The CW-XMU System For CHiME-6 Challenge

Xuerui Yang[1], Yongyu Gao[1], Shi Qiu[1], Song Li[2], Qingyang Hong[2], Xuesong Liu[1], Lin Li[2], Dexin Liao[2], Hao Lu[2], Feng Tong[2], Qiuhan Guo[2], Huixiang Huang[2], Jiwei Li[1]

[1]CloudWalk Technology Co., Ltd., Shanghai, China
[2]Xiamen University, Xiamen, China

[1]{yangxuerui, gaoyongyu, qiushi, liuxuesong, lijiwei}@cloudwalk.cn
[2]{lilin, qyhong}@xmu.edu.cn

Abstract

In this paper, we present Cloudwalk Technology and Xiamen Universitys joint effort for CHiME-6 Challenge to recognize highly-overlapped and very natural conversational speech in dinner party environment. We explored DNN-HMM hybrid system for track 1 rank A and end-to-end model for track 1 rank B. In addition, we also explore different data augmentation approaches and front-end speech enhancement methods to further improve the accuracy of speech recognition systems. We investigated vairous algorithms in speech diarization systems for track 2. Our system came up with 41.65% WER for development set and 40.24%WER for evaluation set in rank A, as well as 40.25% WER for development set and 39.62%WER for evaluation in rank B for track 1. For track 2 category A, results are 57.72%DER, 61.85%JER and 77.5%WER for development set, as well as 65.36%DER, 67.32%JER, 72.52%WER for evaluation sets.

Index Terms: speech recognition, CHiME-6 challenge, dinner party

1. Introduction

CHiME-6 features two tracks: multiple-array speech recognition (track 1) and multiple-array diarization with recognition (track 2). We participate in both catergory A and B for track 1 and catergorty A for track 2. We are going to demonstrate our front-end system in Section 2.1, back-end system in Setion 2.2, some efforts on language model for catergory B in Section 2.3 and experiments for track 2 in Section 3.

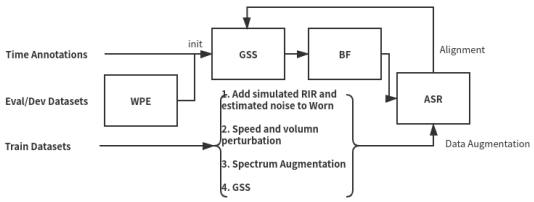

Figure 1: *System Overview*

2. Track1

In this section, we will introduce front-end enhancement, data augementation, the modification on acoustic model training and lattice combination for track1. A simple overview of our system based on official language modeling is illustrated from Figure 1.

2.1. Front-end Speech Enhancement

Our front-end framework is based on dereverberation, guided source separation and mask-based beamforming.

2.1.1. Deverberation

Weighted prediction error WPE [1] has been proved that it can effectvely improve real-life speech recognition performance, we employed multi-channel WPE [2] [1] to do speech dereveberation. We used nara_wpe [1] with same configurations as baseline system for dereverberation.

2.1.2. Guided Source Separation

Complex Angular Central Gaussian Mixture Model(cACGMM) has been explored that it can significantly solve the guided source separation problem [3] [4].

However, due to unmarked silence between words, the annotions provided from datasets are not perfectlly precised. Hence, we followed up frame-level alignment for GSS from well-trained ASR model [5] [6]. The model we used for alingment is trained.

2.1.3. Beamforming

Given the estimated mask from Guided Source Separation, Minimum Variance Distortionless Response beamformer with speaker-aware complex Gaussian mixture models has been applied. We used the baseline cgmm-mvdr for pb_chime5 tool [7].

2.2. Back-end ASR For RankA

2.2.1. Acoustic Model

We experimented various network structures including tdnnf, residual cnn, fsmn and self-attention for hybird system. For the final result, minimum Bayesian risk lattice combination is applied and the lattices are from these acoustic models.

- **Res-cnn-tdnnf-self-attention:** 6-layers convolution neural network with skip connections every two layers, 15-layer tdnnf and a time-restricted self-attention block.
- **Res-cnn-fsmn:** 6-layers convolution neural network with skip connections every two layers, 10-layer pyra-

midal fsmn block.

- **Spec-aug-cnn-tdnnf:** Spec augmentation is applied in front of 6-layers CNN block, followed by 15-layer tdnnf.
- **Multi-cnn-tdnnf-self-attention:** Three 6-layers cnn blocks with different Convolution kernel size simultaneously concate by 15-layer tdnnf and time-restricted self-attention block.

2.2.2. Neural-Network Alignment

Due to the complicated representation of audio data, GMM could be less inaccurate for phonetic alignment. Thus, we trained a chain model without subsampling for alignment which has shown improvement.

2.2.3. Data Augmentation

Tranditional speed and volume perturbation are applied in the training pipeline. Since worn data is relatively clean. To simulate real mic array audio, reverberant and noise augmentation are added. The RIR was generated according to training data floorplan's configurations. SpecAug which gives masks on the spectrogram of input utterances, are also performed during training. In addition, 24 and 12 micorphones GSS has been employed to augment training data. Two types of data are prepared, worn + guided speech separated training data and pure guided speech separated training data.

2.2.4. Other Tricks

- **Strict cleanup:** After decoding with training data, several utterances were found not match to the transcriptions, we remove part of them by high WER;
- **Chain-model tree leaves:** Various senones as modelling units are experimented, and found 5000 better than the baseline.

2.3. RankB

2.3.1. Language model rescore

In rank B, 4-stage pipeline are used. First, the lattice generated from HCLG.fst are rescored through a 4-gram language model. Then, a pruned lstm-based lattice rescore is applied. Finally, the lattice will be sent to an n-best rescore model.

2.3.2. End-to-End ASR

We introduce the CTC loss function to assist Transformer in learning the speech-to-text alignment. A RNNLM is trained for decoding stage. The training data of our end-to-end Hybrid CTC-Transformer System contains two parts: one is straight from the kaldi baseline, and the other was generated by performing WPE, Beamformit, and WPE+Beamformit on the multi-array training set. We also use SpecAugment to improve the robustness of the end-to-end speech recognition systems.

3. Track2

In this section, we will introduce the acoustic features, embedding extractors, and clustering algorithms used in track2.

3.1. Acoustic features

We expeimented different configurations of Mel frequency cepstral coefficient features (MFCC) and Filterbank features(FB).

3.2. Embedding extractors

We have considered two different embedding extractors. The first embedding extractor we used is the official pretrained diarization model. X-vector DNN is trained with the VoxCeleb data and PLDA model is trained with the CHiME-6 data. We used the same data to train the second model, with a different architecture and feature. In order to make speaker embedding obtain better distinction between classes, we chose the factorized time delay deep neural network (F-TDNN) architecture. It has excellent performance in speaker recognition tasks. Based on experience, we choose to use 40-FB with more detailed information to train this model.

3.3. Clustering algorithms

In addition to the AHC used by the official baseline, we also explored the spectral clustering. Because the number of speakers in each sentence in ChiME6 is confirmed to be 4, the threshold is also determined accordingly.

3.4. VB refinement

After segment level clustering, because embedding segments are too quantized, we use VB to refine the mark boundaries. The parameters are re-learned with VoxCeleb data using 23-MFCC.

4. Results

The results of hybird system for Track 1 rank A are 41.65 WER% and 40.24%WER in development sets and evaluation, along with 56.9% and 50.6% in dev and eval for End-to-End system in rank B track 1.

Track	Rank	Dev (WER %)	Eval (WER %)
1	A	41.65	40.24
1	B	40.25	39.62
1	B	56.9	50.6

Table 1: Track 1 results

The results of track 2 category A are 57.72%DER, 61.85%JER and 77.5%WER for development set and 65.36%DER, 67.32%JER, 72.52%WER for evaluation sets.

Baseline	Development Set			Evaluation Set		
	DER%	JER%	WER%	DER%	JER%	WER%
Category A	57.72	61.85	77.52	65.36	67.32	72.52

Table B: Track 2 category A results

5. Conclusions

In this paper,we did various investigations on both track 1 and track 2 that all giving a better results than baseline.

6. References

[1] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *13. ITG Fachtagung Sprachkommunikation (ITG 2018)*, Oct 2018.

[2] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio Speech & Language Processing*, vol. 20, no. 10, pp. 2707–2720.

[3] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016.

[4] C. Boeddecker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," 09 2018.

[5] C. Zorila, C. Boeddeker, R. Doddipatla, and R. 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)*, Dec 2019. [Online]. Available: http://dx.doi.org/10.1109/ASRU46091.2019.9003785

[6] N. Kanda, C. Böddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong ASR backend: Hitachi/paderborn university joint investigation for dinner party ASR," *CoRR*, vol. abs/1905.12230, 2019. [Online]. Available: http://arxiv.org/abs/1905.12230

[7] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-End Processing for the CHiME-5 Dinner Party Scenario," in *CHiME5 Workshop*, 2018.