The USTC-NELSLIP Systems for CHiME-6 Challenge

Jun Du[1], *Yan-Hui Tu*[1], *Lei Sun*[1], *Li Chai*[1], *Xin Tang*[1], *Mao-Kui He*[1], *Feng Ma*[1], *Jia Pan*[1], *Jian-Qing Gao*[1], *Dan Liu*[1], *Chin-Hui Lee*[2], *Jing-Dong Chen*[3]

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

{jundu, tuyanhui}@ustc.edu.cn

Abstract

This technical report describes our submission to the 6th CHiME Challenge. The submitted systems for CHiME-6 cover both the multiple-array speech recognition track and multiple-array diarization and recognition track. For each track, the results corresponded to Category A and Category B are reported. The main technique points of our submission include the deep learning based iterative speech separation, training data augmentation via different versions of the official training data, SNR-based array selection, front-end model fusion, acoustic model fusion. Tested on the development and eval test set, our best system takes the first place among all submitted systems in both two tasks of track 1.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

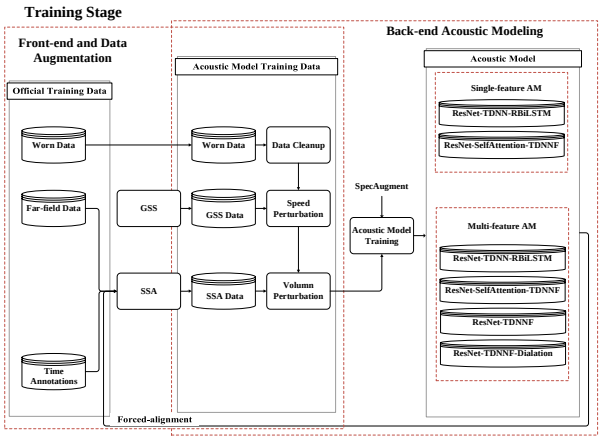

Figure 1: *An illustration of overall framework.*

1. Track1: Multiple-Array Speech Recognition Track

First of all, due to rules defined by official, systems are allowed to exploit knowledge of the utterance start and end time, the utterance speaker label and the speaker location label for track 1. It's allowed to use binaural data and far-field data in the training set.

1.1. System overview

The overall framework of track 1 is shown in Fig. 1. As we can see, it contains several main parts including multi-channel based WPE denoising, space-and-speaker-aware iterative mask estimation (SSA-IME), beamforming and acoustic model training. For the front-end, we fisrt apply a conventional multi-channel noise reduction using log-spectral amplitude [1] which is based on weighted prediction error (WPE) [2]. With the denoised data, we can build the following SSA model which is based on deep-learning techniques. After all, each method of these frond-end techniques can provide processed data of official original training data, add increase the diversity of original data. Using the final augmented data, five types of acoustic model are trained as the back-end system.

The decoding phase is divided into four successive steps, namely, beamforming initialization, SSA-based signal statistics estimation, beamforming, and recognition. First, beamformed speech is initialized and a T-F mask of test speech is obtained by cACGMM-based beamforming [3] using time annotation as initial prior values. Then, the mask estimated by our SSA model is used to improve the initial mask where the SSA model uses the features of the initial beamformed speech. And the ASR-based voice activity detection (VAD) information from the segmentation results of a recognizer with beamformed speech [4] also can be used to improve the initial mask. Next, the improved mask is used as the initial values of the cACGMM-based approach to generate the estimated mask which steers the beamforming, thereby obtaining the beamformed speech for ASR.

1.2. Front-end

1.2.1. Multi-channel preprocessing

For CHiME-6 challenge, we first utilize a multi-channle preprocessing step by traditional methods of signal process, which doesn't rely on training. It uses log-spectral amplitude [1] which is based on generalized weighted prediction error (WPE) [2]. The goal of this step is to suppress some obvious noises and output the single-channel signals for the following stage. The preprocessing is simple but important for our entire system.

1.2.2. SSA model training

In this section, we will describe the training process of the SSA model in detail. To improve the mask estimation accuracy, a neural-network-based mask estimator learned from a multi-feature concatenation data set is proposed. The beamformed STFT features, $\hat{\boldsymbol{S}}_{\mathrm{BF}}^{\mathrm{T}_i}$, are composed of the elements. Unlike conventional regression model for mask estimation, the beamformed features of four speakers are used together as the input of the BLSTM-based regression model. Specifically, $\log(|\hat{\boldsymbol{S}}_{\mathrm{BF}}^{\mathrm{T}_i}|^2)$ $(i = 1, 2, 3, 4)$ denotes the log-power spectral (LPS) features of four speakers on a whole utterance. And $\hat{\boldsymbol{\varphi}}_{\mathrm{BF}}^{\mathrm{D}_j}$ $(j = 1, 2, 3)$ denotes the inter-phase difference (IPD) between

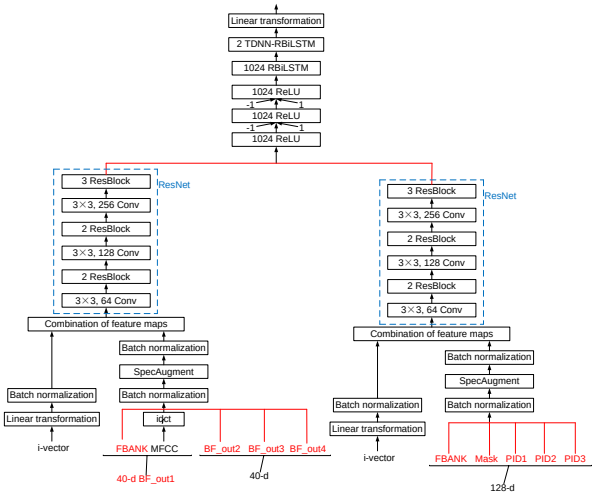

Figure 2: *An illustration of multi-feature acoustic model.*

a target speaker and three other interfering speakers on a whole utterance, which contains the spatial information between different speakers. Based on the above introduction, the BLSTM-based regression model can learn both space and speaker information at the same time. Therefore, we defined this regression model as space-and-speaker-aware (SSA) model which is also a speaker-independent speech separation model.

To train the BLSTM-based SSA model, the learning targets are used because they are calculated by beamformed features which are more reliable than the conventional masks. The optimization function of the BLSTM-based model is defined as:

$$E_{\text{SSA}} = \sum_{i=1}^{4} \sum_{t,f} \left(\hat{M}_{\text{SSA}}^{\text{T}_i}(t,f) - M_{\text{BF}}^{\text{T}_i}(t,f) \right)^2 \quad (1)$$

where $\hat{M}_{\text{SSA}}^{\text{T}_i}(t,f)$ and $M_{\text{BF}}^{\text{T}_i}(t,f)$ are the BLSTM estimated mask and the reference mask, respectively. By using E_{SSA}, the model can not only distinguish four speaker as much as possible by taking advantage of the space and speaker information but also yield robust and refined masks. After training, the one single SSA model of all four speakers can be generated. More detailed can refer to [5].

1.2.3. Beamforming

Motivated by Guided Source Separation (GSS) [6] demonstrated to be very effective for the difficult CHiME-6 dinner party recognition task, the estimated masks from SSA model introduced above are adopted to improve the a complex Angular Central Gaussian Mixture Model (cACGMM)-based [3] mask estimation. It's the first time to address five factors simultaneously in realistic conditions, namely noises, target speaker and interference speakers. The masks are adopted as the initialization state for the EM algorithm. Final masks are sent to generalized eigenvalue decomposition (GEVD) beamformer [7]. Because there are huge differences between the different channels, the channel selection [8] is adopted according to the signal to interference plus noise ratio (SINR). The mean SINR is calculated among all the channels, and only the channels with a higher SINR than the mean SINR are selected. Also the online beamforming [9] are also adopted to update the beamforming weights by continuous blocks. Finally, four beamformers are adopted, denoted as Multi_array, Selected_online, Selected_offline and Single_array. Note only the Selected_online is online, and the others are offline.

So far, the entire frond-end stage finishes and outputs the separated waveform for recognition.

1.3. Acoustic model

In the back-end, we use six different kinds of acoustic models. The features of the first two acoustic models consist of 40-dimensional mel-frequency cepstral coefficients (MFCCs) appended with 100-dimensional i-vectors being extracted on top of PCA-reduced spliced-MFCC features for speaker adaptation [10]. The features of the others four acoustic models consist of four speaker 40-dimensional FBANK, target speaker 40-dimensional MFCC, target speaker 128-dimensional FBANK, target speaker 128-dimensional mask and the three 128-dimensional inter-phase difference (IPD) features of different speakers. Fig. 2 give an example of multi-feature AM.

The CNN-TDNN-RBiLSTM architecture was proposed in [11, 12], which consists of a CNN, TDNN, and RBiLSTM [11]. This architecture is the main contribution of acoustic modeling for the Hitachi/JHU CHiME-5 system that achieved the second-best result in the CHiME-5 Challenge. ResNets [13] are popular in computer vision due to their increasing number of convolutional layers and ease of optimization, achieving a better performance in almost all the standard image recognition datasets. Accordingly, in our system, we use the ResNet to replace the simple CNN to build the acoustic model architectures with the combination with TDNN/TDNNF and RBiLSTM. More detailed can refer to [14]. They are listed as follows:

* ResNet-TDNNF (Multi-feature): 20-layer ResNet + 11-layer TDNNF
* ResNet-TDNNF-Dialation (Multi-feature): 20-layer ResNet + 12-layer TDNNF-Dialation
* ResNet-TDNN-RBiLSTM (Single/Multi-feature): 20-layer ResNet + 2-layer TDNNF-RBiLSTM
* ResNet-SelfAttention-TDNNF (Single/Multi-feature): 20-layer ResNet + 5-layer SelfAttention-TDNNF

2. Track2: Speaker Diarization and Speech Recognition

2.1. Multi-channel Speech Separation

In this section, we will describe the training process of the multi-channel speech separation (MCSS) model in detail. Considering the distribution of microphone arrays in CHiME-6 data, there are 6 microphone arrays with 4 channels for each array corresponding to $I = 6$ and $J = 4$. $\log(|X_{i,1}(t,f)|^2)$ $(i = 1, ..., I)$ denotes the log-power spectral (LPS) features of mixture from channel 1 for each array. And $\hat{\varphi}_i^{\text{D}_d}(t,f)$ $(i = 1, ..., I; d = 1, ..., J-1)$ denotes the inter-phase difference (IPD) between the channel 1 and three other channels in the i-th array. Based on the above introduction, the BLSTM-based regression model can learn the spatial information. Then, the deep model outputs multiple speakers and noise masks. Mixture is processed with weighted prediction error (WPE) [2] and cACGMM. We initialize the cACGMM parameters with prior hard masks indicating existence of speakers and noise. The soft masks of cACGMM outputs are more reliable than the conventional defined mask and taken as the learning targets. The op-

timization function of the BLSTM-based model is defined as:

$$E_{\text{MCSS}} = \sum_{l=1}^{L=4} \sum_{t,f} (\hat{M}^{\text{TS}_l}_{\text{MCSS}}(t,f) - M^{\text{TS}_l}_{\text{cACGMM}}(t,f))^2 + \sum_{t,f} (\hat{M}^{\text{N}}_{\text{MCSS}}(t,f) - M^{\text{N}}_{\text{cACGMM}}(t,f))^2 \quad (2)$$

where $\hat{M}^{*}_{\text{MCSS}}(t,f)$ and $M^{*}_{\text{cACGMM}}(t,f)$ are the BLSTM estimated masks and the reference masks of target speakers (TS) or noise (N), respectively.

2.2. Speaker Diarization

The diarization stage illustrated in Figure 3 includes the next few steps. In the beginning, we perform the CHiME-6 baseline [15] SAD on the four speakers beamformed data. Experiments show clustering all of SAD segments results in a much lower class purity. Because the front-end separation algorithm performs not very well on the overlapping speech and the beamformed speech still contains other speakers' speech. Currently, the best selection strategy is keeping the longest target speaker duration and removing the others' within a time period. Then, we consider extracting x-vectors with the most popular residual network architecture named as ResNet34. The settings for ResNet34 training are similar to [16]. ReLU activation follows each convolutional layer. We apply batch normalization technique to stabilize and speed up network convergence. The x-vectors with 512 dimensions are extracted on the selected speech. Finally, spectral clustering (SC) [17], as its name implies, making use of the spectrum (or eigenvalues) of the similarity matrix of the data, is adopted to cluster the extracted x-vector with a certain number of classes. More detailed can refer to [18].

2.3. Speech Recognition

ASR stage of Figure 3 is a description of speech recognition system for CHiME-6. All of the details can be found in [19]. We explore data augmentation approaches including GSS-based training data augmentation [20] and spectral augmentation [21]. Advanced neural network acoustic models trained according to the Lattice Free Maximum Mutual Information (LF-MMI) criterion [22] whose main modules include the deep convolutional residual network (ResNet), factorized time delay neural network (TDNNF) [23] and residual bidirectional long short-term memory (RBiLSTM) [24], realignment achieved by the acoustic model trained using the cross-entropy criterion on GMM alignments [25], long-term decoding, and lattice-level fusion of acoustic models based on Minimum Bayes Risk (MBR) decoding [26].

3. Experiments on Track 1

3.1. Front-end experiments

First of all, we present the front-end results on official baseline in single-array track. Factored Time Delay Neural Neural Network (TDNNF) recipe [27] using lattice-free maximum mutual information (LF-MMI) training, is used here. The trainining data keeps the same with official recipe in KALDI [28], which uses both binaural data and far-filed data with speech perturbation. The front-end uses the GSS enhancement refined by time annotations from ASR output [29] as a default multichannel speech enhancement approach. More details can be found in [29]. First, compared with the GSS-based approach WER of 48.43% , our implemented four versions yield comparable WERs, as listed in Fig. 4. Second, lattice fusion followed by MBR decoding is performed to combine recognition results from the four enhancements, which achieves more than 2% absolute WER reduction over the best single enhancement and yields the recognition result of 43.24% on the development set.

So far, testing data processed by those frond-end processing procedures is fixed in the rest of this paper.

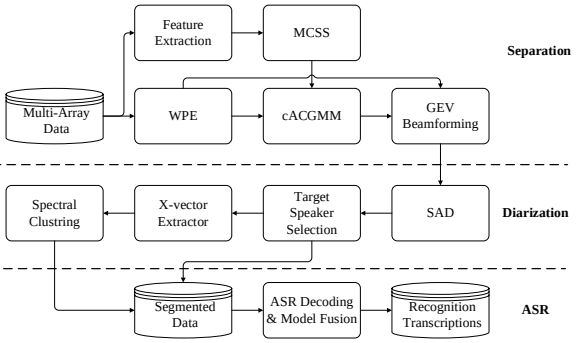

Figure 3: *Framework of our proposed CHiME-6 Track 2 system.*

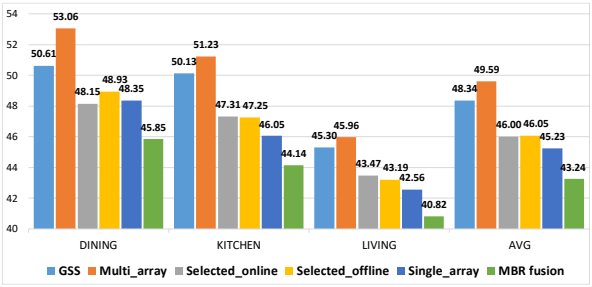

Figure 4: *WERs comparison between different front-end approaches on development set, for Category A. Note the official-AM trained with original training data is adopted.*

3.2. Acoustic models

As shown in Fig. 5, we have compared WERs of acoustic models and model ensembling on development set, for Category A. The result of official baseline acoustic model is shown in blue bar. We first use the newly fixed training data with new model architectures instead of LF-TDNN, including ResNet-TDNN-RBiLSTM, ResNet-SelfAttention-TDNNF, ResNet-TDNNF and ResNet-TDNNF-Dialation. They are all built with lattice-free maximum mutual information (LF-MMI) training method by KALDI tookit. As we can see, the ResNet-TDNNF yields better results than ResNet-TDNN.

Although the performance of four CNNs is comparable due to their big architecture similaritities, lattice fusion followed by MBR decoding is performed to combine recognition results from different models trained on the four kinds of architectures, which achieves more than about 2% absolute WER reduction over the best single system. Compared with official acoustic model, the final WER is reduced from 43.24% to 30.00%, indicating a relative reduction of 30.62%. This large improvement

Table 1: *Results of the best system tested on the development and eval test set for multiple-array speech recognition. WER (%) per session and location together with the overall WER.*

Category	Session		Dining	Kitchen	Living	Overall
A	Dev	S02	33.25	33.43	28.97	30.00
		S09	28.45	27.70	26.73	
	Eval	S01	25.48	42.94	37.94	30.88
		S21	25.56	34.87	25.86	
B	Dev	S02	34.66	34.86	29.50	30.77
		S09	29.10	27.74	27.22	
	Eval	S01	25.01	42.66	37.44	30.50
		S21	25.14	34.84	25.34	

can be attributed to both data augmentation , acoustic modeling and ensembling.

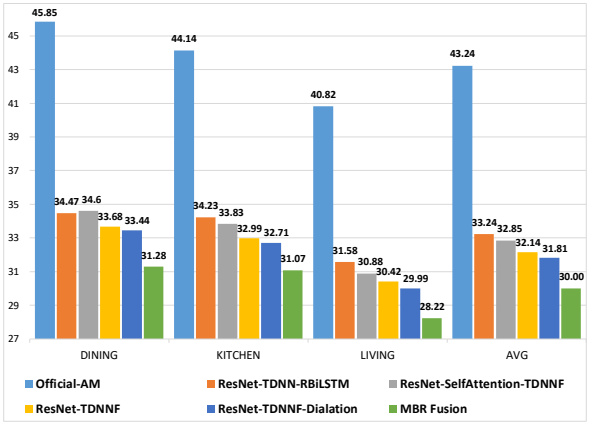

Figure 5: *WERs comparison between acoustic models and model ensembling on development set, for Category A. Note the official-AM is trained with original training data, while others are trained using newly fixed training data.*

3.3. Results summary

To summarize, in the following tables we present the performance details of our best system tuned on the development test set, with its corresponding results on the evaluation test set. The only difference between *Category A* and *Category B* is the language model, which yields slightly better results when using simple RNN-based model. After all, the final results take the first place among all submitted system in both two tasks of Track 1.

4. Experiments on Track 2

4.1. Separation model training

Multi-channel speech separation model takes the inputs of LPS features of six arrays' channel-1 (CH-1) and IPD features between CH-1 and (CH-2, CH-3, CH-4) outputs 4 speakers and noise masks for initializing the cACGMM parameters. A series of frames are extracted by a left-to-right window with 64ms frame length and 16ms frame shift and the dimension of extracted LPS and IPD is compressed from 513 to 128. All LPS and IPD features is spliced together with a 3072-dimensional $(128 * 6 + 128 * 3 * 6)$ vector as the model inputs. Simulated training data is used for fine-tuning a 2-layer BLSTM with 1024 cells for each layer. Then the BLSTM outputs with 640-dimensional $(128 * 5)$ labels including four speakers and noise masks. The soft learning target masks are obtained by performing cACGMM on the dereverberated (WPE) simulated mixture where cACGMM parameters are initialized with hard masks indicating whether the target speaker or noise exists.

ResNet x-vector extractor model is trained with the VoxCeleb data [30]. 40-dimensional filter-bank features as inputs are extracted by a left-to-right window with 25ms frame length and 10ms frame shift. The ResNet model generates a 512-dimensional x-vector with 200 feature frames each time.

Table 2: *Diarization results for CHiME-6 Track 2*

Enhancement & Diarization	Dev.		Eval.	
	DER	JER	DER	JER
BeamformIt+x-vectors+AHC	63.42%	70.83%	68.20%	72.54%
BeamformIt+RN x-vectors+SC	58.15%	59.28%	66.59%	71.72%
MCSS+RN x-vectors+SC	56.69%	58.49%	65.37%	64.25%

Table 3: *ASR results for CHiME-6 Track 2*

Enhancement	Diarization	Acoustic Model	Dev. WER	Eval. WER
BeamformIt	x-vectors+AHC	TDNN-F	84.25%	77.94%
BeamformIt	x-vectors+AHC	MBR Fusion	74.47%	74.79%
BeamformIt	RN x-vectors+SC	MBR Fusion	70.64%	71.72%
MCSS	RN x-vectors+SC	MBR Fusion	68.22%	68.48%

4.2. Results

The CHiME-6 Track 2 baseline system [15] provides a BeamformIt-based speech enhancement front-end. Then, diarization using AHC on the PLDA scoring of x-vector extracted by TDNN is conducted, followed by a two-stage decoding with the TDNN-F acoustic model. The baseline system only performs SAD, speaker diarization and ASR for the U06 array for simplicity.

As shown in Table 2, the baseline enhancement and diarization system are denoted as 'BeamformIt+x-vectors+AHC'. For comparison, both BeamformIt and our neural mask-based beamforming approach using MCSS are followed by the speaker diarization using SC on ResNet (RN) x-vector, which are denoted by 'BeamformIt+RN x-vectors+SC' and 'NMBB+RN x-vectors+SC' respectively. Obviously, diarization with SC on ResNet x-vector brings about 1.61% absolute reduction of DER on the evaluation set (68.20% vs. 66.59%). And neural mask-based beamforming reduces DER by an absolute 1.22% on the evaluation set (66.59% vs. 65.37%).

Table 3 shows the ASR results on different front-ends and back-ends. First, doing MBR decoding and fusion with multiple acoustic models reduces WER on evaluation set to 74.79% from 77.94% obtained by the baseline. Then, SC of the ResNet x-vectors on the selected SAD segments further decrease WER to 71.72%. Finally, our neural mask-based beamforming using MCSS yields the best WER 68.48%, mainly for the front-end enhancement aiming at removing the overlapping segments which have negative effects on for both speaker diarization and speech recognition.

5. References

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

[2] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *ITG 2018, Oldenburg, Germany*, 2018.

[3] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," *european signal processing conference*, pp. 1153–1157, 2016.

[4] Y. Tu, J. Du, L. Sun, F. Ma, H. Wang, J. Chen, and C. Lee, "An iterative mask estimation approach to deep learning based multi-channel speech recognition," *Speech Communication*, vol. 106, pp. 31–43, 2019.

[5] Y. Tu, J. Du, J. Pan, F. Ma, and C. Lee, "A space-and-speaker-aware iterative mask estimation approach for multi-channel speech recognition in chime-6 challenge," in *Submitted to Proc. Interspeech 202*, 2020.

[6] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[7] E. Warsitz and M. R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio Speech & Language Processing*, vol. 15, no. 5, pp. p.1529–1539, 2007.

[8] K. Wojcicki and P. C. Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2904–2913, 2012.

[9] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[10] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. Černockỳ, "ivector-based discriminative adaptation for automatic speech recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 152–157.

[11] N. Kanda, R. Ikeshita, S. Horiguchi, Y. Fujita, K. Nagamatsu, X. Wang, V. Manohar, N. E. Y. Soplin, M. Maciejewski, S.-J. Chen *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.

[12] N. Kanda, Y. Fujita, S. Horiguchi, R. Ikeshita, K. Nagamatsu, and S. Watanabe, "Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6630–6634.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] L. Chai, J. Du, D. Liu, Y. Tu, and C. Lee, "Acoustic modeling for multi-array conversational speech recognition in the chime-6 challenge," in *Submitted to Proc. Interspeech 202*, 2020.

[15] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "Chime-6 challenge:tackling multispeaker speech recognition for unsegmented recordings," 2020.

[16] H. Zeinali, S. Wang, A. Silnova, P. Matejka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *ArXiv*, vol. abs/1910.12592, 2019.

[17] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[18] Y. Tu, J. Du, L. Sun, T. Gao, Y. Fang, and C. Lee, "Deep mask-based beamforming for multi-array speaker diarization and speech recognition in the chime-6 challenge," in *Submitted to Proc. Interspeech 202*, 2020.

[19] L. Chai, J. Du, D. Liu, Y. Tu, and C.-H. Lee, "Acoustic modeling for multi-array conversational speech recognition in the chime-6 challenge," *submitted to Interspeech 2020*, 2020.

[20] C. Zorila, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in asr training and test for chime-5 dinner party transcription," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2019.

[21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH 2019*, 2019.

[22] 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 *Proc. of INTERSPEECH*, 2016, pp. 2751–2755.

[23] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Interspeech 2018*, 2018, pp. 3743–3747.

[24] N. Kanda, R. Ikeshita, S. Horiguchi, Y. Fujita, K. Nagamatsu, X. Wang, V. Manohar, N. E. Yalta Soplin, M. Maciejewski, S.-J. Chen, A. S. Subramanian, R. Li, Z. Wang, J. Naradowsky, L. P. Garcia-Perera, and G. Sell, "The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays," in *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, 2018, pp. 6–10.

[25] A.-r. Mohamed, F. Seide, D. Yu, J. Droppo, A. Stoicke, G. Zweig, and G. Penn, "Deep bi-directional recurrent networks over spectral windows," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 78–83.

[26] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, pp. 802–828, 10 2011.

[27] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[28] 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*, no. CONF. IEEE Signal Processing Society, 2011.

[29] C. Zorila, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in asr training and test for chime-5 dinner party transcription," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 47–53, 2019.

[30] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, p. 101027, 2019.