A TWO-STAGE SINGLE-CHANNEL SPEAKER-DEPENDENT SPEECH SEPARATION APPROACH FOR CHIME-5 CHALLENGE

Lei Sun[1], *Jun Du*[1], *Tian Gao*[12], *Yi Fang*[2], *Feng Ma*[2], *Jia Pan*[2], *Chin-Hui Lee*[3]

[1]University of Science and Technology of China, Hefei, Anhui, China
[2]iFlytek Research, Hefei, Anhui, P. R. China
[3]Georgia Institute of Technology, Atlanta, Georgia, USA

sunlei17@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

In this paper, we design a two-stage single-channel speaker-dependent speech separation approach for the CHiME-5 Challenge, targeting the problem of far-field and multi-talker conversational speech recognition in dinner party scenarios involving background noises, reverberations and overlapping speech. First, we make detailed analysis of the CHiME-5 data and observe problems of inaccurate human annotations and low-resource useable data for target speakers. Motivated by this, we conduct a first-stage speaker-dependent speech separation with a learning target for aggressive segregation to generate more and purer target speech data. Then a second-stage speaker-dependent speech separation with a new learning target is performed to obtain the final speech masks, which can be directly fed to back-end acoustic model. Compared with the official baseline, our proposed approach can yield an absolute word error rate reduction of 5.3%, namely from 81.3% to 76.0% in development test set. To the best of our knowledge, it is the first time to discuss a feasible method of single-channel speaker-dependent speech separation for such a challenging task although we make an assumption of oracle speaker diarization following the challenge rules. By integrating this crucial technique, our submitted systems achieved the first place of all four tasks in the CHiME-5 challenge.

Index Terms— CHiME-5 challenge, speaker-dependent speech separation, multiple speakers, robust speech recognition

1. INTRODUCTION

After decades of development, the field of speech techniques has reached the stage of generating practical products [1]. However, there are still problems to be solved, such as background noises, conversational style, overlapping speech and so on. Automatic speech recognition (ASR), as the basis of other speech applications, should face the challenge first. Since the inception of the CHiME [2], the challenge series have been dedicated to solving the problem of speech recognition in everyday situations. Judging from previous results, the most effective solution is quite complicated and requires the combination of multiple technologies [3], such as single- or multi-microphone enhancement and separation, robust acoustic modeling, language modeling and etc. At the same time, the rules of the game are getting closer to the most challenging "cocktail party problem" [4]. As a result, the speech data in the latest CHiME-5 challenge is directly recorded from dinner party scenario, while there are no restrictions on speaking content and style. To achieve that, all sessions are captured from twenty actual homes. Typically, every session contains four persons, and they are well known to each other to make sure the dialogues are smooth and natural. A set of six Microsoft Kinect devices are strategically placed in three kinds of locations, including kitchen, dining room, living room. From the acquired data, twenty sessions are split into training set, development test set and evaluation test set, which contains 16, 2 and 2 sessions, respectively. More details can be found in [2].

For ASR system under such realistic conditions, the big challenge lies mainly in two points. The first problem, as usual, is about far-field speech processing. Second, unlike reading-style speech, the complexity of conversational speech can greatly increase the instability of an ASR system. Besides casual contents, too many overlapping regions will weaken the discriminating ability of acoustic models. Hence, those key problems make CHiME-5 extremely challenging compared with former ones. Fortunately, the speaker segmentation information derived from human annotations is provided and could be used by following the challenge rules. In other words, it is allowed to exploit knowledge of the utterance start and end time and the utterance speaker label, namely an oracle diarization [2]. Accordingly, we would like to consider building speech separation system to effectively extract target speech from potentially noisy, reverberated and overlapped speech.

Speaking of single-channel speech separation, one mainstream kind of unsupervised approaches is based on computational auditory scene analysis (CASA) [5], which uses the psychoacoustic cues such as pitch, onset/offset, temporal continuity, harmonic structures, and modulation correlation, and segregate a voice of interest by masking the interfering sources [6]. Recently, deep learning based supervised methods have shown great performance in source separation areas. Depending on different learning targets, the methods can be divided into mask-based methods [7] and regression-based methods [8]. As a compromise, Weninger et al. [9, 10] modified the learning targets to combine both advantages of mask-based methods and regression-based methods. In [11], different learning targets has been compared in the field of speech enhancement. For multi-talker speech separation without considering noises, many researches like deep clustering (DC) [12] and attractor network (DANet) [13], focus on finding great embedding space of mixture signals, where T-F units belong to the same speaker form a cluster. Moreover, permutation invariant training (PIT) [14] is proposed to address the label permutation problem in speaker independent multi-talker speech separation problem.

However, most of those approaches could not work due to the difficulty of the CHiME-5 data. In this study, we present a novel two-stage single-channel speaker-dependent speech separation approach designed for the CHiME-5 challenge. First, we make detailed analysis of real-recorded data in CHiME-5 which is far more challenging

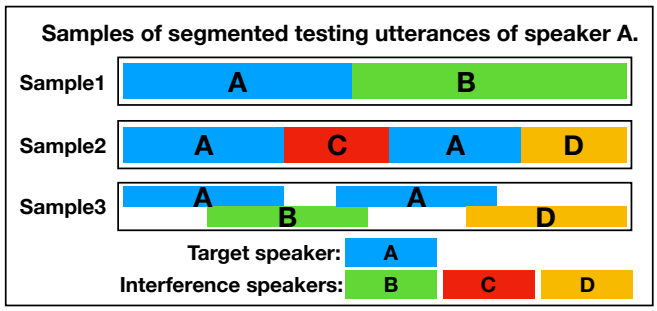

Fig. 1. An illustration of segmented utterances in one session. Silence/non-speech regions are emitted here. Speaker A is the target speaker, others are the interference speakers.

than data in previous researches, and present our motivation. Then, we will elaborate the procedures of our proposed speech separation framework for CHiME-5. Finally, we directly send the separated speech to official back-end acoustic model, and evaluate the performance in terms of word error rate (WER).

2. MOTIVATION

The CHiME-5 challenge features two tracks, we focus on single-array track where only one reference array can be used to recognize a given test utterance. Following the challenge rules, each utterance is segmented according to human annotations and the speaker identity information can be used in the testing phase.

However, the oracle speaker segmentation information is inaccurate. One reason is that the manually marked boundaries can not achieve a high solution on the frame level due to quick speaker turns. Meanwhile, overlapping speech is an unavoidable issue in conversational dialogues. After statistical analysis of official transcriptions, 97.3% of total 7,440 utterances in development test set contain overlapping regions. As illustrated in Fig. 1, suppose we have a session containing 4 speakers, namely A, B, C and D. Here the target speaker is set to speaker A, and speakers B, C, D become the interference speakers. In this way, the interference speech recognized by the ASR system will produce inevitable insertion errors. Thus, a speech separation system is necessary to extract speech from the target speaker for improving the recognition performance.

Conventionally to simulate mixed signals of multiple speakers, only non-overlapping regions can be used as source data. Details of speaker information in the CHiME-5 development set are listed in Table 1. The last two columns present the total speech duration and non-overlapping duration of each speaker, respectively. Obviously, the percentage of effective non-overlapping speech data is low. Moreover, silence/non-speech regions also take a certain percentage making the valid speech less. To fully utilize the limited data, a speaker-dependent speech separation model is an appropriate choice which can avoid the permutation problem as discussed in PIT [14]. Also, unavoidable environment noises will hurt the embedding process in approaches like DC [12] and DANet [13]. Similar to our own attempts, it is reported in [15] that all those methods failed in CHiME-5 data.

Motivated by those analyses, we propose a novel two-stage speaker-dependent speech separation framework based on several state-of-the-art methods [16, 17, 18] for the CHiME-5 challenge, which can well address the issue of low-resource non-overlapping data. To the best of our knowledge, it is the first time to discuss a feasible method of speech separation approach to address such challenging realistic data like the CHiME-5.

Table 1. Details of speaker information in CHiME-5 development set, according to oracle human transcriptions. Possible existing regions of silence and non-speech are not excluded from the duration length.

Dev Set (Session)	Speaker	Gender	Total Duration (minute)	Non-overlapping (minute)
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

3. THE PROPOSED TWO-STAGE APPROACH

As illustrated in Fig. 2, the overall diagram of our front-end via a two-stage single-channel speaker-dependent speech separation approach consists of three main modules, namely array preprocessing, first-stage speech separation, and second-stage speech separation, which are elaborated in the following subsections.

3.1. Array preprocessing

As shown in the official report [2], the CHiME-5 baseline uses a weighted delay-and-sum beamformer [19] as a default multichannel speech enhancement approach. Alternatively, we design our own multichannel preprocessing for better suppression of background noises and reverberations before speech separation. Firstly, we employ the generalized weighted prediction error (GWPE) [20] algorithm upon multiple signals of the reference array, which is commonly used as a dereverberation preprocessor. Without regard to noises, the resulted frequency-domain features are adopted by an auxiliary function based independent vector analysis (IVA) [21] method. Lastly, a noise reduction method by multi-channel post-filtering [22] is used to suppress both stationary and non-stationary background noises without distorting the speech signal components. The final output single-channel data, namely 'array preprocessing' data, is utilized as a prerequisite for subsequent operations.

3.2. The first-stage speech separation

We first use non-overlapping 'array preprocessing' data of each speaker to simulate mixed speech. To build the training set of data pairs of target speech and mixed speech, the utterances of target speaker utterances are corrupted with speech from interference speakers at several SNR levels, i.e., -5dB, 0dB, 5dB, 10dB and 15dB. Thus, each speaker can get its own training data among one session. Since the array-processed data is not clean enough, we need to design an approach to make an aggressive segregation of the target speaker, namely suppressing the interference speech as much as possible. Accordingly, we adopt the intermediate mapping (IM) as in [9, 10, 11] based on a bi-directional long short-term memory (BLSTM) model, aiming to fully utilize the advantage of mapping-based and masking-based learning targets [11] of deep models, namely log-power spectral (LPS) features [8] and ideal ratio mask

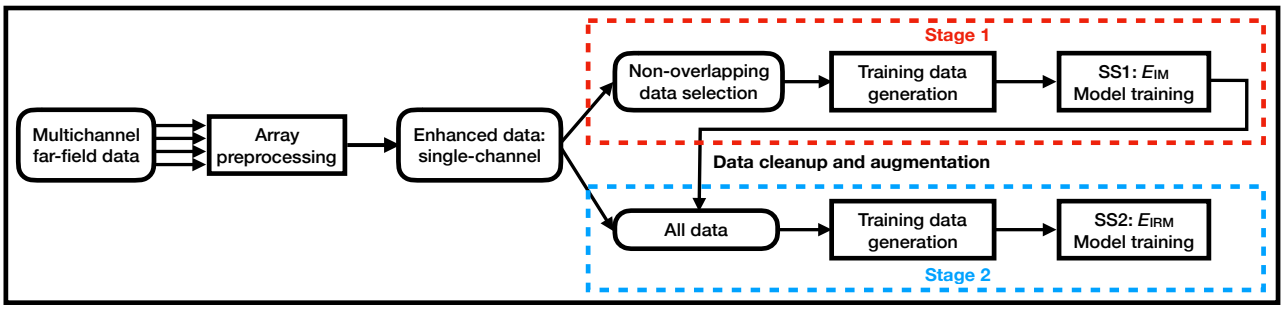

Fig. 2. The diagram of our proposed two-stage single-channel speaker-dependent speech separation system designed for CHiME-5 challenge.

(IRM) [7]. IRM ranging from 0 and 1, which is inspired by auditory masking phenomenon in audition and the exclusive allocation principle in auditory scene analysis [5], is defined as:

$$z^{\text{IRM}}(t,f) = \frac{S(t,f)}{S(t,f)+N(t,f)} \tag{1}$$

where $S(t,f)$ and $N(t,f)$ represent the power spectra of the speech and noise signals at the T-F unit (t,f), respectively. Then IM-based approach estimates the IRM by optimizing the BLSTM parameters via the minimum mean square error (MMSE) between the masked and the reference LPS features [9, 10]:

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

where $\hat{z}^{\text{IRM}}(t,f)$ is the BLSTM estimated IRM which is combined with the logarithm operation and the input noisy LPS features $x^{\text{LPS}}(t,f)$ to generate the masked LPS features. $\bar{z}^{\text{LPS}}(t,f)$ are the reference clean LPS features at the T-F unit (t,f). By using IM, it 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 speech separation model of the first stage could be generated, denoted as **SS1**.

As analyzed in Section 2, useable non-overlapping data size is small, especially for speakers like P07, P08 and P26. Such insufficient data can not satisfy the data demands for training a speaker-dependent model. Thus, we directly use SS1 models on all original data to extract the speech of a specific speaker. It cleans the signals by suppressing interference speech and augments the useable data size by including both non-overlapping and overlapping data. The speech diversity of each speaker has been enhanced. Overall, the role of first-stage speech separation is the data cleanup and augmentation as shown in Fig. 2.

3.3. The second-stage speech separation

In the second stage, data separated by SS1 models is used to simulate new training data set again with the same amount of data pairs as in the first stage. We use **SS2** to denote the separation model of the second stage. In **SS2** model training phase, we choose the original IRM as our training target of BLSTM model because it leads to better speech intelligibility and less speech distortions. It is more appropriate and stable in terms of final ASR performance. Thus, the corresponding MMSE objective function for optimizing the parameters of BLSTM is:

$$E_{\text{IRM}} = \sum_{t,f} \left(\hat{z}^{\text{IRM}}(t,f) - \bar{z}^{\text{IRM}}(t,f) \right)^2 \tag{3}$$

where $\hat{z}^{\text{IRM}}(t,f)$ and $\bar{z}^{\text{IRM}}(t,f)$ are the BLSTM estimated and the reference IRMs, respectively. In testing phase, each utterance is first processed by its corresponding SS2 speech separation model of each speaker, and then sent to back-end ASR system to generate the recognition results of the corresponding speaker.

4. EXPERIMENTS

4.1. Speech separation model training

The development set contains two separate sessions, namely S02 and S09, each session contains 4 speakers. In other words, to make speaker-dependent models, we need to make total 8 customized models in development set. As described in Section 3.2, 'array preprocessing' data was first derived by array preprocessing. Non-overlapping segments were selected and used to make 50,000 utterances of training data for SS1 models. Rather than only using the selected segments, estimated IRMs of all full sentences were generated by inferencing corresponding speaker's 'SS1' model. Data cleanup and augmentation were accomplished after recovering waveforms from masked spectral features. Next, the separated utterances were used to make another 50,000 utterances for training SS2 models. We used a two-layer BLSTM as the speech separation model for both SS1 and SS2, each direction with 512 cells. 257-dimensional LPS features were utilized here as the acoustic features to facilitate recovering waveforms, 7-frame expansion was used in the input. The computational network toolkit (CNTK) [23] was adopted for training. After separation stage, the resulting waveforms were sent to back-end.

4.2. Acoustic model training

To better illustrate the effectiveness of our proposed system, we used the official time delay neural network (TDNN) recipe [24] with lattice-free maximum mutual information (LF-MMI) training via KALDI toolkit [25]. Mel-frequency cepstral coefficients (MFCCs) and i-vectors were adopted as input features. The data used in acoustic model training was only from the official training set, both binaural data from binaural microphones and far-field data from reference microphone array. Three different levels of speed perturbation were conducted to augment data size, which are 0.9, 1.0, 1.1. Eventually, the training data size was about 310 hours. When decoding, we also used the official 3-gram language model. The whole training process was the same as in official baseline [2], with the WER of **81.3%** on the development test set.

4.3. Results

Table 2 presents comprehensive WER comparisons among different stages of our proposed speech separation system on session 02, which is half of the whole development test set. Here we fix the back-end acoustic model and evaluate different versions of processed data. Moreover, the individual results for all 4 speakers, including P05, P06, P07, P08, are also presented to show the effectiveness of our speaker-dependent strategy.

Table 2. WER comparison of official baseline using BeamformIt and different versions of processed data in S02.

S02 WER(%)	Official Baseline	Our Array Preprocessing	SS1	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**

Several observations could be made here. First, our array-preprocessing data yields slightly better results than the official baseline using the BeamformIt method [19]. Although this gain is not significant, array-preprocessing is quite important for the following speech separations to work well as we do not need to explicitly consider strong noises and reverberations in the separation stage. Next, the separated speech by SS1 reduces the WER of all speakers except P07. There are two main problems of the first-stage speech separation. One is that the training target in SS1 brings speech distortions, the other one can be ascribed to the small source data size. For instance, P07 only has about 5.6 minutes of non-overlapping data which is insufficient for training a speaker-dependent model. After data cleanup and augmentation, we can attain relatively more and purer speech for every speaker, which makes the second stage yielding better results. In the last column, SS2 reduces the average WER to 74.8%, an absolute WER reduction of 6.4% in comparison to the official baseline method, which is a quite significant gain for the single-channel separation approaches among all submitted systems in CHiME-5 challenge. By comparing results between SS1 and SS2, significant improvements appear mainly in P07 and P08, whose useable data is limited in the first-stage speech separation.

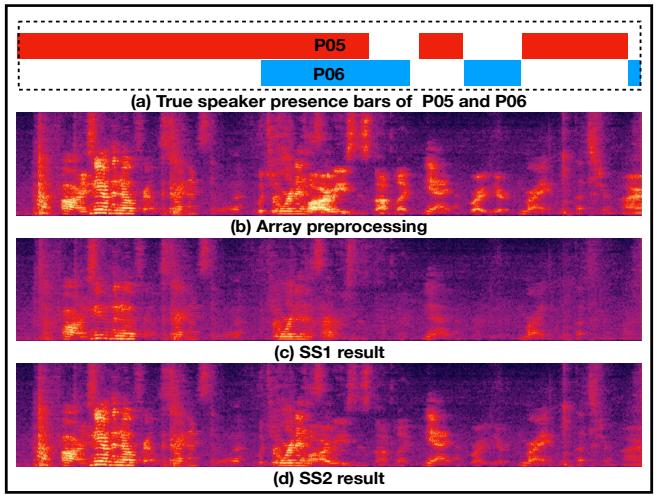

Fig. 3. An utterance of speaker P05 in session 02.

To better illustrate the effectiveness of our speech separation system, an utterance selected from session 02 is presented in Fig. 3. According to the oracle annotation, red bars indicate the target speaker 'P05' while blue bars denote the corresponding interference speaker 'P06'. Several observations can be made. First, compared with the spectrogram of 'Array preprocessing', speech processed by 'SS1' models removes most of interference part both on overlapping regions and non-overlapping regions. And it also introduces some speech distortions to target speaker. As discussed above, the spectrogram of 'SS1' fully meets our expectation of data cleanup. Due to different learning targets introduced in Section 3, the final processed speech makes trade-off between speech distortion and speech intelligibility, in consideration of ASR performance. We can observe that the power and strength of interference speech regions are largely impaired.

Overall results of development test set are listed in Table 3. Compared with S02, performance gains are relatively small in S09. On the one hand, the recording quality in S09 is worse than S02, speech sounds are imperceptible even by human auditory sensation. On the other hand, 4 speakers in S09 are all female, which are quite challenging to separate them from each other [26]. Based on 'Array preprocessing', the ASR performance has been progressively improved by two stages. According to different scenarios, results in living room are better than dining room and kitchen, partially due to less environmental noises. Overall, compared with the WER of 81.3% reported from the official baseline [2], the final results yield an absolute WER reduction of 5.3%.

Table 3. Overall WER comparison on the development test set.

Dev Set WER(%)	Session	Dining	Kitchen	Living	Ave	Overall
Array preprocessing	S02	79.4	86.9	75.5	80.3	80.2
	S09	82.6	81.1	77.2	80.1	
SS1	S02	77.8	83.7	73.3	78.0	78.5
	S09	83.6	78.9	76.6	79.3	
SS2	S02	**74.4**	**81.8**	**69.2**	**74.8**	**76.0**
	S09	**81.6**	**77.3**	**76.4**	**78.1**	

5. CONCLUSION AND FUTURE WORK

Recognition of overlapping speech is the most tough nut in the CHiME-5 challenge. We propose a two-stage speaker-dependent speech separation approach for CHiME-5 data which improves the ASR performance of official baseline system in terms of WER, from 81.3% to 76.0%. Furthermore, by integrating this unique method, our final system achieves the first place in all four tasks among all submitted systems in the CHiME-5 challenge. However, it still holds the privilege of using oracle speaker segmentation and speaker identity. In future studies, it is of great significance to continue exploring the speech separation and speech enhancement [27] under realistic conditions without oracle labels, such as a comprehensive system containing both robust speaker diarization and source separation.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No.61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No.17030901005. This work was also funded by Huawei Noah's Ark Lab.

7. 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 *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[2] 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.

[3] Y.H. Tu, J. Du, L. Sun, F. Ma, and C.H. 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, 2017.

[4] E.C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[5] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[6] M. Wu, D.L. Wang, and G.J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[7] D.L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

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

[9] F. Weninger, J.R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014, pp. 577–581.

[10] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J.R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[11] L. Sun, J. Du, L.R. Dai, and C.H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*. IEEE, 2017, pp. 136–140.

[12] J.R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[13] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 246–250.

[14] D. Yu, M. Kolbæk, Z.H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 241–245.

[15] I. Medennikov, I. Sorokin, A. Romanenko, and etc., "The STC System for the CHiME 2018 Challenge," in *spandh.dcs.shef.ac.uk/chime_workshop/programme.html*, 2018.

[16] J. Du, Y.H. Tu, Y. Xu, L.R. Dai, and C.H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 473–477.

[17] T. Gao, J. Du, L.R. Dai, and C.H. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," *Speech Communication*, vol. 95, pp. 28–39, 2017.

[18] Y. Wang and D.L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[19] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[20] 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.

[21] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 2011, pp. 189–192.

[22] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1149–1160, 2004.

[23] F. Seide and A. Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 2135–2135.

[24] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, and S. Wang, Y.and Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Interspeech*, 2016, pp. 2751–2755.

[25] 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," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[26] Y. Wang, J. Du, L.R. Dai, and C.H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1535–1546, 2017.

[27] Q. Wang, S. Wang, F. G, C.W. H, J. Lee, L. G, and C.H. Lee, "Two-Stage Enhancement of Noisy and Reverberant Microphone Array Speech for Automatic Speech Recognition Systems Trained with Only Clean Speech," *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2018.