are three-fold. First, the natural conversation contains casual contents, sometimes occupied by laugh and coughing. Speaker interference is common in conversational speech as well, which causes degradation on speech

\* work done during internship at Tencent AI lab.

† corresponding author

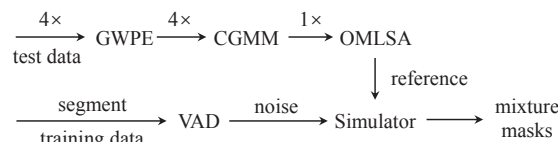


Figure 1: Flow chart of data processing and simulation

recognition. Second, hardware devices, far-field wave propagation and ambient noises cause audio clipping, signal attenuation and noise corruption, respectively. Furthermore, the lack of the clean speech for supervised training greatly limits the algorithm design and external datasets are not allowed according to the rule of CHiME-5. By considering these aspects, robust front-end processing of target speaker enhancement is critical for improving the ASR performance.

Recent studies have made great efforts in multi-channel speech enhancement [7, 8, 9, 15] and most of them depended on the Time-Frequency (TF) masks. Deep learning based beamforming became the most popular approach since CHiME-3 and CHiME-4 challenge [16]. However, in CHiME-5, it's difficult to train a speech enhancement mask estimator and obtain accurate predictions due to the lack of the oracle clean data required by supervised training. On the other hand, there are many limitations on performing recently proposed monaural blind speech separation methods, e.g., DPCL [1], uPIT [2], because it's necessary to do speaker tracking due to the permutation issue. The number of speakers is also a prerequisite for those monaural speech separation approaches, while it is infeasible in CHiME-5 challenge. However, considering that the target speaker ID is given in each utterance, we tried speaker-dependent (SD) separation in [14] and Du et al. used a speaker dependent system along with a two-stage separation method in [17].

In this paper, we focus on single-array track (only one reference array used) and achieve significant improvement with the following contributions. First, we process data by making use of GWPE [18], CGMM [8, 19] and OMLSA [20] to further remove the interference in the non-overlapped data segments, which are used as the training target in the SD models. In [14], suffering from low-quality training targets, the system just achieved 2% absolute reduction on WER. Second, inspired by [21, 22, 23], we incorporate i-vectors as auxiliary features, which aims to extract the target speaker. With the speaker-aware (SA) training technique, we achieve much better results using only one mask estimation model. Third, we investigate the beamforming performance, and observe that with more accurate speaker masks, generalized eigenvalue (GEV) [24] beamformer performs better than minimum variance distortionless response (MVDR) [25] beamformer. Finally, we report 10% absolute WER reduction on the development set and 20% with our improved acoustic model (AM) which is based on the fac-

tored form of time-delay neural network (TDNN-F) [26]. Compared with the single systems submitted for CHiME-5, our proposed system outperform most of them. And compared to [17], where a set of separation models were trained and a two-stage separation is performed, our SD method has low computational complexity apparently.

## 2. Proposed System

In this section, we will discuss the data processing, speaker-aware (SA) training and beamforming used in our new system and indicate how we boost previously submitted speaker-dependent front-end.

### 2.1. Data processing

In order to simulate training data for speaker-dependent models, we use non-overlapped utterances as reference, which can be segmented according to the provided annotations. However, those segments are not guaranteed to be in high signal-to-noise ratio (SNR) and may contain strong background noise, especially in the kitchen room. These issues can lead to inaccurate training targets (e.g., IRM), which may result in slow convergence and bad performance of the separation model. In order to further remove noise in those segments and improve the quality of training targets, we utilize complex Gaussian mixture model (CGMM) to estimate speech masks in a unsupervised manner and perform MVDR beamforming to suppress background noise. Following the suggestions from [27], GWPE is applied on multi-channel signals to reduce potential reverberations before beamforming, which are also proved to benefit ASR performance in the following experiments.

We use a two-component CGMM, i.e., speech and noise, and TF-masks are computed as the following posterior

$$\lambda_{t,f}^k = \frac{\alpha_k p(\mathbf{y}_{t,f} | \Theta_k)}{\sum_c \alpha_c p(\mathbf{y}_{t,f} | \Theta_c)} \quad k \in \{n, s\}, \quad (1)$$

where  $p(\mathbf{y}_{t,f} | \Theta_k) = \mathcal{N}(\mathbf{y}_{t,f} | 0, \phi_{t,f}^k \mathbf{R}_f^k)$ . Following [7, 19], speech and noise covariance matrices are estimated via

$$\Phi_f^k = \frac{1}{\sum_t \lambda_{t,f}^k} \sum_t \lambda_{t,f}^k \mathbf{y}_{t,y} \mathbf{y}_{t,y}^H \quad k \in \{n, s\}, \quad (2)$$

where  $(\cdot)^H$  means conjugate transpose. For MVDR beamforming, steer vector  $\mathbf{d}_f$  at each frequency is required, and the principal eigenvector of  $\Phi_f^s$  is an ideal estimation based on the fact that covariance matrices of the directional target is close to a rank-one matrix. With  $\Phi_f^n, \mathbf{d}_f$ , weights of MVDR is computed as

$$\mathbf{w}_f^{\text{MVDR}} = \frac{(\Phi_f^n)^{-1} \mathbf{d}_f}{\mathbf{d}_f^H (\Phi_f^n)^{-1} \mathbf{d}_f}. \quad (3)$$

Considering that the enhanced speech obtained by beamforming always contains residual noise, we continue to perform single-channel denoising. One typical statistical method is OMLSA [20], which was proposed for single-channel robust speech enhancement. Although it may introduce speech distortion, it reduces the background noise and keeps the TF regions of speech with higher energy, further improving the accuracy of target mask computation, especially in noise dominant TF bins.

As shown in Fig.1, with those processed non-overlapped segments as reference (clean) data, we perform data simulation, mask computation, network training, etc. in the following steps.

### 2.2. Speaker aware training

Some of recent blind speech separation methods need to know the number of speakers in the mixture and can not assign output to specific speaker properly. Here it's not suitable to use them in CHiME-5 challenge which requires to recognize the speech of target speaker in the given utterances. Under such circumstances, there are two optional methods for the front-end separation system. One is to make use of speaker information and condition the speech separation, similar to [22]. Another one is to train a set of models for each known speaker, like the one we used in [14] and also in [17]. In fact, the first one is more applicable to real scenarios because it can generalize to unseen speakers if model is well trained and it also can avoid the permutation problem at the same time.

Our motivation is to use i-vectors as speaker features to bias the prediction of the target masks. We tried two typical TF-masks, i.e. IRM and PSM, which are defined as

$$\begin{aligned} \mathbf{m}_{\text{IRM}} &= |\mathbf{s}_t| / (|\mathbf{s}_t| + |\mathbf{n}|), \\ \mathbf{m}_{\text{PSM}} &= |\mathbf{s}_t| \cos(\angle \mathbf{y} - \angle \mathbf{s}_t) / |\mathbf{y}|, \end{aligned} \quad (4)$$

where  $\mathbf{y}, \mathbf{s}_t, \mathbf{n}$  are short-time Fourier transform (STFT) of mixture, target speaker and noise component respectively, which satisfies the equation  $\mathbf{y} = \mathbf{s}_t + \mathbf{n}$ . When simulating the training data, we mix target speaker with background noise as well as one or two interference speakers at various SNRs. Considering that PSM is unbounded and may be negative, we truncated its value between 0 and 1. Neural networks are trained by minimizing the mean square error

$$\mathcal{L}_{\text{MSE}} = \|\hat{\mathbf{m}} - \mathbf{m}_t\|_2^2. \quad (5)$$

In the training stage, for a given noisy utterance and a specific speaker, i-vectors are computed on random selected segment from target speaker's non-overlapped set, similar to [22]. And during testing, we use average results instead of random one to get a robust and stable prediction.

### 2.3. Beamforming

Beamformer is a linear spatial filter applied on microphone signals, which suppresses energy on non-target directions and produces an enhanced output. On frequency domain it could be described as

$$s_{t,f} = \mathbf{w}_f^H \mathbf{y}_{t,f}, \quad (6)$$

where  $\mathbf{w}_f$  is a complex valued vector on the frequency  $f$ . In Section 2.1 we introduce MVDR beamforming, which is a special case of parameterized multi-channel Wiener filter (PMWF)

$$\mathbf{w}_f^{\text{PMWF}-\beta} = \frac{(\Phi_f^n)^{-1} \Phi_f^s}{\beta + \text{tr}[(\Phi_f^n)^{-1} \Phi_f^s]} \mathbf{u}_r \quad (7)$$

with  $\beta = 0$ .  $\mathbf{u}_r$  is a vector indicating reference microphone, which can be manually specified or chosen by the estimation of the posterior SNR [28]. When  $\beta = 1$ , it equals to multi-channel Wiener filter (MCWF), another widely used beamforming in signal processing.

In [7], GEV beamformer, which is obtained by Max-SNR criterion and avoids matrix inversion in the computation, provides better results than MVDR. The beamforming filter is designed to maximize expected SNR at each frequency:

$$\mathbf{w}_f^{\text{GEV}} = \arg \max_{\mathbf{w}} \frac{\mathbf{w}^H \Phi_f^s \mathbf{w}}{\mathbf{w}^H \Phi_f^n \mathbf{w}}, \quad (8)$$

Table 1: *The description of the training data*

Data ID	Description	Duration
1	worn (cleaned+sp)	64h×3
2	100k far-field (cleaned+sp)	39h×3
3	reverberate on 1	64h×3
4	100k far-field (cgm+mvdr)	35h
5	100k far-field (gwpe,ch1)	35h

Table 2: *Performance of different acoustic models*

Structure	Data	WER%
baseline 9-TDNN	1+2	80.28%
baseline 9-TDNN	1+2+3	79.13%
9-TDNN+1BLSTM	1+2+3	77.15%
12-TDNN-F	1+2+3	70.02%
5CNN+9-TDNN-F	1+2+3	68.72%
5CNN+9-TDNN-F	1+2+3+4	<b>68.43%</b>
Original submission [14]	-	70.49%

which can be solved by forming a generalized eigenvalue problem with  $\Phi_f^s$  and  $\Phi_f^n$ . To produce a distortionless speech signal at the beamformer output, [24] also provides several post-filtering algorithms to normalize GEV coefficients. In our experiments, we adopt *Blind Analytical Normalization* (BAN) by default.

### 3. Experiments

#### 3.1. Acoustic model

The performance of acoustic models we tuned on the development data is given in Table 2, with the description of the training data listed in Table 1. All models are trained with lattice-free maximum mutual information (LF-MMI, [29]) criterion via KALDI [30] toolkit. Mel-frequency cepstral coefficients (MFCCs) and online i-vectors are adopted as input features. In addition to the training data used in official baseline (1 + 2), we include reverberated data (3) and enhanced data (4 + 5) processed by GWPE<sup>1</sup> and CGMM-MVDR<sup>2</sup>. To simulate the reverberated audio samples in 3, we take the room impulse response (RIR) dataset released in [31] but only use the portion of *small room* because it has a similar room size as CHiME-5.

Our best configuration follows the successful practise of CNN-TDNN-F structure in our original system [14]. As can be seen in Table 2, our boosted version of TDNN-F acoustic model brings 12% absolute WER reduction compared to the official TDNN, which also surpasses our previous submitted result. In the following sections, we will mainly focus on the performance of the front-end and evaluate the results with our own acoustic model (see Table 4).

#### 3.2. Data processing

The non-overlapped segments of each speaker we used are listed in Table 3 and short segments (less than 2s) are discarded. The noise files come from non-speech intervals and a energy based VAD is used to filter out possible silence segments. Based on those processed segments and the background noise files, we

<sup>1</sup>[https://github.com/funcwj/setk/blob/master/scripts/run\\_gwpe.sh](https://github.com/funcwj/setk/blob/master/scripts/run_gwpe.sh)

<sup>2</sup>[https://github.com/funcwj/setk/blob/master/scripts/run\\_cgmm.sh](https://github.com/funcwj/setk/blob/master/scripts/run_cgmm.sh)

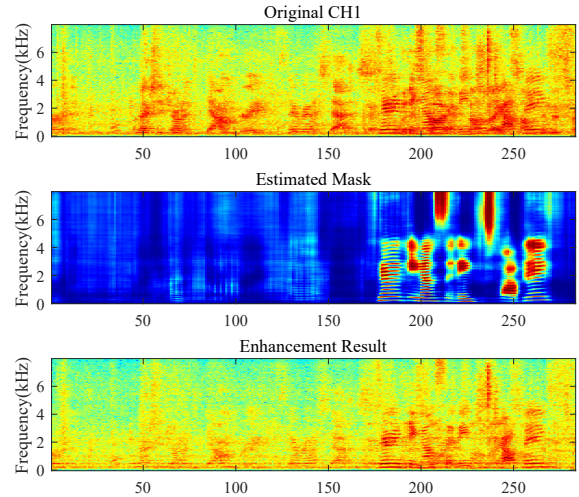


Figure 2: *A beamforming example and predicted target speaker masks. Although the interference speaker occupies most of the time in the utterance, the estimation of the target TF-mask is accurate. The last row plots the spectrogram of the enhancement output, where the interference speech is well suppressed.*

Table 3: *The number of non-overlapped segments per speaker used in data simulation on the development set*

P05	P06	P07	P08	P25	P26	P27	P28
161	251	132	108	121	78	92	169

simulate the data for speaker-dependent model training as depicted in Fig.1.

To demonstrate the effectiveness of the data processing discussed in Section 2.1, we first evaluate the ASR performance of GWPE followed by CGMM-MVDR. As can be seen in Sys-5 in Table 4, compared to the original CH1 (Sys-1), the data processing step brings 4% absolute WER reduction.

To evaluate the necessity of conducting single-channel denoising for each speaker, we simulate two sets of data (each with ~25h) depending on whether to apply OMLSA after GWPE and CGMM-MVDR, which are denoted as  $SD_A$  (without) and  $SD_B$  (with), respectively. For each set, we mix target speaker with 1 or 2 interference speakers as well as background noise randomly with SDR between 0 and 10dB and SNR between -5 and 10dB. We adopt a  $2 \times \text{TDNN} - 3 \times \text{BLSTM}$  structure with a sigmoid output layer to estimate speaker masks and use IRM as training targets. 513-dimensional log power spectrogram features are extracted as input, with utterance level cmvn applied.

In Table 4, we can see that with the processing step GWPE and CGMM-MVDR,  $SD_A$  gives 4% absolute WER reduction compared to official baseline (Sys-3) and including OMLSA as a further step yields better results, showing in Sys-8. Both models surpass our previous results without the data processing steps. To illustrate the necessity of speaker separation, we also train a denoising network (DN) in Sys-6 for comparison, which only predicts masks of speech instead of target speaker. Without the target information in the estimated masks, DN only produces a similar result as CGMM, which inspires us to focus on separation more than enhancement or denoising in CHiME5 challenge.

Table 4: WER (%) of each speaker on the development set with CNN-TDNN-F acoustic model

Sys	Input	Mask	#Models	Beamformer	P05	P06	P07	P08	P25	P26	P27	P28	Total
0	CH1-4	[14]	8	MVDR	71.57	63.46	70.13	72.53	69.77	79.18	70.06	53.57	68.66
1	CH1	×	-	×	75.21	66.74	71.77	83.64	66.73	79.38	71.09	53.39	70.71
2	GWPE-CH1	×	-	×	72.22	64.40	69.78	78.45	66.61	78.97	68.55	52.65	68.52
3	CH1-4	×	-	WDS	72.13	66.09	69.20	78.10	67.65	79.93	68.02	50.78	68.72
4	CH1-4	CGMM	-	MVDR	70.74	61.08	67.36	78.76	66.03	79.88	67.50	49.88	66.91
5	GWPE	CGMM	-	MVDR	70.59	61.15	67.08	78.25	63.94	78.85	66.34	49.69	66.36
6	GWPE	DN	1	MVDR	70.74	63.15	68.12	78.63	63.89	79.34	64.63	49.06	66.81
7	GWPE	SD <sub>A</sub>	8	MVDR	68.16	59.72	64.90	74.06	62.92	78.72	65.39	47.33	64.47
8	GWPE	SD <sub>B</sub>	8	MVDR	67.57	60.05	63.92	72.21	62.13	76.77	65.08	46.24	<b>63.75</b>
9	GWPE	SD <sub>B</sub>	8	GEV	65.95	59.55	63.57	65.85	67.34	76.93	72.62	50.93	64.43
10	GWPE	SA	1	MVDR	66.72	59.74	64.53	71.87	61.36	76.23	63.95	46.42	63.37
11	GWPE	SA	1	GEV	63.12	59.62	62.21	62.92	59.93	71.51	67.17	46.72	<b>61.31</b>
12	GWPE	SA++	1	MVDR	66.07	58.89	63.60	69.69	60.20	74.79	63.45	46.18	62.41
13	GWPE	SA++	1	PMWF-1	65.12	58.03	63.46	69.66	60.97	76.12	64.28	46.09	62.31
14	GWPE	SA++	1	GEV	62.45	58.11	61.60	61.64	57.99	69.90	65.70	46.54	<b>60.16</b>

### 3.3. Speaker-aware training

Our motivation is to train a speaker-independent target separation networks, which includes target speaker’s embeddings as auxiliary input and outputs mask estimation of the speaker. Unfortunately, the model trained on the training set can not exceed the results mentioned above. Here we apply the idea of speaker aware training on our speaker-dependent models, as the discussed in Section 2.2. In our experiments, we adopt the same network structure with the SD models, but concatenate i-vectors in the second layer of TDNN and the following BLSTM layer to bias the prediction of target masks. The i-vectors used here are extracted from non-overlapped segments shown in Section 3. During the test stage, we average the i-vectors on those utterances and get one fixed embedding for each speaker. From Table 4, SA gives similar result as SD<sub>B</sub> with MVDR beamforming, but yields a significant improvement on GEV beamformer, which brings a notable WER reduction on speaker P25 ~ P28. Compared to MVDR, GEV beamformer is more sensitive to TF-masks and may distort target speech and degrade ASR performance seriously if mask is estimated inaccurately.

Based on SA, we utilize two strategies to further improve the performance of the speaker-aware separation and denote it as SA++ in the table. The first is to initialize network with a pre-trained model on the training data, considering that the number of speakers and non-overlapped segments on the development is quite limited. Another one is to replace IRM with truncated PSM, which has been proved to be effective in monaural speech enhancement. We give an example in Fig.2. Row 2 shows the output masks of SA++ given the log power spectrogram of mixture in row 1, which masks out the interference speakers well<sup>3</sup>. We also compare GEV with other forms of beamforming (e.g. MCWF, MVDR in [28]), but the results are not further improved.

Table 5 compares the proposed system with the performance of other teams, under the circumstance that not using system combination. We get a 20% absolute WER reduction in total compared to official result and outperform most of other teams. Although it’s inferior to the USTC-iFlytek’s, our system perform separation only once and has low computational complexity and model size apparently. The details on each session and location over official AM and ours are given in Table 6. Even based on the official backend, our SD separation front-

<sup>3</sup>Samples at <https://funcwj.github.io/online-demo/page/chime5>

Table 5: Single system comparison with other teams

Team	WER (%)
USTC-iFlytek [17]	57.10
Ours	60.16
JHU [32]	62.09
Toshiba [33]	63.30
STC [23]	63.30
RWTH-Paderborn [34]	68.40
Official [13]	80.28

Table 6: WER (%) summary on the official &amp; our AM

AM	Sess	Din	Kit	Liv	Avg	Total
Baseline	S02	70.82	79.79	62.11	70.26	70.46
	S09	74.58	71.29	67.38	68.06	
Ours	S02	61.58	70.70	52.58	60.61	<b>60.16</b>
	S09	63.24	59.21	56.22	59.21	

end contributes a 10% WER reduction, which is a significant improvement on this challenging task.

## 4. Conclusions

In this work, we continue to optimize the performance of our speaker-dependent separation system submitted to CHiME-5 challenge. We utilize multi-channel dereverberation and enhancement, followed by single-channel denoising, to improve the quality of the training targets. To crack the data scarcity problem in CHiME-5, we apply the idea of speaker-aware training on our speaker-dependent models and reduce the number of the front-end models to one, while bringing significant ASR improvement. Experiments show that with well-tuned beamforming, our system improves the ASR performance from 80.28% official baseline to 70.46% in terms of WER. And tested on our own acoustic backend, our system achieves 60.16% WER on the development set, without using any fusion techniques.

## 5. Acknowledgement

This research work is supported by the National Natural Science Foundation of China (No.61571363).

## 6. References

- [1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*. IEEE, 2016, pp. 31–35.
- [2] M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.
- [3] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.
- [4] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.
- [5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.
- [6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 483–492, 2016.
- [7] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*. IEEE, 2016, pp. 196–200.
- [8] 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 Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.
- [9] Z.-Q. Wang and D. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust asr," in *ICASSP*. IEEE, 2018, pp. 5709–5713.
- [10] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.
- [11] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 1, pp. 102–111, 2017.
- [12] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 7, pp. 1492–1501, 2017.
- [13] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.
- [14] Z. Zhao, J. Wu, and L. Xie, "The nwpu system for chime-5 challenge," in *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, 2018, pp. 16–18.
- [15] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.
- [16] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.
- [17] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "The ustc-ifytek system for chime-4 challenge," pp. 36–38, 2016.
- [18] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.
- [19] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *ICASSP*. IEEE, 2016, pp. 5210–5214.
- [20] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.
- [21] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Interspeech*, 2017, pp. 2655–2659.
- [22] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voice-filter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.
- [23] I. Medennikov, I. Sorokin, A. Romanenko, D. Popov, Y. Khokhlov, T. Prisyach, N. Malkovskii, V. Bataev, S. Astapov, M. Korenevsky *et al.*, "The stc system for the chime 2018 challenge," in *The 5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018)*, *Interspeech*, 2018.
- [24] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.
- [25] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.
- [26] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech*, 2018.
- [27] L. Drude, C. Boeddeker, J. Heymann, R. Haeb-Umbach, K. Kinoshita, M. Delcroix, and T. Nakatani, "Integrating neural network based beamforming and weighted prediction error dereverberation," in *Interspeech*, 2018.
- [28] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks," in *Interspeech*, 2016, pp. 1981–1985.
- [29] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar *et al.*, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Interspeech*, 2016, pp. 2751–2755.
- [30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.
- [31] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*. IEEE, 2017, pp. 5220–5224.
- [32] N. Kanda, R. Ikeshita, S. Horiguchi, Y. Fujita, K. Nagamatsu *et al.*, "The hitachi/jhu chime-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays," in *The 5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018)*, *Interspeech*, 2018.
- [33] R. Doddipatla, T. Kagoshima, C.-T. Do, P. Petkov, C. Zorila *et al.*, "The toshiba entry to the chime 2018 challenge," in *The 5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018)*, *Interspeech*, 2018.
- [34] M. Kitzka, W. Michel, C. Boeddeker, J. Heitkaemper, T. Menne, R. Schlüter, H. Ney *et al.*, "The rwth/upb system combination for the chime 2018 workshop," in *The 5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018)*, *Interspeech*, 2018, pp. 53–57.