The ZTSpeech System for CHiME-5 Challenge: A Far-field Speech Recognition System with Front-end and Robust Back-end

Chenxing Li[1,2]*, Tieqiang Wang*[1,2]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China
[2]University of Chinese Academy of Sciences, Beijing, P.R.China

{lichenxing2015, wangtieqiang2015}@ia.ac.cn

Abstract

In this paper, we describe our ZTSpeech for two tracks of CHiME-5 challenge. For front-end, our experiments conduct the comparisons between several popular beamforming methods. Besides, we also propose a omnidirectional minimum variance distortionless response (OMVDR) followed by weighted prediction error (WPE). Furthermore, we investigate the impact of data augmentation and data combinations. For back-end, several acoustic models (AMs) with different architectures are deeply investigated. N-gram-based and recurrent neural network (RNN)-based language models (LMs) are both evaluated. For single-array track, by combining the most effective approaches, our final system can achieve 11.94% promotion on performance in evaluation set, from 73.27% to 61.33%. For multiple-array track, our final system can achieve 12.29% improvement in evaluation set, from 73.30% to 61.01%.

1. Introduction

Recently, the performance of automatic speech recognition (ASR) has been significantly improved by deep neural networks (DNNs). However, the performance of the far-field recognition is still limited, which gradually attracts more attention. Several approaches have been proposed to draw this issue, which mainly focus on developing more powerful front-ends, more robust DNN-based AM and RNN-based LM. Besides, some researchers focus on the end-to-end far-field speech recognition, which integrated fronts-ends and back-ends under one jointly-trained framework.

Referring to front-ends, beamforming is the most popular choice. Specifically, weighted delay and sum (WDAS) [1], minimum variance distortionless response (MVDR) [2], parameterized multi-channel wiener filter (PMWF) [3] and generalized sidelobe canceller (GSC) [4] are commonly deployed. These methods are designed under different criteria, which represents different degrees of the trade-offs between distortion and noise reduction. Recently, data-driven-based masking approaches use time-frequency masks to estimate spatial correlation matrix. Technically, complex Gaussian mixture models (cGMM) [5] and network-based methods [6, 7, 8] have reported the state-of-the-art performance. Traditional methods, such as WPE [9, 10] and DNN-based methods [11, 12], are widely utilized. Speech enhancement [13, 14, 15] and speech separation [16, 17] also provide effective solutions.

For back-ends, DNN-based AMs have achieved the state-of-the-art performance in speech recognition (DNN-HMM-based AMs [18, 19, 20] and time-delayed neural network (TDNN) with lattice-free maximum mutual information training (LF-MMI) [21]). Attention-based [22, 23] and connectionist temporal classification-based [24, 25, 26] end-to-end methods gradually attract more attention. The performance of LMs also has been improved by RNN [27, 28]. On the contrary, some researchers focus on the end-to-end fashion, which fused the front end and back end integrally. Some methods use a stronger DNN-based AM to process the raw multi-channel waveforms [29, 30], and some focus on jointly training speech enhancement and AM [31, 32, 33, 34].

CHiME challenges [35, 36, 37, 38] provide an excellent platform to evaluate the performance of signal enhancement and noise-robust AMs for ASR systems. However, the previous challenges are restricted by the limited scale of data, single-speaker environment and fixed distance between arrays and source. CHiME-5 [39] provides a large-scale corpus of real multi-speaker conversational speech in multiple places. This dataset is derived from everyday scenario, and the proposed systems based on this dataset have more practical value.

Our goal is to build a system for far-field multi-channel speech recognition, which involves front-end and back-end techniques. Our contributions are as follows: (1). We evaluate the performance of classical beamforming methods on CHiME-5 dataset. Simultaneously, OMVDR-WPE is proposed. (2). We explore how the performance varies to different combinations and augmentation of our data. (3). We incorporated LSTM and BLSTM into LF-MMI TDNN to explore the impact of different AMs on performance. (4). The role of different LMs is also investigated. We evaluate the performance via Word Error Rate (WER). For single-array track, Our OMVDR-WPE achieves 0.89% improvement compared with WDAS. Compared with baseline, experimental results show that our ZTSpeech achieves 9.92% improvement in development set, from 81.07% to 71.15%, and 11.94% in evaluation set, from 73.27% to 61.33%. For multiple-array track, compared with baseline, experimental results show that our ZTSpeech achieves 8.85% improvement in development set, from 82.73% to 73.88%, and 12.29% in evaluation set, from 73.30% to 61.01%.

The rest of this paper is organized as follows. Section 2 introduces the system and describes the algorithms in this paper. Experimental results for single array track are presented in Section 3. Section 4 details the experimental results of multiple-array track. Finally, Section 5 provides the conclusion.

2. System Overview

The proposed ZTSpeech consists of 2 parts: front-end and back-end. Each processing step is detailed in the following sections.

2.1. Front end

2.1.1. Omnidirectional beamforming

The traditional MVDR is designed to choose the coefficients of the filter which can minimize the output power. It has the constraint that the desired speech signal is not affected. MVDR

problem for choosing the weights is written as:

$$\min_{W} E||W^H X||^2, \ s.t. \ W^H d = 1, \tag{1}$$

where W denotes the filter, X is input signal and d is steering vector. Solved by Lagrange multipliers, W comes from:

$$W = \frac{\Phi_{NN}^{-1} d}{d^H \Phi_{NN}^{-1} d}, \tag{2}$$

where Φ_{NN}^{-1} is the noise correlation matrix, and H denotes conjugate transposition. The performance of MVDR relies heavily on the estimation of Φ_{NN}^{-1} and d. If a segment disturbed by other speakers, traditional methods will give wrong directions in some frames. For network-based mask estimation methods, a large scale of parallel dataset is required to train the network. But there is no clean speech matched with noisy one practically. OMVDR calculates W for all directions and provides multiple enhanced speech. Speech with the highest energy is regarded enhanced. When ambient noise is weaker than the speaker's voice, the dominant speech can be enhanced despite the short-term loud noise and the human interference. When positions keep stable without overlap, this method can separate speech successfully. OMVDR enhances speech in fixed direction, which avoids the inaccurate direction estimation.

In this experiment, since speech is collected by linear arrays, we choose 37 directions of arrival which distributed from 0 degrees to 180 degrees with 5 degrees step. The speech among 37 enhanced segments with the highest energy is considered to be the speech we need.

2.1.2. WPE-based speech dereverberation

WPE uses an autoregressive generative model for the acoustic transfer functions (ATFs) and models the spectral coefficients of the desired speech signal using a Gaussian distribution. Dereverberation is then performed by maximum likelihood (ML) estimation of all unknown model parameters. In an enclosed place, the reverberant speech signal captured by M microphones are typically modeled in the short-time Fourier transform (STFT) domain as:

$$x_{t,f}^m = \sum_{l=0}^{L_h-1} (h_{l,f}^m) s_{t-l,f} + e_{t,f}^m, \tag{3}$$

where $h_{l,f}^m$ models the ATF between the speech source and m-th microphone in STFT domain. L_h denotes the length of ATF and H denotes the complex conjugate operator. The additive term $e_{t,f}^m$ jointly represents modeling errors and the additive noise signal. The formula can be rewritten as:

$$x_{t,f}^m = d_{t,f}^m + \sum_{l=D}^{L_h-1} (h_{l,f}^m) s_{t-l,f} + e_{t,f}^m, \tag{4}$$

where $d_{t,f}^m$ is composed of the anechoic speech and early reflections at the m-th microphone and D corresponds to the duration of the early reflections. For simplification, the signal observed at the first microphone ($m = 1$) can be written in:

$$x_{t,f}^1 = d_{t,f} + \sum_{m=1}^{M} (g_f^m) x_{t-D,f}^m, \tag{5}$$

and the dereverberated signal can be estimated as:

$$d_{t,f} = x_{t,f}^1 - \sum_{m=1}^{M} (g_f^m) x_{t-D,f}^m, \tag{6}$$

Therefore, dereverberation can be performed by estimating the regression vectors g_f^m and calculating an estimate of the desired speech signal $d_{t,f}$.

2.2. Back end

2.2.1. Acoustic Model

The baseline framework uses an advanced LF-MMI-based TDNN. In our experiment, we integrate long short-term memory neural network (LSTM) and its bi-directional version (BLSTM) into TDNN. And LF-MMI-based TDNN with different configurations are investigated. Specifically, TDNN-a has 9 TDNN layers with 512 nodes per layer, which is the same as baseline. TDNN-b has 11 TDNN layers with 1280 nodes per layer. And one linear layer with 256 nodes is added between every two TDNN layers. TDNN-c has 11 TDNN layers with 1536 nodes per layer. And two linear layers with 256 nodes are added between every two TDNN layers. LSTM-TDNN-a has 6 TDNN layers with 700 nodes per layer followed by 3 LSTM layers with 700 nodes per layer. LSTM-TDNN-b has the same structure with TDNN-c but with 4 extra LSTM layers. BLSTM-TDNN-a has 3 TDNN layers with 1024 nodes per layer followed by 3 BLSTM layers with 1024 nodes per layer.

2.2.2. Language Model

Firstly, several Good Turning-based, Kneser-Ney-based and Max Entropy-based 3-gram, 4-gram and 5-gram LMs are trained. The LMs with the minimum perplexity (PPL) are chosen and the search graphs are created by these LMs. The graphs are then rescored by RNN-based and LSTM-based LMs. Specifically, RNN-LM-a has 1 layer with 30 nodes. LSTM-LM-a has 2 LSTM layers with 200 nodes per layer. LSTM-LM-b has 2 LSTM layers with 400 nodes per layer.

2.3. Experimental Setup

In our study, development set contributes to controlling the learning rates and evaluating different models. The final results are all evaluated on evaluation set. Speech signal is conveyed via frames. For each frame, acoustic features are generated based on 80-dimensional log-mel filterbank features and 3-dimensional pitch features [40]. The alignments are generated by a pre-trained GMM-HMM system. LMs are trained on transcription texts of the training set and trained by SRILM [41] and Tensorflow [42]. AMs are trained by Kaldi [43].

3. Experimental Results on Single-array-based Speech Recognition

3.1. Speech Enhancement

In this section, several beamforming methods [44] have been applied to enhancing the data. For comparison, AM is trained via baseline script and keeps fixed. The training data is unenhanced while the development set is enhanced. The experimental results are shown in Table 1.

Table 1 tells that cGMM-based methods produce worse results. The execution order of WPE and beamforming methods also has an effect on the performance. Multi-channel WPE may degrades speech quality and has a bad affects on subsequent MVDR. Superdirective MVDR (SMVDR) does not suppress white noise sufficiently, and the noise covariance matrix cannot be estimated in real time. OMVDR-WPE achieves best results with 0.89% improvement.

Table 1: *Comparison of beamforming methods in WER (%)*

System	Dev Set (%)
WDAS	81.07
GSC	80.79
cGMM-MVDR	88.95
cGMM-PMWF	85.51
WPE-SMVDR	87.20
SMVDR-WPE	83.43
OMVDR-WPE	80.18

3.2. Data Selection and Augmentation

In baseline, only speech recorded by the binary 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 and investigate the effect.

Firstly, we explore whether enhance training data is meaningful or not. In this experiment, training data is enhanced by BeamformIt [1]. Secondly, we investigate the impact of data augmentation. And 100000, 300000, 500000 utterances are used to train the acoustic model respectively. The experimental results are shown in Table 2.

Table 2: *Comparison of data augmentation in WER (%)*

System	Data Combinations	Data Size	Dev Set (%)
Baseline	Original	100k	81.07
System1	Enhanced	300k	79.44
System2	Original+Enhanced	300k	79.65
System3	Original+Enhanced	500k	79.90

Compared with baseline, system 1 promotes performance by 1.63%. This owes to the larger train set, which means more complex conversation scenarios and acoustic information can be modeled by AM. At the same time, due to the training data and the evaluation data are matched, the performance is further improved. Compared with system 1, the performance of system 2 and 3 degrades. This is may be caused by the random initialization, which imports fluctuations in model performance. At the same time, it also shows that the performance is basically saturated in a certain amount of data.

3.3. Acoustic Model

According to the results in section 3.1 and 3.2, we conducted two training datasets. Data 1 is the same as the system 1 in section 3.2. For data 2, training data consists of the binary-microphone close-talk speech, WDAS-based enhanced speech (500k) and OMVDR-WPE-based enhanced speech (50k). Two development sets are conducted which enhanced by BeamformIt and OMVDR-WPE respectively. We evaluate several LF-MMI-based TDNN and LSTM-TDNN AMs with different structures. The results are shown in the following Table 3.

For data 1, the performance of WDAS-based dev set is better than OMVDR-WPE-based dev set. This is because data 1 is enhanced by WDAS, which matches with training data. Among AMs, TDNN-c is the best. For data 2, the performance gap between WDAS-based and OMVDR-WPE-based dev set becomes smaller. Adding OMVDR-WPE-enhanced data to training set can effectively improve the performance of OMVDR-WPE-based dev set. Due to time constraint, we have not enhanced all training set by OMVDR-WPE. We believe that the performance of OMVDR-WPE-based dev set can be improved if the training set is fully enhanced by OMVDR-WPE.

Table 3: *Comparison of different AMs in WER (%)*

Data	System	Dev Set (%)	
		WDAS	OMVDR-WPE
Data 1	TDNN-a	79.44	79.87
	TDNN-b	73.59	75.79
	TDNN-c	71.81	74.37
	LSTM-TDNN-a	77.58	81.36
	LSTM-TDNN-b	74.50	76.58
	BLSTM-TDNN-a	78.36	84.05
Data 2	TDNN-a	79.90	80.13
	TDNN-c	73.29	73.94

3.4. Language Model

Based on the best AM, TDNN-c, we explore the impact of LMs. First, we explore the system performance under different N-gram LMs. Several 3-gram, 4-gram and 5-gram LMs are trained. Max-entropy-based LMs achieve the minimum PPL, which are utilized in the following experiments. The experimental results are shown as follows:

Table 4: *Comparison of LMs in WER (%)*

System	PPL	Dev Set (%)
3-gram	154.5547	71.77
4-gram	154.7304	71.81
5-gram	155.1294	71.66
3-gram+RNN-LM	–	71.36
3-gram+LSTM-LM-a	–	71.18
3-gram+LSTM-LM-b	–	71.15

From Table 4, 3-gram-based LM achieves the minimum PPL and well WER. For ranking B, we use RNN-based LMs to rescore the 3-gram LM. The experimental results are shown in Table 4.

In brief, for ranking A, our system has WDAS-based front-end, TDNN-based AM and 3-gram-based LM. Compared with baseline, this system achieves 9.41% WER improvement in development set, from 81.07% to 71.66%, and 11.26% in evaluation set, from 73.27% to 62.01%. Our best system has WDAS-based front-end, TDNN-based AM and LSTM-based LM. Compared with baseline, this system achieves 9.92% WER improvement in development set, from 81.07% to 71.15%, and 11.94% in evaluation set, from 73.27% to 62.01%. Thus, our best result for single-array track is detailed in Table 5.

4. Experimental Results on Multiple-array-based Far-field Speech Recognition

4.1. Speech Enhancement

In this experiment, we use the same beamforming methods as in section 3.1. At the same time, the AM is trained via baseline script. The experimental results are shown in Table 6.

In multiple-array track, Table 6 shows that cGMM-based

Table 5: *Results for the best system. WER (%) per session and location together with the overall WER.*

Rank	Session		K.	D.	L.	Overall
Rank A	Dev	S02	80.62	71.91	68.04	**71.66**
		S09	71.10	69.14	66.48	
	Eval	S01	68.72	54.90	73.51	**62.01**
		S21	66.21	55.09	59.06	
Rank B	Dev	S02	80.48	71.36	67.75	**71.15**
		S09	69.84	69.11	65.46	
	Eval	S01	68.76	53.90	73.56	**61.33**
		S21	65.55	54.07	58.05	

Table 6: *Comparison of beamforming methods in WER (%)*

System	Dev Set (%)
WDAS	82.73
GSC	82.35
cGMM-MVDR	83.04
cGMM-PMWF	86.11
MVDR-WPE	83.18

methods still perform poorer than WDAS. GSC achieves 0.38% improvement. For simplicity, in the following experiments, the multi-channel data is enhanced by BeamformIt.

4.2. Data Selection and Augmentation

Mismatch between training/evaluation data causes performance degrades. In this section, we explore the impact of data augmentation and combinations. Similar to section 3.2, we augment training data by enhancing train set. And we select 100000, 300000, 500000 utterances to train AM respectively. The experimental results are shown in Table 7.

Table 7: *Comparison of data augmentation in WER (%)*

System	Data Combinations	Data Size	Dev Set (%)
Baseline	Original	100k	82.73
System1	Enhanced	300k	81.44
System2	Original+Enhanced	300k	81.62
System3	Original+Enhanced	500k	81.71

By adding enhanced data to training set, the performance is improved. Model performance is further improved as the amount of data increases. Compared with baseline, system 1 achieves the best performance with 1.29% WER improvement.

4.3. Acoustic Model

We conducted one training set in multiple-array case. The data combines binary-microphone close-talk speech and WDAS-based enhanced speech (300k). We evaluate several LF-MMI-based TDNN and LSTM-TDNN AMs with different structures. The results are shown in Table 8. Here, TDNN-c achieves the best results again. It gains 6.77% WER improvement compared with TDNN-a.

4.4. Language Model

Based on the best AM, TDNN-c, we explore the impact of different LMs. First, we explore the system performance under

Table 8: *Comparison of different AMs in WER (%)*

System	Dev Set (%)
TDNN-a	81.44
TDNN-b	76.13
TDNN–c	74.67
LSTM-TDNN-a	80.77
LSTM-TDNN-b	80.77
BLSTM-TDNN-a	83.69

Table 9: *Comparison of LMs in WER (%)*

System	PPL	Dev Set (%)
3-gram	154.5547	74.67
4-gram	154.7304	74.69
5-gram	155.1294	74.75
3-gram+RNN-LM	–	74.27
3-gram+LSTM-LM-a	–	73.94
3-gram+LSTM-LM-b	–	73.88

different N-gram LMs. Max entropy-based 3-gram, 4-gram and 5-gram LMs has the minimum PPL among all LMs. The experimental results are shown in Table 9:

In Table 9, 3-gram-based LM has the minimum PPL and WER, which is used for ranking A. For ranking B, we use RNN-based LMs to rescore 3-gram LM. The experimental results are shown in Table 9.

In brief, for ranking A, our system has WDAS-based front-end, TDNN-based AM and 3-gram-based LM. Compared with baseline, this system achieves 8.06% WER improvement in development set, from 82.73% to 74.67%, and 11.53% in evaluation set, from 73.30% to 61.77%. Our best system has WDAS-based front-end, TDNN-based AM and LSTM-based LM. Compared with baseline, this system achieves 8.85% WER improvement in development set, from 82.73% to 73.88%, and 12.29% in evaluation set, from 73.30% to 61.01%. Thus, our best result for multiple-array track is detailed in Table 10.

Table 10: *Results for the best system. WER (%) per session and location together with the overall WER.*

Rank	Session		K.	D.	L.	Overall
Rank A	Dev	S02	79.82	73.33	72.72	**74.67**
		S09	73.33	72.53	74.01	
	Eval	S01	67.99	54.64	73.10	**61.77**
		S21	65.84	54.44	59.63	
Rank B	Dev	S02	79.45	73.48	73.00	**73.88**
		S09	70.66	69.99	72.16	
	Eval	S01	67.72	53.61	72.91	**61.01**
		S21	65.13	53.74	58.45	

5. Conclusion and Discussion

In this paper, we introduce ZTSpeech system for CHiME-5 challenge. By using fixed AM, our proposed OMVDR achieves 0.89% WER improvement compared with WDAS. Afterwards, the performance of the system is further improved by data augmentation and enhancement. Our final system can achieve 11.94% performance improvement for single-array track and

12.29% for multi-array track.

In CHiME-5, a lot of speech segments are interfered by other speakers. At the same time, because speakers do not face to arrays when talking. The speech received by arrays may not come from direct paths, which degrades the performance of source direction of arrival. For front-end, we have tried various methods. Classical beamforming methods do not perform well. DNN-based beamforming does not utilized because parallel corpus is not available. We also experiment with single-channel and multi-channel-based unsupervised speech enhancement. Due to time constraint, we do not fine tune models, and the performance fails to exceed the baseline. We will try to generate parallel dataset by using room impulse response and try DNN-based approaches. At the same time, we will continue to explore unsupervised speech enhancement, which have more practical values. When using the same AM, our OMVDR-WPE performs better than WDAS. However, we do not use this method in the subsequent experiments because we do not enhance all train set by OMVDR-WPE. We will then try to enhance all training set to further investigate this method.

For back-end, when the expressiveness of AM is powerful enough, the shortcomings of the front-end can be compensated to some extent. We only tried LF-MMI-based TDNN models. In the near future, we will try end-to-end methods.

6. References

[1] 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.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[3] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

[4] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[5] 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, pp. 5210–5214.

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 196–200.

[7] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks." in *INTERSPEECH*, 2016, pp. 1981–1985.

[8] Y. Zhou and Y. Qian, "Robust mask estimation by integrating neural network-based and clustering-based approaches for adaptive acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 85–88.

[10] ——, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[11] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, 2017.

[12] C. Li, T. Wang, S. Xu, and B. Xu, "Single-channel speech dereverberation via generative adversarial training," in *INTERSPEECH*, 2018.

[13] 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*, 2015, pp. 91–99.

[14] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *INTERSPEECH*, 2017.

[15] Z. Chen, Y. Huang, J. Li, and Y. Gong, "Improving mask learning based speech enhancement system with restoration layers and residual connection," in *INTERSPEECH*, 2017.

[16] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," *INTERSPEECH*, 2017.

[17] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[18] A. Graves, *Supervised sequence labelling with recurrent neural networks*. Springer, 2012.

[19] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," *Computer Science*, pp. 338–342, 2014.

[20] T. Sercu and V. Goel, "Advances in very deep convolutional neural networks for LVCSR," *INTERSPEECH*, 2016.

[21] 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.

[22] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[23] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 4945–4949.

[24] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep rnn models and WFST-based decoding," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 167–174.

[25] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[26] H. Liu, Z. Zhu, X. Li, and S. Satheesh, "Gram-CTC: Automatic unit selection and target decomposition for sequence labelling," *International Conference on Machine Learning*, 2017.

[27] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Annual Conference of the International Speech Communication Association*, 2010.

[28] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Annual Conference of the International Speech Communication Association*, 2012.

[29] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4624–4628.

[30] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani *et al.*, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 30–36.

[31] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5075–5079.

[32] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition." in *INTERSPEECH*, 2016, pp. 1976–1980.

[33] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5325–5329.

[34] B. Liu, S. Nie, Y. Zhang, D. Ke, S. Liang, and W. Liu, "Boosting noise robustness of acoustic model via deep adversarial training," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[35] 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.

[36] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The 2nd 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 126–130.

[37] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The 3rd 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 504–511.

[38] 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.

[39] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The 5th 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," *INTERSPEECH*, 2018.

[40] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 2494–2498.

[41] A. Stolcke, "SRILM-An extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[42] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[43] 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 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[44] H. Xiang, B. Wang, and Z. Ou, "The THU-SPMI CHiME-4 system: Lightweight design with advanced multichannel processing, feature enhancement, and language modeling," in *CHiME-4 Workshop*, 2016.