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

Chenxing Li\textsuperscript{1,2}, Tieqiang Wang\textsuperscript{1,2}

\textsuperscript{1}Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China
\textsuperscript{2}University of Chinese Academy of Sciences, Beijing, P.R.China
{lchenxing2015, 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)\cite{1}, minimum variance distortionless response (MVDR)\cite{2}, parameterized multi-channel wiener filter (PMWF)\cite{3} and generalized sidelobe canceller (GSC)\cite{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)\cite{5} and network-based methods\cite{6, 7, 8} have reported the state-of-the-art performance. Traditional methods, such as WPE\cite{9, 10} and DNN-based methods\cite{11, 12}, are widely utilized. Speech enhancement\cite{13, 14, 15} and speech separation\cite{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\cite{18, 19, 20} and time-delayed neural network (TDNN) with lattice-free maximum mutual information training (LF-MMI)\cite{21}). Attention-based\cite{22, 23} and connectionist temporal classification-based\cite{24, 25, 26} end-to-end methods gradually attract more attention. The performance of LMs also has been improved by RNN\cite{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\cite{29, 30}, and some focus on jointly training speech enhancement and AM\cite{31, 32, 33, 34}.

CHiME challenges\cite{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\cite{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:

\[ \min_W E[|W^H X|^2], \text{ s.t. } W^H d = 1, \]  

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

\[ W = \frac{\Phi_{MN}^{-1} d}{d^H \Phi_{NN}^{-1} d}, \]  

where \( \Phi_{MN} \) is the noise correlation matrix, and \( H \) denotes conjugate transposition. The performance of MVDR relies heavily on the estimation of \( \Phi_{MN}^{-1} \) 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 as 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:

\[ x_{t,f}^m = \sum_{l=0}^{L_s-1} (h_{l,f}^m) s_{t-l,f} + e_{t,f}^m, \]  

where \( h_{l,f}^m \) models the ATF between the speech source and \( m \)-th microphone in STFT domain. \( L_s \) 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:

\[ x_{t,f}^m = d_{t,f}^m + \sum_{l=0}^{L_s-1} (h_{l,f}^m) s_{t-l,f} + e_{t,f}^m, \]  

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:

\[ x_{t,f}^1 = d_{t,f} + \sum_{m=1}^{M} (g_{f}^m) x_{t-D,f}^m, \]  

and the dereverberated signal can be estimated as:

\[ d_{t,f} = x_{t,f}^1 - \sum_{m=1}^{M} (g_{f}^m) x_{t-D,f}^m, \]

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 rate 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 degrade speech quality and has a bad affect 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 (%)

<table>
<thead>
<tr>
<th>System</th>
<th>Dev Set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WDAS</td>
<td>81.07</td>
</tr>
<tr>
<td>GSC</td>
<td>80.79</td>
</tr>
<tr>
<td>cGMM-MVDR</td>
<td>88.95</td>
</tr>
<tr>
<td>cGMM-PMWF</td>
<td>85.51</td>
</tr>
<tr>
<td>WPE-SMVDR</td>
<td>87.20</td>
</tr>
<tr>
<td>SMVDR-WPE</td>
<td>83.43</td>
</tr>
<tr>
<td>OMVDR-WPE</td>
<td>80.18</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>System</th>
<th>Data Combinations</th>
<th>Data Size</th>
<th>Dev Set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Original</td>
<td>100k</td>
<td>81.07</td>
</tr>
<tr>
<td>System1</td>
<td>Original+Enhanced</td>
<td>300k</td>
<td>79.44</td>
</tr>
<tr>
<td>System2</td>
<td>Original</td>
<td>100k</td>
<td>79.65</td>
</tr>
<tr>
<td>System3</td>
<td>Original+Enhanced</td>
<td>500k</td>
<td>79.90</td>
</tr>
</tbody>
</table>

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 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.

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

<table>
<thead>
<tr>
<th>Data</th>
<th>System</th>
<th>Dev Set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDNN-a</td>
<td>79.44</td>
<td>79.87</td>
</tr>
<tr>
<td>TDNN-b</td>
<td>73.59</td>
<td>75.79</td>
</tr>
<tr>
<td>TDNN-c</td>
<td>71.81</td>
<td>74.37</td>
</tr>
<tr>
<td>LSTM-TDNn-a</td>
<td>77.58</td>
<td>81.36</td>
</tr>
<tr>
<td>LSTM-TDNn-b</td>
<td>74.50</td>
<td>76.58</td>
</tr>
<tr>
<td>BLSTM-TDNn-a</td>
<td>78.36</td>
<td>84.05</td>
</tr>
</tbody>
</table>

Table 4: Comparison of LMs in WER (%)

<table>
<thead>
<tr>
<th>System</th>
<th>PPL</th>
<th>Dev Set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-gram</td>
<td>154.5547</td>
<td>71.77</td>
</tr>
<tr>
<td>4-gram</td>
<td>154.7304</td>
<td>71.81</td>
</tr>
<tr>
<td>5-gram</td>
<td>155.1294</td>
<td>71.66</td>
</tr>
<tr>
<td>3-gram+RNN-LM</td>
<td>–</td>
<td>71.36</td>
</tr>
<tr>
<td>3-gram+LSTM-LM-a</td>
<td>–</td>
<td>71.18</td>
</tr>
<tr>
<td>3-gram+LSTM-LM-b</td>
<td>–</td>
<td>71.15</td>
</tr>
</tbody>
</table>

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...
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 (%)

<table>
<thead>
<tr>
<th>System</th>
<th>Data Combinations</th>
<th>Data Size</th>
<th>Dev Set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Original</td>
<td>100k</td>
<td>82.73</td>
</tr>
<tr>
<td>System1 Enhanced</td>
<td>300k</td>
<td>81.44</td>
<td></td>
</tr>
<tr>
<td>System2 Original+Enhanced</td>
<td>300k</td>
<td>81.62</td>
<td></td>
</tr>
<tr>
<td>System3 Original+Enhanced</td>
<td>500k</td>
<td>81.71</td>
<td></td>
</tr>
</tbody>
</table>

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 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:

Table 9: Comparison of LMs in WER (%)

<table>
<thead>
<tr>
<th>System</th>
<th>PPL</th>
<th>Dev Set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-gram</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>4-gram</td>
<td>154</td>
<td></td>
</tr>
<tr>
<td>5-gram</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>3-gram+RNN-LM</td>
<td>3-gram+LSTM-LM-a</td>
<td></td>
</tr>
<tr>
<td>3-gram+LSTM-LM-b</td>
<td>3.43</td>
<td></td>
</tr>
<tr>
<td>3-gram+LSTM-LM-a</td>
<td>73.94</td>
<td></td>
</tr>
<tr>
<td>3-gram+LSTM-LM-b</td>
<td>73.88</td>
<td></td>
</tr>
</tbody>
</table>

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 multi-array track is detailed in Table 10.

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Session</th>
<th>K.</th>
<th>D.</th>
<th>L.</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Dev</td>
<td>S02</td>
<td>S09</td>
<td></td>
<td>74.67</td>
</tr>
<tr>
<td></td>
<td>Eval</td>
<td>S01</td>
<td>S09</td>
<td></td>
<td>61.77</td>
</tr>
<tr>
<td>B</td>
<td>Dev</td>
<td>S02</td>
<td>S09</td>
<td></td>
<td>73.88</td>
</tr>
<tr>
<td></td>
<td>Eval</td>
<td>S01</td>
<td>S09</td>
<td></td>
<td>61.01</td>
</tr>
</tbody>
</table>

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. Afterward, 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 utilize 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


