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

Chenxing Li, Shuang Xu, Tieqiang Wang, Bo Xu
Institute of Automation, Chinese Academy of Sciences
lichenxing2015@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 9.92% promotion on performance 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, our final system can achieve 8.85% improvement in development set, from 82.73% to 73.88%.

Introduction

In recent years, the performance of automatic speech recognition (ASR) has been significantly improved due to the success of deep neural networks. However, the recognition performance under far-field cases is still limited, which gradually attracts more and more attention.

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:

- Classical beamforming methods are evaluated on CHiME-5 dataset. Besides, OMVDR-WPE is proposed.
- We explore how the performance varies to different combinations and augmentation of our data.
- We incorporated LSTM and BLSTM into LF-MMI TDNN to explore the impact of different AMs on performance.
- The role of different LMs is also investigated.

System Overview

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.
- OMVDR calculates W for all directions and provides multiple enhanced speech. The speech with the highest energy is selected as the final enhanced speech.
- In this experiment, 37 directions of arrival are selected, which distributed from 0 degrees to 180 degrees with 5 degrees step.

WPE-based speech dereverberation

WPE uses an autoregressive generative model for the acoustic transfer functions and models the spectral coefficients of the desired speech signal using a Gaussian distribution. Dereverberation is then performed by maximum likelihood 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:

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

Dereverberated signal can be estimated as:

$$d_{t,f} = x_{t,f}^{1} - \sum_{m=1}^{M} (g_{f}^{m}) x_{t-D,f}^{m}. \quad (2)$$

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

Acoustic Model

- LF-MMI-based TDNN is utilized in this experiment.
- LSTM and BLSTM are integrated into TDNN.
- 3 TDNN-based and 3 LSTM-TDNN-based AMs have been conducted.

Language Model

- Several Good Turning-based, Kneser-Ney-based and Max Entropy-based 3-gram, 4-gram and 5-gram LMs are conducted.
- LM with the minimum PPL is rescored by RNN-based and LSTM-based LMs.

Experimental Setup

- Acoustic features are generated based on 80-dimensional log-mel filterbank features and 3-dimensional pitch features.
- The alignments are generated by a pre-trained GMM-HMM system.
- LMs are trained on transcription texts of the training set.

ZTSpeech for Far-field Speech Recognition

Speech Enhancement

- Several popular beamforming methods have been applied to enhancing data.
- AM is trained via baseline script and keeps fixed.
- Training data is unenhanced while the development set is enhanced.

System	Dev Set (%)	
	S-array	M-array
WDAS	81.07	82.73
GSC	80.79	82.35
cGMM-MVDR	88.95	83.04
cGMM-PMWF	85.51	86.11
WPE-SMVDR	87.20	–
SMVDR-WPE	83.43	–
OMVDR-WPE	80.18	83.18

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

- For single-array track, OMVDR-WPE achieves the best results with 0.89% improvement.

Data Selection and Augmentation

- BeamformIt is applied to enhancing training data. (OMVDR-WPE is omitted in this section.)
- Impact of data augmentation is evaluated.

System	Data Combinations	Data Size	Dev Set (%)	
			S-array	M-array
Baseline	Original	100k	81.07	82.73
System1	Enhanced	300k	79.44	81.44
System2	Original+Enhanced	300k	79.65	81.62
System3	Original+Enhanced	500k	79.90	81.71

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

- Larger train set introduces more complex conversation scenarios and acoustic information, which can be modeled by AM.
- Due to the training data and the development data are matched, the performance is further improved.

Acoustic Model

- Two training datasets are conducted. Data 1 consists of original and WDAS-based enhanced data. Compared with Data 1, Data 2 has additional OMVDR-WPE-based enhanced data.
- Several LF-MMI-based TDNN and LSTM-TDNN AMs with different structures are applied.

Data	System	Dev Set (%)		
		S-array		M-array
		WDAS	OMVDR-WPE	WDAS
Data 1	TDNN-a	79.44	79.87	81.44
	TDNN-b	73.59	75.79	76.13
	TDNN-c	**71.81**	74.37	**74.67**
	LSTM-TDNN-a	77.58	81.36	80.77
	LSTM-TDNN-b	74.50	76.58	75.92
	BLSTM-TDNN-a	78.36	84.05	83.69
Data 2	TDNN-a	79.90	80.13	–
	TDNN-c	73.29	73.94	–

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

- TDNN-c achieves the best results, which is selected for the following experiments.

Language Model

- System performance under different N-gram LMs is explored.
- RNN-based LMs are used to rescore the 3-gram LM.
- Only WDAS-based deveopment set is evaluated.

System	PPL	Dev Set (%)	
		S-array	M-array
3-gram	154.5547	71.77	74.67
4-gram	154.7304	71.81	74.69
5-gram	155.1294	71.66	74.75
3-gram+RNN-LM	–	71.36	74.27
3-gram+LSTM-LM-a	–	71.18	73.94
3-gram+LSTM-LM-b	–	**71.15**	**73.88**

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

- Max-entropy-based LMs achieve lower PPL than Good Turning-based and Kneser-Ney-based LMs.
- Max-entropy-based 3-gram LM achieves the minimum PPL. And LM rescored by LSTM-LM-b achieves the best performance.

Detailed Results for the best system.

Track	Rank	Session		K.	D.	L.	Overall
S-array	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	
M-array	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	

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

Conclusions

We introduce ZTSpeech system for CHiME-5 challenge. By using fixed AM, OMVDR-WPE achieves 0.89% WER improvement compared with WDAS. The performance of the system is further improved by data augmentation and enhancement. By combining the most effective AM and LM, for single-array track, our final system can achieve 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, our final system can achieve 8.85% improvement in development set, from 82.73% to 73.88%.

Forthcoming Research

For front-end, classical beamforming methods do not perform well. DNN-based beamforming is omitted 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. We will continue to explore unsupervised speech enhancement, which have more practical values.

References

[1] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The 5th 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines. *INTERSPEECH*, 2018.

[2] Jacob Benesty, Jingdong Chen, and Yiteng Huang. *Microphone array signal processing*, volume 1. Springer Science & Business Media, 2008.

[3] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Annual Conference of the International Speech Communication Association*, 2010.

[4] 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*, pages 2751–2755, 2016.