ACOUSTIC MODELING FOR MULTI-ARRAY CONVERSATIONAL SPEECH RECOGNITION IN THE CHIME-6 CHALLENGE

Li Chai[1], *Jun Du*[1], *Di-Yuan Liu*[2], *Yan-Hui Tu*[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

2021 IEEE Spoken Language Technology Workshop (SLT) | 978-1-7281-7066-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/SLT48900.2021.9383628

ABSTRACT

This paper presents our main contributions of acoustic modeling for multi-array multi-talker speech recognition in the CHiME-6 Challenge, exploring different strategies for acoustic data augmentation and neural network architectures. First, enhanced data from our front-end network preprocessing and spectral augmentation are investigated to be effective for improving speech recognition performance. Second, several neural network architectures are explored by different combinations of deep residual network (ResNet), factorized time delay neural network (TDNNF) and residual bidirectional long short-term memory (RBiLSTM). Finally, multiple acoustic models can be combined via minimum Bayes risk fusion. Compared with the official baseline acoustic model, the proposed solution can achieve a relatively word error rate reduction of 19% for the best single ASR system on the evaluation data, which is also one of main contributions to our top system for the Track 1 tasks of the CHiME-6 Challenge.

Index Terms— Data augmentation, acoustic modeling, model ensemble, multi-talker recognition, CHiME-6 Challenge

1. INTRODUCTION

Despite recent advances made in automatic speech recognition (ASR) after the introduction of deep neural network (DNN) based acoustic models [1, 2], noise, reverberation and speech from other talkers still cause severe degradations in the ASR performances. In particular, multi-talker speech recognition that aims at recognizing the individual speech sources from overlapped speech is one of the most challenging issues for ASR [3–5] due to the difficulty of separating target speech from other interfering speech signals. A classic scenario is far-field speech recognition in a daily home environment, such as a dinner party [4, 5].

The latest 6th CHiME Speech Separation and Recognition Challenge (CHiME-6) [5] was held to encourage researchers interested in providing advanced solutions for distant multi-array conversational speech recognition in everyday home environments. The CHiME series of challenges is very helpful in promoting the development of the state-of-the-art ASR for diverse environments. The CHiME-6 Challenge revisits the previous CHiME-5 Challenge [4] and further considers multi-microphone conversational speech diarization and recognition. Speech materials are the same as the previous CHiME-5 recordings except for an additional accurate array synchronization. The corpus essentially congregates all possible acoustic issues in real life including mixtures of noises, reverberation and overlapping speech and thus poses a big challenge to the current ASR technologies.

One way to improve distant multi-array conversational speech recognition is to improve the robustness of acoustic models with data augmentation techniques [6–8], better training objectives [9–13], improved acoustic model architectures [14–16], etc. More specifically, a series of data generation methods that derive far-field data from existing close-talk sets via simulation were introduced in [6, 7] to augment the training set for improving the robustness of acoustic models. In addition, SpecAugment [8] operating on the log mel spectrogram of the input audio is an effective data augmentation method and can be applied in an online manner during training. Its implementation does not require any additional data. As for the objectives for acoustic model training, new sequence training criteria such as Connectionist Temporal Classification [17] and Lattice-Free Maximum Mutual Information (LF-MMI) [12] are commonly used to improve the recognition performance and significantly outperform the cross-entropy (CE) criterion. Advanced DNN architectures have been developed to increase the robustness of acoustic models, such as a factorized time delay neural network (TDNNF) [18] that is a factored form of TDNN [19] and a combination of convolutional neural network (CNN), long short-term memory (LSTM)/bidirectional LSTM (BiLSTM)/residual BiLSTM (RBiLSTM) [20] and DNN/TDNN architecture [15, 20]. Another way to improve the robustness of acoustic models is adopting speech enhancement techniques. The speech enhancement method

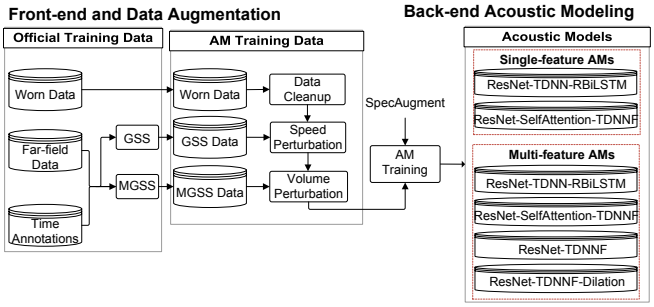

Fig. 1. The flowchart of the overall ASR system for Track 1 in the CHiME-6 Challenge.

based on spatial Gaussian mixture model (GMM) blind source separation, named Guided Source Separation (GSS) [21] is demonstrated to be very effective for the difficult CHiME-6 dinner party recognition task.

The CHiME-6 Challenge contains two tracks, Track 1: multiple-array speech recognition and Track 2: multiple-array diariazation and recognition. In this paper, we focus on Track 1: multiple-array speech recognition where annotations can be used to recognize a given test utterance. We present our main contributions of acoustic modeling for multi-array multi-talker ASR in the CHiME-6 Challenge for Track1. We explore different data augmentation methods including the SpecAugment and GSS-enhanced data. In addition, two novel architectures of neural network are designed by different combinations of deep residual networks (ResNet) [22, 23], TDNNF/TDNN, self-attention [24] and RBiLSTM, denoted as ResNet-SelfAttention-TDNNF and ResNet-TDNN-RBiLSTM, respectively. Experiments demonstrate that they significantly outperform the CNN-TDNNF [25] and CNN-TDNN-RBiLSTM [20] models commonly used in CHiME-5 Challenge, and also extremely significantly outperform the official TDNNF model. All the acoustic models are trained using the LF-MMI criterion on the GMM alignments. After realignment obtained from the ResNet-TDNN-RBiLSTM model trained using the CE criterion on the GMM alignments, better recognition results can be achieved. Finally, to combine the recognition results from acoustic models with different architectures and trained on the two kinds of alignments, lattice fusion followed by Minimum Bayes Risk (MBR) decoding [26] is performed.

2. CHIME-6 MULTIPLE-ARRAY SPEECH DATA

The CHiME-6 Challenge revisits the CHiME-5 Challenge and further considers the problem of distant multi-microphone conversational speech diarization and recognition in everyday home environments. There are some differences between the two challenges. First, the CHiME-6 Challenge uses the same recordings as the CHiME-5 Challenge except for an accurate array synchronization done by *frame-dropping* and *clock-drift*. In addition, GSS-based speech enhancement is applied to multiple arrays and achieves good recognition results. Finally, the TDNNF acoustic model trained with LF-MMI is adopted to replace the TDNN acoustic model in the baseline system of CHiME-5. The training set denoted as "worn_simu_u400k" for acoustic model training consists of worn binaural microphone recordings, a subset of 400k utterances from the array microphones and simulated data generated using the worn data and point source noises extracted from the noise regions in the CHiME-6 corpus. Data cleanup is done by the hidden Markov model (HMM)-GMM ASR system.

Fig. 1 shows our overall ASR system which achieves the best recognition performance of the Track 1 among submitted systems in the CHiME-6 Challenge. The training data consists of worn binaural microphone recordings after data cleanup, GSS-enhanced data and our proposed modified GSS (MGSS)-enhanced data, where the first two data sets are processed by a three-fold speed perturbation and all of them are processed by a volume perturbation using a random factor in [0.125, 2.0]. Moreover, the SpecAugment data augmentation technique applied for spectral perturbation of the input audio is used for augmenting the training set and demonstrated to be very effective for the CHiME-6 dinner party recognition task. In addition, there are six acoustic models used in our ASR system consisting of two single-feature acoustic models and four multi-feature acoustic models. In this paper, we only explore the two single-feature acoustic models due to the space limitation here. We will disclose more details of the multi-feature acoustic models and our proposed MGSS algorithm in our future work.

3. ACOUSTIC MODELING

3.1. Training Data Augmentation

[21] proposed a novel speech enhancement method, named GSS, which achieved a significant improvement for evaluation data in multiple array settings. Moreover, [25] demonstrated that using GSS-enhanced data in training improves ASR results significantly. Therefore, we incorporate the enhanced data after applying multi-array GSS data cleaning into the training set for acoustic model training. In this study, we select the worn microphone recordings and the GSS-enhanced data as the training set denoted as "worn_gss". The MGSS-enhanced data will be disclosed in our future work. The speed and volume perturbations as used in the CHiME-6 baseline recipe is also used to improve the robustness of the acoustic models. Experiments demonstrate that the acoustic model trained on this dataset produces better recognition results compared with that trained on the official training set "worn_simu_u400k". In addition, we found that the SpecAugment technique [8] applied for spectral perturbation is very effective for the CHiME-6 task, which

can achieve further performance gains. It is implemented during training.

3.2. Acoustic Neural Network Architectures

The baseline ASR system in the CHiME-6 Challenge uses a TDNNF which is a factored form of TDNN [19] introduced in [18]. The TDNNF has 15 layers with a hidden dimension of 1536 and a bottleneck dimension 160; each layer also has a resnet-style bypass-connection from the previous layer's output, and a "continuous dropout" schedule [27]. It is trained using the LF-MMI criterion. The acoustic feature vector consists 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 [28]. CNNs have been previously shown to improve ASR robustness [29]. Therefore, combining CNN and TDNNF layers is a promising approach to improve the baseline system. [25] has demonstrated that the CNN-TDNNF model whose architecture consists of 6 CNN layers followed by 9 TDNNF layers outperforms the TDNNF model for the CHiME-5 scenario. Accordingly, in this study, the CNN-TDNNF is explored for the CHiME-6 scenario. In addition, the CNN-TDNN-RBiLSTM architecture was proposed in [20, 30], which consists of a CNN, TDNN, and RBiLSTM [20]. 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. The model has single-channel and multi-channel input branches. We only explore the part of the single-channel input branch in the CHiME-6 task. Note that log mel-filterbank is used as the input in addition to MFCCs and i-vectors.

ResNets [31] 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. Moreover, ResNets used as acoustic model architectures can improve robustness against noisy conditions given that they are capable to more effectively model the speech variability of data [22]. Accordingly, in this study, we use the ResNet to replace the simple CNN to build the acoustic model architectures with the combination with TDNN/TDNNF and RBiLSTM, etc. We will give more details of our proprosed acoustic model architectures in Section 3.2.1 and Section 3.2.2.

3.2.1. ResNet-TDNN-RBiLSTM

Here, we propose a new acoustic model consisting of a ResNet, TDNN, and RBiLSTM. An overview of the acoustic model is depicted in Fig. 2. In this figure, the red blocks represent a CNN, the blue blocks represent a RBiLSTM proposed in [20], the green blocks represent the batch normalization [32], the yellow blocks represent a TDNN, the grey blocks represent a rectified linear unit (ReLU) activation

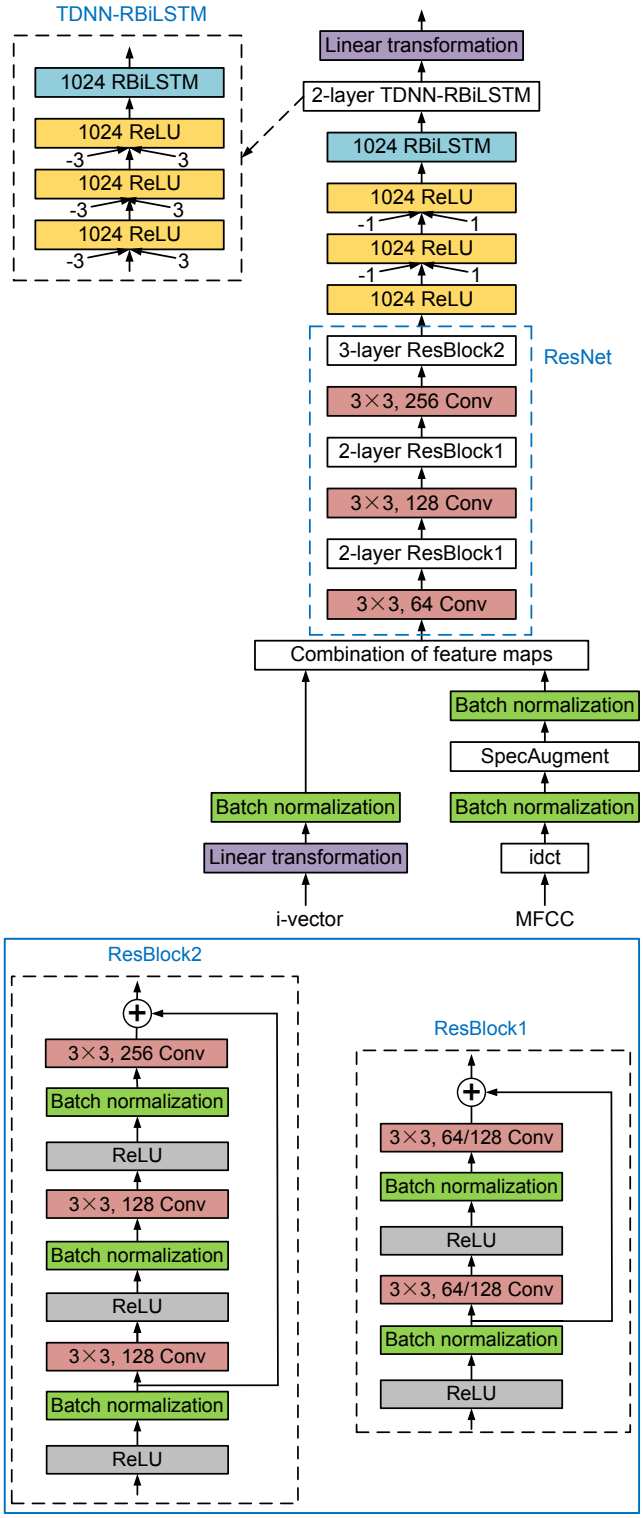

Fig. 2. The ResNet-TDNN-RBiLSTM architecture. A number with an arrow indicates a time splicing index, which forms the basis of TDNN [19]. The details of the residual block (ResBlock) in the ResNet are shown in the blue solid box.

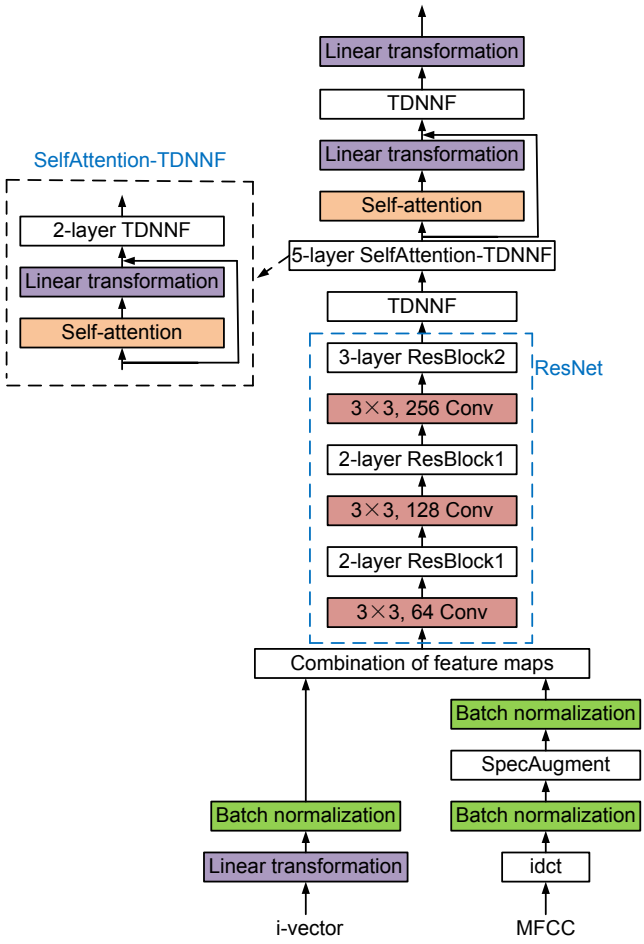

Fig. 3. The ResNet-SelfAttention-TDNNF architecture.

function, and the purple blocks represent a fully connected layer. The blue dotted box represents the ResNet consisting of convolution layers and several residual blocks [23] whose details are shown in the blue solid box. It is trained using the LF-MMI criterion and the Kaldi toolkit [33]. 40-dimensional MFCCs appended with 100-dimensional i-vectors are used as the inputs. Note that the SpecAugment in Kaldi is implemented as an nnet layer which does the augmentation on-the-fly during training. We denote this architecture as ResNet-TDNN-RBiLSTM.

3.2.2. ResNet-SelfAttention-TDNNF

We also propose a new acoustic model consisting of a ResNet, self-attention [24] and TDNNF. An overview of the acoustic model is depicted in Fig. 3. In this figure, the orange blocks represent a self-attention. The TDNNF layer has a hidden dimension of 2560 and a bottleneck dimension 512. This model is denoted as ResNet-SelfAttention-TDNNF. It is trained using the LF-MMI criterion and the Kaldi toolkit. The acoustic feature vector consists of 40-dimensional MFCCs appended with 100-dimensional i-vectors.

3.3. Realignment

In the baseline recipe, a HMM-GMM system is used as a seed system to get alignments for acoustic neural network training. The GMM stages include standard triphone-based acoustic model building with various feature transformations including linear discriminant analysis, maximum likelihood linear transformation, and feature space maximum likelihood linear regression with speaker adaptive training. In this study, we find that realignment [34] obtained from the acoustic model trained using the CE criterion on GMM alignments can bring better recognition results.

3.4. MBR Fusion

The different acoustic neural network architectures and alignments have complementarity. Lattice fusion followed by MBR decoding [26] is used to combine the recognition results from acoustic models with different architectures and trained on different alignments.

4. EXPERIMENTAL RESULTS

In this section, we report experimental results using the acoustic modeling described in Section 3. All the experiments were conducted on the Kaldi toolkit. The development data (DEV) and evaluation data (EVAL) are processed by the GSS enhancement refined using time annotations from ASR output [25]. The 3-gram language model provided by the official recipe is used in this paper.

4.1. Effect of Training Data Augmentation

We make a word error rate (WER) comparison among different training data settings for acoustic models. Row 2 in Table 1 shows recognition results of the ASR systems trained using the training set "worn_gss" consisting of the worn microphone recordings and the GSS-enhanced data as described in Section 3.1. Row 1 in Table 1 shows recognition results of the official training set "worn_simu_u400k" which is five times the amount of the training set "worn_gss". Yet the model trained on the training set "worn_gss" gives an around 1.5% absolute reduction in WER compared with that trained on the training set "worn_simu_u400k". Furthermore, when combining segments of the training set "worn_gss" whose durations are shorter than a specified minimum segment length, i.e., 2s, a 1.8%/1.3% absolute improvement in WER is obtained for the development/evaluation set as shown in row 3 of Table 1. Rows 3 and 4 show that replacing the TDNNF with the CNN-TDNNF model yields about 1.1%/1.6% absolute WER reduction for the development/evaluation set. Row 5 uses SpecAugment data augmentation which achieves more than 2% absolute WER reduction for both the development (DEV) and evaluation (EVAL) sets. For the remainder of the paper we report results of the acoustic

Table 1. WER(%) results on DEV and EVAL (in parentheses) sets for acoustic models trained with different training settings.

Training data	Data augm.	AMs	WER(%)
worn_simu_u400k	-	TDNNF	47.27 (50.34)
worn_gss	-	TDNNF	45.69 (48.86)
worn_gss_com2s	-	TDNNF	43.83 (47.52)
worn_gss_com2s	-	CNN-TDNNF	42.76 (45.89)
worn_gss_com2s	SpecAugment	CNN-TDNNF	40.70 (43.73)

Table 2. WER(%) comparison for different acoustic models.

AMs	DEV(%)	EVAL(%)
CNN-TDNNF	40.70	43.73
CNN-TDNN-RBiLSTM	40.95	43.94
ResNet-TDNN-RBiLSTM	38.86	41.37
ResNet-SelfAttention-TDNNF	38.90	42.35

models using SpecAugment data augmentation trained with the setting of "worn_gss_com2s".

4.2. Effect of Neural Network Architecture

Table 2 shows the results of acoustic models with different neural network architectures. We can make some observations. First, rows 1 and 2 show that the CNN-TDNNF acoustic model achieves slightly better recognition results than the CNN-TDNN-RBiLSTM acoustic model. Second, by comparing rows 1 and 4, it's observed that an absolute 1.8%/1.4% WER reduction is achieved for the development/evaluation set when replacing the CNN layers of CNN-TDNNF model with the ResNet shown in the blue dotted box of Fig. 2, interleaving the self-attention layers into the TDNNF layers and increasing the hidden dimension from 1536 to 2560 and the bottleneck dimension from 160 to 512 for the TDNNF. Third, the ResNet-TDNN-RBiLSTM model achieves the best results, which yields an absolute 9% WER reduction compared with the official baseline system shown in the row 1 of Table 1.

4.3. Effect of Realignment

Aforementioned acoustc models are trained using the LF-MMI criterion on the GMM alignments as used in the baseline recipe. We trained the ResNet-TDNN-RBiLSTM model using the CE criterion on the GMM alignments and then used it to do realignment. Table 3 shows the improvements of recognition results after realignment for different acoustic models. It's clearly observed that the realignment brings around 0.7%/0.5% and 0.8%/1.4% absolute reductions in WER to the ResNet-TDNN-RBiLSTM and ResNet-SelfAttention-TDNNF models on the development/evaluation set, respectively. Overall, the ResNet-TDNN-RBiLSTM model after realignment achieves the best single system result on the evaluation set, which improves the WER from 50.34% of the official baseline result to 40.81%.

Table 3. WER(%) Comparison for acoustic models trained on speech data with different alignments.

Configuration	DEV(%)	EVAL(%)
ResNet-TDNN-RBiLSTM	38.86	41.37
+ realign	38.19	**40.81**
ResNet-SelfAttention-TDNNF	38.90	42.35
+ realign	38.10	40.96
MBR fusion	34.43	37.30

4.4. Effect of MBR Fusion

Lattice fusion followed by MBR decoding is performed to combine recognition results from different models trained on the two kinds of alignments, i.e., GMM alignments and realignment. In Table 3, it's observed that MBR fusion achieves more than 3% absolute WER reduction over the best single system and yields the recognition result of 34.43%/37.30% on the development/evaluation set.

5. SUMMARY

In this paper, we present the acoustic modeling efforts in developing our CHiME-6 ASR system for Track 1 tasks. We explored data augmentation approaches for improving robustness against noisy conditions and found that SpecAugment data augmentation is effective and achieves a 2% absolute WER reduction. In addition, we investigated various acoustic neural network architectures, and yielded the best single system result using the ResNet-TDNN-RBiLSTM model trained using the LF-MMI criterion. Furthermore, after realignment obtained from the ResNet-TDNN-RBiLSTM model trained with the CE criterion on the GMM alignments, another 0.5% absolute WER reduction for the ResNet-TDNN-RBiLSTM model was achieved on the evaluation data. Finally, lattice fusion followed by MBR decoding is adopted to combine recognition results from different models trained on the two kinds of alignments, which achieved more than 3% absolute WER reduction. Our front-end acoustic signal processing effort, a key to our overall Track 1 ASR system, is described in another companion paper submitted to the same conference.

6. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N

Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Dong Yu and Li Deng, *AUTOMATIC SPEECH RECOGNITION.*, Springer, 2016.

[3] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao, and Fil Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," *Proc. Interspeech 2018*, pp. 3038–3042, 2018.

[4] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth'chime'speech separation and recognition challenge: dataset, task and baselines," *Proc. Interspeech 2018*, pp. 1561–1565, 2018.

[5] Shinji Watanabe, Michael Mandel, Jon Barker, and Emmanuel Vincent, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[6] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev 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.

[7] Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," 2017.

[8] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[9] Naoyuki Kanda, Yusuke Fujita, and Kenji Nagamatsu, "Lattice-free state-level minimum bayes risk training of acoustic models.," in *Interspeech*, 2018, vol. 2018, pp. 2923–2927.

[10] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks.," in *Interspeech*, 2013, vol. 2013, pp. 2345–2349.

[11] Hang Su, Gang Li, Dong Yu, and Frank Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6664–6668.

[12] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi.," in *Interspeech*, 2016, pp. 2751–2755.

[13] Naoyuki Kanda, Yusuke Fujita, and Kenji Nagamatsu, "Investigation of lattice-free maximum mutual information-based acoustic models with sequence-level kullback-leibler divergence," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 69–76.

[14] Vijayaditya Peddinti, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, "Low latency acoustic modeling using temporal convolution and lstms," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2017.

[15] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim 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.

[16] Dong Yu, Wayne Xiong, Jasha Droppo, Andreas Stolcke, Guoli Ye, Jinyu Li, and Geoffrey Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention.," in *Interspeech*, 2016, pp. 17–21.

[17] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[18] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *Interspeech*, 2018, pp. 3743–3747.

[19] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] Naoyuki Kanda, Rintaro Ikeshita, Shota Horiguchi, Yusuke Fujita, Kenji Nagamatsu, Xiaofei Wang, Vimal Manohar, Nelson Enrique Yalta Soplin, Matthew Maciejewski, Szu-Jui 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.

[21] Christoph Boeddeker, Jens Heitkaemper, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[22] F de-la Calle-Silos, C Peláez-Moreno, and A Gallardo-Antolın, "Deep residual networks with auditory inspired features for robust speech recognition," .

[23] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[24] Daniel Povey, Hossein Hadian, Pegah Ghahremani, Ke Li, and Sanjeev Khudanpur, "A time-restricted self-attention layer for asr," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5874–5878.

[25] Catalin Zorila, Christoph Boeddeker, Rama Doddipatla, and Reinhold 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.

[26] Haihua Xu, Daniel Povey, Lidia Mangu, and Jie Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[27] Vimal Manohar, Szu-Jui Chen, Zhiqi Wang, Yusuke Fujita, Shinji Watanabe, and Sanjeev Khudanpur, "Acoustic modeling for overlapping speech recognition: Jhu chime-5 challenge system," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6665–6669.

[28] Martin Karafiát, Lukáš Burget, Pavel Matějka, Ondřej Glembek, and Jan Černockỳ, "ivector-based discriminative adaptation for automatic speech recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 152–157.

[29] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[30] Naoyuki Kanda, Yusuke Fujita, Shota Horiguchi, Rintaro Ikeshita, Kenji Nagamatsu, and Shinji 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.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[33] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[34] Abdel-rahman Mohamed, Frank Seide, Dong Yu, Jasha Droppo, Andreas Stoicke, Geoffrey Zweig, and Gerald 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.