INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Acoustic Model Ensembling Using Effective Data Augmentation for CHiME-5 Challenge

Feng Ma[1], *Li Chai*[1], *Jun Du*[1], *Diyuan Liu*[2], *Zhongfu Ye*[1], *Chin-Hui Lee*[3]

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

fengma@iflytek.com, cl122@mail.ustc.edu.cn, {jundu,yezf}@ustc.edu.cn, chl@ece.gatech.edu

Abstract

CHiME-5 is a research community challenge targeting the problem of far-field and multi-talker conversational speech recognition in dinner party scenarios involving background noises, reverberations and overlapping speech. In this study, we present five different kinds of robust acoustic models which take advantages from both effective data augmentation and ensemble methods to improve the recognition performance for the CHiME-5 challenge. First, we detail the effective data augmentation for far-field scenarios, especially the far-field data simulation. Different from the conventional data simulation methods, we use a signal processing method originally developed for channel identification to estimate the room impulse responses and then simulate the far-field data. Second, we introduce the five different kinds of robust acoustic models. Finally, the effectiveness of our acoustic model ensembling strategies at the lattice level and the state posterior level are evaluated and demonstrated. Our system achieves the best performance of all four tasks among submitted systems in the CHiME-5 challenge.

Index Terms: CHiME-5 challenge, acoustic modeling, data augmentation, multi-talker conversational speech recognition

1. Introduction

Although significant advancement has been made in automatic speech recognition (ASR) after the introduction of deep neural network (DNN) based acoustic models [1, 2], far-field recognition still remains a challenging problem due to its specific factors such as reverberation, noisiness, simultaneous speech of multiple speakers, etc. A popular and effective approach to render ASR robust against adverse acoustic distortions is to train the acoustic model using a multi-condition training set that matches the final test conditions as much as possible. Besides, many approaches focusing on developing more powerful front-ends and back-ends have been proposed to handle this problem. Front-end based approaches operate on the signal or the feature, and attempt to remove the corrupting noises, interfering speakers or reverberation from the observations before recognition [3, 4]. Back-end based approaches leave the observations unchanged and instead update the acoustic/language model parameters to match the degraded speech [5, 6]. However, the performance gap between far-field and close-talking set-ups is still large as demonstrated in the AMI meeting transcription task [7] and the REVERB challenge task [8].

In recent years, several techniques have been proposed on robust acoustic modeling. Advanced DNN architectures have been developed to increase the robustness of the acoustic model, such as a novel highway long short-term memory (LSTM) network introduced in [2], a very deep convolutional neural network (DCNN) introduced and developed in [9–12] and a combined CNN, LSTM and DNN architecture that named CLDNN proposed in [13]. In addition to DNN architectures, recent studies have shown that robustness of DNN-based acoustic models largely depends on the quality of the training data [14]. Typically, using a training set that matches the final test conditions results in largest improvements in performance. However, it may not be practical to obtain such a set in many cases. A series of simulated data generation methods that deriving far-field data from existing close-talk sets via simulation were introduced in [15,16]. The quality of the model trained on derived sets depends on how good the simulation is, and how closely it captures the wide variety of the test conditions.

Recently, the latest CHiME-5 challenge [17] was held to encourage people who are interested to provide best solutions for distant multi-microphone issue in everyday home environment. Different from the previous CHiME challenges [18–21] which are restricted by the limited scale of data, single-speaker environment and fixed distance between arrays and sources, the CHiME-5 challenge provides a large-scale corpus of multi-speaker conversational speech recorded via commercially available microphone arrays in multiple realistic homes. The corpus provided by this challenge essentially congregates all possible acoustic issues in real life including noises, reverberation and overlapping speech and thus poses a large challenge to current ASR systems. Therefore, the proposed ASR systems based on this dataset have more practical value.

In this paper, we detail our back-end system for the CHiME-5 challenge. First, we introduce our data augmentation methods, especially the far-field data simulation method. Data simulation is a common way to augment the training set of the acoustic model for improving its environmental robustness. The two common data simulation methods in [15] and [16] were used by the second place [22] and the third place [23] in this challenge, respectively. Different from these conventional simulation methods, we use a signal processing method originally developed for channel identification to estimate the room impulse responses and then simulate the far-field data. Experiments demonstrate its great effectiveness. After data augmentation, the amount of the final acoustic model training set is 534 hours, which is much smaller than that used by the second place [22]. The effective data augmentation makes an important contributor to our ultimate best ASR system. Second, we introduce five kinds of acoustic models used in this challenge. All of them are conventional DNN/HMM

http://dx.doi.org/10.21437/Interspeech.2019-2601

(hidden Markov model) hybrid acoustic model. The first two are based on lattice-free maximum mutual information (LF-MMI) [24] training, including a conventional BLSTM network and CNN-TDNN-LSTM network both with the input combining 40-dimensional Mel-frequency cepstral coefficients (MFCC) features and 100-dimensional i-vector. The later three are based on frame-level cross-entropy criterion, including an improved CLDNN based on the conventional CLDNN [13], a deep fully CNN [25] and a deep fully CNN with gate on feature map all with the input combining 40-dimensional log mel-filterbank (LMFB) feature and raw waveform. Finally, acoustic model ensembling strategies at the lattice level and the state posterior level are detailed in experimental section. They achieve extremely significant ASR performance improvements. Our final ASR system achieves the best performance for all four tasks among all submitted systems.

2. Data Augmentation

As shown in Figure 1, the training data of acoustic models consists of four parts, namely 64 hours of original binaural data, 250 hours of simulated far-field data after multichannel preprocessing, 110 hours of far-field data after multichannel preprocessing and 110 hours of the final resulting far-field data after the entire front-end processing. The details for the later two training data parts processed by different front-end stages could be found in [26]. Here, we mainly present our simulation techniques to generate large-scale simulated data for augmenting the training set of acoustic models.

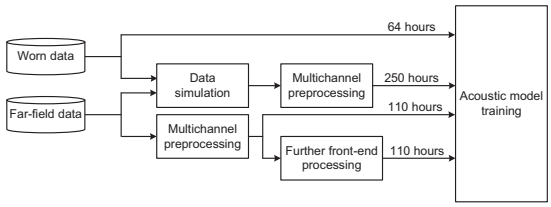

Figure 1: *An illustration of the input data for acoustic models.*

Far-field speech signals provided by CHiME-5 challenge are corrupted by noises, interfering speakers and reverberation, which lead to very poor signal quality. Our experiments reveal that for multi-condition training the ASR performance of both array and binaural microphone development set will be degraded when the far-field speech data increases to a certain amount. This implies that the far-field data with extremely poor quality is useless and even does harm to the acoustic model training. Accordingly, the available far-field data for multi-condition training becomes limited. This motivates us to simulate far-field data with adjustable quality levels using binaural signals to augment the multi-condition training set.

Assuming that there are K sound sources and M microphones, the received far-field signal at microphone m is expressed by the following equation:

$$\begin{aligned} x_m &= \sum_{k=1}^{K} s_{km}^{\text{far}} + n_m = \sum_{k=1}^{K} s_k * h_{km} + n_m \\ &= \sum_{k=1}^{K} s_k * h_{kj} * h_{kj}^{-1} * h_{km} + n_m \\ &= \sum_{k=1}^{K} s_{kj}^{\text{near}} * h_{kjm} + n_m, \end{aligned} \tag{1}$$

Algorithm 1 Simulation of far-field speech

Step1: Use oracle human transcriptions of official training set to select non-speech segments and speech segments including only one speaker. Then obtain K parallel sound source sets from the two near-field microphones and four far-field microphones denoted as $S = \{S_1, S_2, ..., S_K\}$, where $S_k = \{s_{k1}^{\text{near}}, s_{k2}^{\text{near}}, s_{k1}^{\text{far}}, s_{k2}^{\text{far}}, s_{k3}^{\text{far}}, s_{k4}^{\text{far}}\}$. Use the official baseline ASR system to eliminate recognized non-speech segments and then obtain the noise set recorded by the four far-field microphones denoted as $N = \{n_1, n_2, n_3, n_4\}$.
Step2: Use the SRO approach in [27] to calibrate the sampling rates and then estimate a set of room impulse responses using S via Eq. (3) for the four microphones denoted as $H = \{H_1, H_2, ..., H_K\}$, where $H_k = \{h_{k11}, h_{k12}, h_{k13}, h_{k14}, h_{k21}, h_{k22}, h_{k23}, h_{k24}\}$.
Step3: Use H, N and near-field data in S to simulate noisy far-field data of four microphones under the signal-to-noise ratios between 0dB and 20dB.

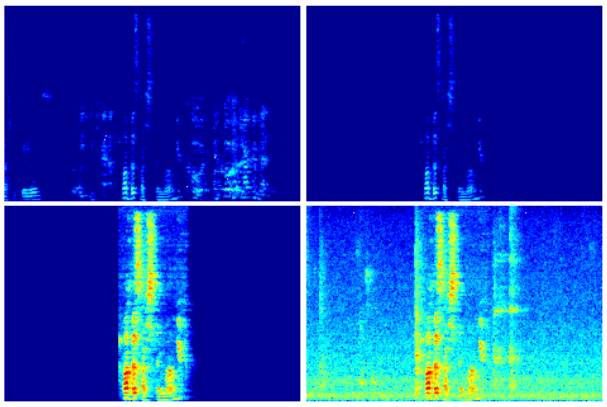

Figure 2: *An illustration of far-field data simulation: near-field speech (upper left), selected near-field speech including only one speaker (upper right), simulated far-field speech (bottom left), simulated noisy far-field speech (bottom right).*

where $s_{kj}^{\text{near}} = s_k * h_{kj}$, $h_{kjm} \triangleq h_{kj}^{-1} * h_{km}$, s_{kj}^{near} is a sound source recorded by j-th near-field microphone, s_{km}^{far} is a sound source recorded by m-th far-field microphone and n_m represents noise recorded by m-th far-field microphone. Then we draw the following conclusion

$$s_{km}^{\text{far}} = s_{kj}^{\text{near}} * h_{kjm}. \tag{2}$$

Accordingly, we can use the parallel near-field and far-field speech segments containing single speaker k extracted from official training set using oracle speaker diarization information to estimate the room impulse response h_{kjm}.

From a signal processing perspective, the difference between s_{kj}^{near} and s_{km}^{far} is mainly caused by the space transmission channel. Hence, the estimation of the room impulse response h_{kjm} could be considered as a classic channel identification problem. The closed-form solution could be obtained by applying the Wiener-Hopf equation as follows:

$$h_{kjm} = R_{\text{nn}}^{-1} R_{\text{fn}}, \tag{3}$$

where R_{nn} represents the autocorrelation matrix of the near-field signal s_{kj}^{near}, R_{fn} represents the cross-correlation matrix of the

far-field signal s_{km}^{far} and the near-field signal s_{kj}^{near}. In this work, sizes of the two correlation matrixes are set to be the number of sampling points for a 500ms speech segment. However, the far-field data and near-field data were recorded by different devices with different sampling clocks, which leads to sampling rate drift and the inaccurate estimation of channel identification. To solve this problem, the sampling rate offset (SRO) approach in [27] is adopted to align the sampling clocks of the two devices and then h_{kjm} is estimated. The detailed simulation procedures are presented in Algorithm 1 and an illustration is shown in Figure 2. Note that the simulated noisy far-field data is further processed by microphone array algorithms before being sent as the inputs of acoustic models.

3. Acoustic Models

In the CHiME-5 challenge, we use five different kinds of conventional DNN/HMM hybrid acoustic models. The first two correspond to a conventional 5-layer BLSTM network and CNN-TDNN-LSTM (2-layer CNN + 9-layer TDNN + 3-layer LSTM) network optimized by LF-MMI criterion [24]. They are trained using Kaldi Toolkit [28] with the input combining 40-dimensional MFCC features and 100-dimensional i-vector. The later three are an improved CLDNN based on the conventional CLDNN [13], 50-layer deep fully CNN [25] and 50-layer deep fully CNN with gate on feature map and all of them are optimized by frame-level cross-entropy criterion. Their inputs are 40-dimensional LMFB features and raw waveforms. Considering that the three CNN acoustic models have comparable recognition performance and there is limited space in this paper, we mainly introduce the improved version of CLDNN. Its architecture is shown in Figure 3, where each BLSTM layer has 1250 cells and a 350-unit projection layer for dimensionality reduction.

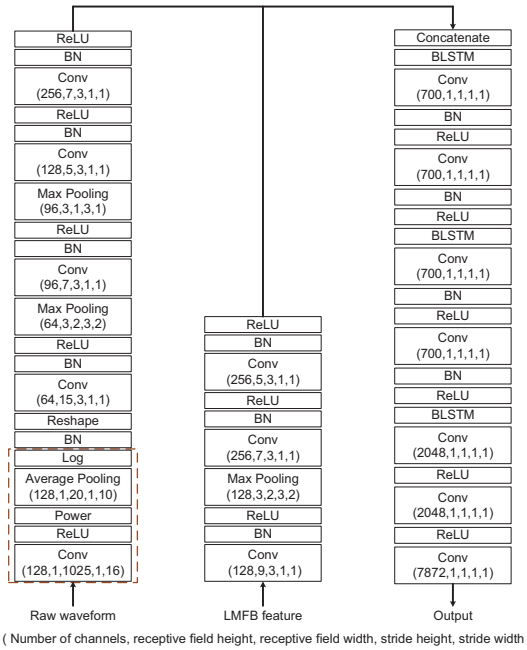

Figure 3: *Architecture of the CLDNN, where BN represents Batch Normalization.*

Learning an acoustic model directly from the raw waveform has been an active area of research. [29] is the first work which

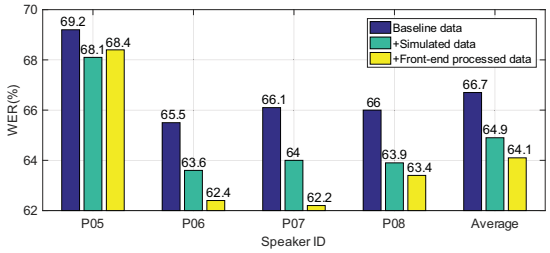

Figure 4: *WER comparison among different training sets of the official acoustic model baseline on S02 of the development set.*

is able to match the performance of raw waveform and the most popular LMFB feature on an LVCSR task using a state-of-the-art CLDNN acoustic model. Moreover, by stacking raw waveform features with LMFB features, a 3% relative reduction in word error rate (WER) is achieved. Accordingly, raw waveform features and LMFB features are both used as the inputs of our CLDNN. Different from the conventional practice of combining LMFB features and learned filterbank features from the raw waveform as input into the CLDNN, we concatenate their high-level representations extracted by some convolutional layers and then send them into the rest of the network. Experiments on the CHiME-5 challenge demonstrate that our CLDNN framework achieves relative WER reduction of 5.4% over the CLDNN that only uses LMFB features as inputs. This implies that the complementarity between the high-level representations of the learned filterbank features and LMFB features is stronger.

4. Experiments

The challenge contains two tracks, where a single-array task contains only the reference array data and a multiple-array task contains data from all six arrays placed in different positions of the home. Each track contains two separate rankings, where Ranking A focuses on acoustic robustness while Ranking B addresses all aspects of the task. More information could be found at the CHiME-5 challenge official website [1]. In this paper, we focus on the Ranking A of single-array track and explore our back-end methods on the development set consisting of two separate sessions, namely Session 02 and Session 09. Each session contains four speakers.

4.1. Evaluation of data augmentation

In baseline, only speech recorded by the binaural microphones and arrays are used for training. Thus it tends to cause mismatch between training data and evaluation data. It is necessary to augment and enhance the training data. We evaluated the effectiveness of the data augmentation methods using the official TDNN recipe with LF-MMI optimization criterion [24] as shown in Figure 4. Note that the evaluation data was preprocessed by the entire front-end before being sent to the acoustic model for recognition. Here, we removed the speech perturbation for official baseline data and then made ASR performance comparison among different training sets of the TDNN acoustic model. Clearly, in Figure 4, the baseline data is 174 hours consisting of 64 hours of original binaural data and 110 hours of far-field data. After adding 120 hours of simulated far-field data, the average WER decreased from 66.7% to

[1] http://spandh.dcs.shef.ac.uk/chime_challenge/index.html

64.9%. Furthermore, the ASR performance improvements were consistent for each speaker. Then 110 hours of far-field data processed by the entire front-end was further added. Accordingly the average WER decreased from 64.9% to 64.1%. Overall, experiments demonstrated the effectiveness of the data augmentation of both simulated far-field data and the far-field data processing by the entire front-end. Finally, the amount of simulated data was increased to 250 hours used as one part of the acoustic model training set. In the following experiments, total 534 hours of training set after data augmentation was used.

4.2. Evaluation of acoustic model ensembling

There are a large amount of overlapping speech segments in the CHiME-5, which dramatically degrade the ASR performance. To alleviate the problem, a two-stage single-channel speaker-dependent speech separation approach was proposed, where non-overlapping part of the multichannel preprocessed data of each speaker was used to build the training set of the speech separation model by mixing with interference speakers and a bi-directional long short-term memory (BLSTM) network was adopted as the model architecture. More details have been described in [30]. The original non-overlapping segments are mostly too short, the simulated mixture utterances cannot be effectively utilized by the BLSTM network to capture long-term sequential information. Accordingly, an alternative training mode is provided that the short segments are concatenated to form long segments and then used for training the speaker-dependent models.

Both the short-segment training mode and the long-segment training mode were used for training the speaker-dependent speech separation models for all the 8 speakers from the development set to provide two separation models for each speaker. Then the resulting parallel enhanced data pairs were sent to the acoustic models and thus two WER scores were obtained for each acoustic model as shown in the first two rows in Table 1. To combine the advantages of the segregated speech from the short-segment models and long-segment models, a fusion strategy via the state posterior averaging was adopted to ensemble the two WER scores for each acoustic model. The corresponding fusion results were listed in the third row in Table 1, where it was observed that consistent better recognition performance was obtained for each acoustic model. In addition, the same fusion strategy was applied to ensemble the two acoustic models trained by the Kaldi Toolkit and the three CNN acoustic models, respectively. Significant WER reductions were achieved as shown in the fourth row in Table 1. Because the number of states of the acoustic models trained by the Kaldi Toolkit and the three CNN acoustic models were different, the fusion strategy at the lattice level rather than the state posterior level was adopted to ensemble the two results obtained from the second fusion step (Fusion2). Finally, we obtained 50.62% of the average WER on the development set for Ranking A of the single-array track, which is the best result among the submitted systems of CHiME-5 challenge. Overall, our fusion strategies conducted step by step were effective and made a great contributor to our ultimate best ASR system.

4.3. Overall comparison

In Table 2, we presented the detailed ASR performance comparison among the official baseline system, the second place system and our best system on the development set for Ranking A of the single-array track. Compared to the second place system, one notable point is that our system achieved more significant recognition performance improvements for the S02 over the S09. Moreover, even poorer recognition performance occurred in the dining condition for S09. This is because for reducing the computation cost our system was mostly tuned on the S02 of the development set. On average, our best system achieved relative WER reduction of 37.7% and 10.1% compared to the official baseline and the second place system respectively. It is worth mentioning that our results approach the binaural microphone results shown in official baseline report [17], namely 47.9% of WERs.

Table 1: *WER(%) comparison among the five acoustic models and model ensembling on the development set for Ranking A of the single-array track, where LF1, LF2, CNN1, CNN2, CNN3 represent the BLSTM, CNN-TDNN-LSTM, CLDNN, 50-layer deep fully CNN and 50-layer deep fully CNN using gating mechanism.*

	LF1	LF2	CNN1	CNN2	CNN3
Long	61.42	58.49	56.26	56.74	56.25
Short	60.77	58.46	56.36	56.76	56.28
Fusion1	59.44	57.13	56.00	56.24	55.79
Fusion2	54.83		52.59		
Fusion3	50.62				

Table 2: *WER(%) comparison among the official baseline, the second place system and our system on the development set for Ranking A of the single-array track.*

System	Session	Kitchen	Dining	Living	Overall
Baseline [17]	S02	87.3	79.5	79.0	81.3
	S09	81.6	80.6	77.6	
Hitachi/JHU [22]	S02	66.4	56.8	50.9	56.4
	S09	55.9	55.9	51.6	
Ours	S02	57.8	49.4	41.8	50.6
	S09	52.4	56.8	51.4	

5. Conclusion

In this study, we detail our back-end system for the CHiME-5 challenge which scored the first place in all four tasks among submitted systems. Three main contributors for the back-end system are introduced, namely data augmentation, robust acoustic models and acoustic model ensembling. More specifically, a new idea that using a signal processing method developed for channel identification to estimate the room impulse responses and then simulate the far-filed data is provided. In addition, to utilize the complementarity between LMFB and learned filterbank from the raw waveform, a new method of combining their high-level representations is proposed. Finally, two acoustic model ensembling strategies at the lattice level and the state posterior level are evaluated and their effectiveness is demonstrated.

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, and Huawei Noah's Ark Lab.

7. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[2] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory rnns for distant speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5755–5759.

[3] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 444–451.

[4] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[5] I. Himawan, P. Motlicek, D. Imseng, B. Potard, N. Kim, and J. Lee, "Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4540–4544.

[6] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5014–5018.

[7] T. Yoshioka and M. J. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Computer Speech & Language*, vol. 31, no. 1, pp. 65–86, 2015.

[8] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[9] M. Bi, Y. Qian, and K. Yu, "Very deep convolutional neural networks for LVCSR," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for LVCSR," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4955–4959.

[11] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[12] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention." in *Interspeech*, 2016, pp. 17–21.

[13] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4580–4584.

[14] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7398–7402.

[15] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[16] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google home," 2017.

[17] 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.

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

[19] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiMEspeech separation and recognition challenge: Datasets, tasks and baselines," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 126–130.

[20] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiMEspeech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[21] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[22] 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.

[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] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI." in *Interspeech*, 2016, pp. 2751–2755.

[25] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8614–8618.

[26] L. Sun, J. Du, T. Gao, Y. Fang, F. Ma, P. Jia, and C.-H. Lee, "A speaker-dependent approach to separation of far-field multi-talker microphone array speech for front-end processing in the CHiME-5 challenge," in *IEEE Journal of Selected Topics in Signal Processing*.

[27] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 571–582, 2016.

[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," IEEE Signal Processing Society, Tech. Rep., 2011.

[29] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[30] L. Sun, J. Du, T. Gao, Y. Fang, F. Ma, J. Pan, and C.-H. Lee, "A two-stage single-channel speaker-dependent speech separation approach for chime-5 challenge," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6650–6654.