The OPPO System for CHiME-6 Challenge

Xiaoming Ren, Huifeng Zhu, Liuwei Wei, Linju Yang, Ming Yu, Chenxing Li, Dong Wei, Jie Hao

Beijing OPPO telecommunications corp., ltd., Beijing, China

renxiaoming@oppo.com, zhuhuifeng@oppo.com, weiliuwei@oppo.com, yanglinju@oppo.com, yuming@oppo.com, lichenxing1@oppo.com, weidong@oppo.com, haojie@oppo.com

Abstract

This paper describes our system and experimental results for the 6th CHiME Challenge. We participate in Track1(ASR only) on Category A and B. Category A is system based on conventional acoustic modeling and official language modeling. The outputs of the acoustic model must remain frame-level tied phonetic (senone) targets and the lexicon and language model must not be changed compared to the conventional ASR baseline. Category B is all other systems. Our system mainly include data preparation, frontend procesing, acoustic modeling, lattice rescoring with RNN Language Model(RNNLM) and system combination. The frontend employs the baseline Guided Source Separation(GSS) [1]. For backend, we use TDNN-F and CNN-TDNNF [2] acoustic models, the systems employs a combination of 8 acoustic models, and finally apply Minimum Bayes Risk(MBR) [3] decoding for multiple lattices of different acoustic models. Comparing with the offical baseline system, our system can get 20.44% and 18.07% relative Word Error Rate(WER) reduction on the dev and eval sets respectively.

1. Introduction

Significant progress in Automatic Speech Recognition (ASR) area is made in recent years. Many ASR tasks have been intensively studied, and human parity level is achieved or even outperform for some of them [4]. However, there are still many challenges for researchers in ASR [5]. In particular, multitalker speech recognition is one of the most difficult tasks for speech recognition [6, 7] because of the difficulty of separating the target speech signal from other interfering speech signals. One example is meeting speech recognition, where it is known that the WERs are still around 30% [8] even with state-of-the-art speech recognizers. Another example is Distant Speech Recognition (DSR) in a daily home environment, such as a dinner party [7], which will be useful for developing intelligent home devices. DSR specific factors such as reverberation, noisiness, overlap speech of several speakers, etc. degrade ASR system performance drastically. To push the boundary of the state-of-the-art ASR for the complicated noisy environments, the CHiME challenge has been held every one or two years [7, 9, 10, 11, 12, 13].

CHiME-6 targets the problem of distant microphone conversational speech recognition in everyday home environments. The main features of the CHiME-6 challenge are:

- simultaneous recordings from multiple microphone arrays;
- real conversation, i.e. talkers speaking in a relaxed and unscripted fashion;
- a range of room acoustics from 20 different homes each with two or three separate recording areas;
- real domestic noise backgrounds, e.g., kitchen appliances, air conditioning, movement, etc.

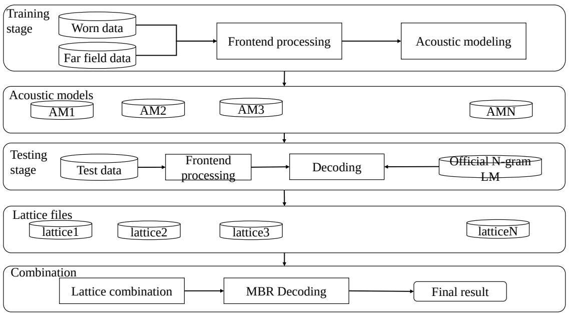

Figure 1: *Framework of system.*

Details on the challenge can be found in [13].

Our system focus on Track1(ASR only) with conventional acoustic model, which include data preparation, frontend, acoustic modeling, language modeling and system combination with MBR decoding. Figure 1 shows the framework of the submission system. With the proposed system, we finally achieve 41.18% and 42.02% WER on the dev and eval sets respectively. The rest of paper is organized as follows, section 2 describes the system in detail. The details of our expermental evaluation are given in section 3.

2. System Description

The overall framework of our system constain data preparation, frontend processing, acoustic modeling, language modeling and decoding, which is described in detail as follows:

2.1. Data Preparation

2.1.1. CHiME-6 Corpus

The dataset for CHiME-6 challenge is same as the dataset for CHiME-5 challenge. The dataset is made up of the recording of 20 separate dinner parties taking place in real homes. Recordings are made in kitchen, dining and living room areas with eachphase lasting for a minimum of 30 mins. Each dinner party has 4 participants. Each party has been recorded with a set of 6 Microsoft Kinect devices and in-ear binaural microphones(worn data). Each Kinect device has a linear array of 4 sample-synchronised microphones. The data is split into trainig, development, and evaluation. These in-ear microphone signals are considered as close talk, and they are only used in training and development.

The training set consist of 16 parties with 32 speakers in total. The number of utterances in the training set is 79, 967 adding up to around 41 hours. Development set has 2 parties with 8 speakers and 7, 437 utterances with nearly 4.5 hours of audio. Similarly 2 parties with 8 speakers and 11, 027 utterances comprising of 5.1 hours of audio is used as the evaluation set.

2.1.2. Data augmentation

To simulate the reverberation conditions, we apply randomly generated impulse responses simulated by the image method by following the small and middle sized room settings in [14]. We also randomly add non-speech region extracted from microphone array training data in order to simulate the noisy condition.

2.1.3. Training data

For the training data, comparing to official baseline, in addition we clean up and augment the data on the following aspects :

- For the worn(L+R) microphone training data, realign original utterance segmentation using ASR model
- We apply only speed perturbation(x3) [15] for the traning data without the volume perturbation
- Clean up the training data by filtering out segments which are less than 1 second
- Remove some noises which can be recognized as words from noises used in Room Impuse Responses(RIR) convolution

With the above data cleanup and data augmentation methods, we obtain about 1400 hours of data as the final training set, which contains the following dataset:

- The realigned worn(L+R)training data
- The far field data enhanced by GSS module
- The worn data and enhanced far field data both convolved with RIRs
- The augmented previous three datasets by speed perturbation

2.2. Frontend processing

Figure 2 shows the framework of the GSS enhancement system. GSS enhancement is a blind source separation technique originally proposed in [1] to solve the speaker overlap problem in CHiME-5. Given a mixture of reverberated speech, GSS aims to separate the sources using a traditional signal processing approach. To separate the different sources, GSS apply the complex Angular Central Gaussian Mixture Model (cACGMM) [16].

To avoid the permutation problem and to simplify the estimation of the model parameters, GSS exploite the time annotations provided by the challenge organizers, which indicates when a particular speaker is active. These source activity patterns guide the estimation of the mixture model parameters and avoid the need to solve the frequency permutation and the global speaker permutation problem.

Masks estimated from the GSS output, are used for beamforming and/or mask-based source extraction. As beamformer we employ the Minimum Variance Distortionless Response (MVDR) beamformer with Blind Analytic Normalization (BAN) [17, 18, 19].

Temporal context [20] also plays an important role in the GSS. Experiments have shown that a large context of 10 or 15 seconds left and right of the considered segment improves the mixture model estimation performance significantly.

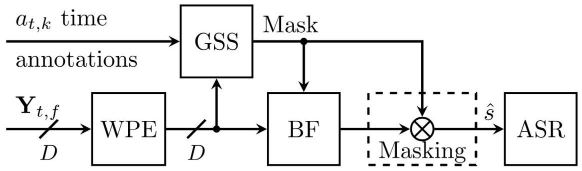

Figure 2: *The front-end consists out of WPE, a spacial mixture model that uses time annotations (GSS), beamforming and masking.*

GSS is performed with the first and last microphones of each array, and estimation performance is better than using them all.

Compared to the official baseline setup which apply GSS in testing stage , we apply the GSS not only in testing stage but also in training stage.

2.3. Acoustic modeling

For acoustic model training, we use two different kinds of acoustic model structures based on lattice-free maximum mutual informaton (LF-MMI) [21] training. They are TDNN-F network and CNN-TDNN-F network with the {40, 80}-dimenstion MFCC and 100-dimenstion online ivector. We train various acoustic models with different parameters and all the acoustic models are trained using Kaldi [22] toolkit. To determine the set of systems which should be combined for optimal performance, a greedy search procedure is employed. They are listed as follows:

- CNN-TDNN-F{1, 2, 3, 4}: GSS module with {10, 15} context, 40-dim MFCC, 6-layer CNN + 19-layer TDNN-F with bottleneck-dim = 512, NUM-PDFS = {2500, 3500}, RIR augment
- CNN-TDNN-F{5, 6}: GSS module with {10, 15} context, 40-dim MFCC, 6-layer CNN + 19-layer TDNN-F with bottleneck-dim = 768, NUM-PDFS = 3500, RIR augment
- CNN-TDNN-F7: GSS module with 15 context, 80-dim MFCC, 6-layer CNN + 19-layer TDNN-F with bottleneck-dim = 512, NUM-PDFS = 3500, RIR augment
- TDNN-F8: GSS module with 15 context, 40-dim MFCC, 25-layer TDNN-F with bottleneck-dim = 512, NUM-PDFS = 2500, RIR augment

2.4. Language modeling

We train recurrent neural network language models (RNNLM) by using the official transcription of training data. We prepare two 2-layer LSTM-based models with projection to rescore the lattice. We submit results without and with RNN-LM as shown in Table 6(Category A without RNN-LM and Category B with RNN-LM).

2.5. Decoding

In decoding phase, we use multiple acoustic models which are described in acoustic modeling section. Firstly, we get the lattices from each acoustic model. Then, we perform lattice fusion followed by MBR decoding [3] to combine recognition results from different models and different versions of GSS.

3. Expermental evaluation

3.1. Acoustic models

For acoustic models, we use official TDNNF model with 15 layers, deeper TDNNF model with 25 layers , CNN-TDNNF model which 6 convolution layers [23] and followed by 19 TDNNF layers. Table 1 compare the three acoustic models using the official training data and frontend module.

Table 1 show that replacing the TDNN-F with the CNN-TDNNF AM yield more than 2 % absolute WER reduction.

Table 1: *WER(%) of different acoustic models on the dev and eval sets*

AM	dev	eval
tdnnf15	51.76	51.29
tdnnf25	50.77	50.30
cnn-tdnnf25	48.53	48.15

3.2. Frontend

In order to match the data in testing stage, we also apply GSS module for all multi-array data in the training stage, instead of randomly selecting 400k utterances from multi-array data in baseline. The result of WER is presented in Table 2. Compared to the official baseline, WER is reduced by 2% absolutely on dev set. We conjecture that the multi-array GSS training data are more compatible with dev and eval dataset. Furthemore, as a result of multi-array GSS in training stage, the amount of training data is reduced from 1500 hours to 240 hours. Consequently, it can speed up acoustic model training.

Table 2: *WER(%) of different frontends on the dev and eval sets*

Frontend	dev	eval
baseline	48.53	48.15
multi-array GSS in training stage	46.54	48.02

3.3. Data augmentation

By applying GSS module in training stage, it significantly reduce the amount of training data. In order to augment the training data, we repalce the L channel worn data with the L+R channel worn data and realigned (L+R) channels worn data respectively, WER can be reduced by 0.99% and 0.72% absolutely on dev and eval sets. After RIR data augmentation, the amount of training data increase by 4 times. It has a total of 1800 hours. After cleaning up, there are 1400 hours left which is equivalent to the baseline. RIR data augmentation greatly improves the performace of our system. The result of WER is presented in Table 3. We achive 0.49% and 1.53% WER absolutely reduction on the dev and eval sets.

Table 3: *WER(%) of different datasets on the dev and eval sets*

data	dev	eval
multi-array GSS + worn(L)	46.54	48.02
multi-array GSS + worn(L+R)	45.55	47.30
multi-array GSS + aligned worn(L+R)	45.80	47.34
multi-array GSS + aligned worn(L+R) + RIR	45.31	45.81

3.4. Feature

WER with different feature-dim are shown in Table 4. Comparing to model which used 40-dimenstion MFCC, we find that using 80-dimenstion MFCC can get 44.99% and 45.28% WER on the dev and eval sets.

Table 4: *WER(%) of different feature-dim on the dev and eval sets*

feature	dev	eval
CNN-TDNN-F4(40-dim)	45.31	45.81
CNN-TDNN-F7(80-dim)	44.99	45.28

3.5. System combination

Finally, our single acoustic model WER are shown in Table 5. We combine lattices produced by multiple acoustic models descibed in section 2.3, and then apply MBR decoding to get the final result. In Track1, for Category A, we get the lattice using offical N-gram LM, combine lattices and apply MBR decoding. At last we achieve the WER of 41.99% and 42.41% on the dev and eval sets. For Category B, The only difference is to rescore the lattices by using RNNLM, we get 41.18% and 42.02% of WER on dev and eval sets.

Table 5: *WER(%) of different acoustic models on the dev and eval sets*

AM	dev	eval
CNN-TDNN-F1(10,2500)	45.20	45.76
CNN-TDNN-F2(10,3500)	45.00	45.58
CNN-TDNN-F3(15,2500)	45.61	45.74
CNN-TDNN-F4(15,3500)	45.31	45.81
CNN-TDNN-F5(10)	45.46	45.80
CNN-TDNN-F6(15)	45.50	46.17
CNN-TDNN-F7	44.99	45.28
TDNN-F8	46.66	47.14

3.6. Results summary

To summarize, the final results of our system in detail on the development and evaluation sets are reported in Table 6. One another point is that RNN-LM is effective for all environments. We are able to confirm the robustness of RNN-LM for the natural conversation.

Table 6: *WERs of the system in Track1(ASR only) for Category A and Category B*

Category	Session		Kitchen	Dining	Living	Ave
A	Dev	S02	47.42	45.57	38.31	41.99
		S09	39.28	43.22	38.80	
	Eval	S01	58.00	35.83	47.80	42.41
		S21	51.49	35.09	34.43	
B	Dev	S02	46.66	45.00	37.47	41.18
		S09	38.48	41.99	38.04	
	Eval	S01	57.56	35.47	47.76	42.02
		S21	50.95	34.95	33.75	

4. Conclusions

In this paper we present our system for the 6th CHiME Challenge. The OPPO systems explore various enhancements, frontends, AM architectures for the final submission system. The system achieve a performance of 20.44% and 18.07% relative Word Error Rate(WER) reduction on the dev and eval sets respectively. The system focuses on acoustic robustness .

5. References

[1] Boeddeker, Christoph, et al. ”Front-end processing for the CHiME-5 dinner party scenario.” CHiME5 Workshop, Hyderabad, India. 2018.

[2] Povey, Daniel, et al. ”Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks.” Interspeech. 2018.

[3] Xu, Haihua, et al. ”Minimum bayes risk decoding and system combination based on a recursion for edit distance.” Computer Speech & Language 25.4 (2011): 802-828.

[4] Saon, George, et al. ”English conversational telephone speech recognition by humans and machines.” arXiv preprint arXiv:1703.02136 (2017).

[5] Peddinti, Vijayaditya, et al. ”Far-Field ASR Without Parallel Data.” INTERSPEECH. Vol. 9. 2016.

[6] Yoshioka, Takuya, et al. ”Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks.” arXiv preprint arXiv:1810.03655 (2018).

[7] Barker, Jon, et al. ”The fifth’CHiME’speech separation and recognition challenge: dataset, task and baselines.” arXiv preprint arXiv:1803.10609 (2018).

[8] Kanda, Naoyuki, et al. ”Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches.” ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[9] Barker, Jon, et al. ”The PASCAL CHiME speech separation and recognition challenge.” Computer Speech & Language 27.3 (2013): 621-633.

[10] Vincent, Emmanuel, et al. ”The second ‘CHiME’speech separation and recognition challenge: An overview of challenge systems and outcomes.” 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013.

[11] Barker, Jon, et al. ”The third ‘CHiME’speech separation and recognition challenge: Dataset, task and baselines.” 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015.

[12] Vincent, Emmanuel, et al. ”The 4th CHiME speech separation and recognition challenge.” URL: http://spandh. dcs. shef. ac. uk/chime challenge Last Accessed on 1 August, 2018 (2016).

[13] Watanabe, Shinji, et al. ”CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings.” arXiv preprint arXiv:2004.09249 (2020).

[14] Ko, Tom, et al. ”A study on data augmentation of reverberant speech for robust speech recognition.” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[15] Ko, Tom, et al. ”Audio augmentation for speech recognition.” Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[16] Ito, Nobutaka, Shoko Araki, and Tomohiro Nakatani. ”Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing.” 2016 24th European Signal Processing Conference (EUSIPCO). IEEE, 2016.

[17] Souden, Mehrez, Jacob Benesty, and Sofiène Affes. ”On optimal frequency-domain multichannel linear filtering for noise reduction.” IEEE Transactions on audio, speech, and language processing 18.2 (2009): 260-276.

[18] Erdogan, Hakan, et al. ”Improved mvdr beamforming using single-channel mask prediction networks.” Interspeech. 2016.

[19] Warsitz, Ernst, and Reinhold Haeb-Umbach. ”Blind acoustic beamforming based on generalized eigenvalue decomposition.” IEEE Transactions on audio, speech, and language processing 15.5 (2007): 1529-1539.

[20] Zorila, Catalin, et al. ”An Investigation into the Effectiveness of Enhancement in ASR Training and Test for CHiME-5 Dinner Party Transcription.” arXiv preprint arXiv:1909.12208 (2019).

[21] Povey, Daniel, et al. ”Purely sequence-trained neural networks for ASR based on lattice-free MMI.” Interspeech. 2016.

[22] Povey, Daniel, et al. ”The Kaldi speech recognition toolkit.” IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF. IEEE Signal Processing Society, 2011.

[23] Ghahremani, Pegah, et al. ”Acoustic Modelling from the Signal Domain Using CNNs.” Interspeech. 2016.