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.

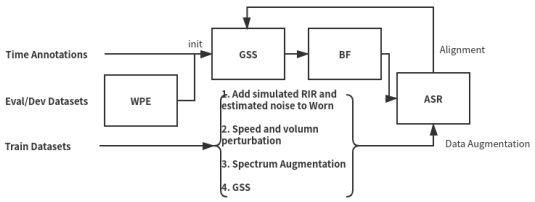

Figure 1: *System Overview*

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 which are 512 sample points for frame length, 128 sample points for frameshift, 10 frames for filter taps and 2 frames delay. Moreover, we have tried coherent-to-diffuse power ratio(CDR) [3] for dereverberation due to source signals were deeply reverberated. We modifed CDR to satisfy 4 channels singals. Table 1 showed that WPE was out performed than CDR.

Dereverberation	**Eval(WER)**
Wpe(Baseline)	52.04%
Cdr	54.19%

Table 1: Recognition rate comparison between WPE and CDR

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 [4] [5].

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 [6] [7]. The model we used for alingment is trained.

In our experiment, the training data was augmented with 12mics and 24mics GSS. For evaluations, the GSS was set with 5 iterations, and realignded by pre-trained ASR model. We used different training data combinations to train models with same structure, as shown in table 2.

Alignment from model	Train data
x	Augmented worn + array +beamformit
y	Augmented worn + gss + array
z	GSS

Table 2: Alignments from different models

Realginments from well-trained ASR models can at least enhance 1.5% absolute WER in our systems. The results in table 3 illustrates alignment from model trained with augmented gss worn and array data provided the best enhancement, and the alignment from model trained with augmented worn, array and beamformit data is better than the alignment from model trained with pure gss is because the former model has better variety.

Model	GSS realignment from Model	WER
MULTI-CNN-TDNNF-ATT-specaug	None	45.11
	x	43.82
	y	43.78
	z	43.86

Table 3: Best single model gss realignment results

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 [8].

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 pyramidal 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.

In the experiments, 40-dim fbank and 120-dim i-vector features are fixed as the network input. Acoustic model structures are one of the most influential factors for the recognition tasks. Thus, we trained several AM structures and some of them out perform the Kaldi baseline TDNNF model in table 4. MULTI-CNN-TDNNF-ATTENTION with spec augment shows the superior result.

MODEL	Eval (WER)
TDNNF-baseline	52.04%
CNN-FSMN-ATT-1c	49.96%
CNN-(14)TDNNF-ATT	49.64%
CNN-(13)TDNNF-ATT	49.84%
MULTI-CNN-TDNNF-ATT-specaug	48.41%
TDNN-RBiLSTM	53.92%

Table 4: Variant AM architectures comparison

During training, LF-MMI and LF-bMMI are used. It is found that half-LF-MMI-half-LF-bMMI training criterion reaches the best result as shown in Figure 5

Training criterion	Eval (WER)
LF-MMI	52.04%
LF-bMMI	52.35%
Hybrid*	51.87%

Table 5: Various training criterion results with Kaldi baseline setup

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

Since worn data is relatively clean, we simulated reverberated speech from the worn data by utilizing room impulse responses (RIR) and point-source noises. In ad- dition, we performed weighted prediction error (WPE), beam- forming, and hybrid speech enhancement on all training data for data augmentation. Finally, speed and volume perturbation- s were utilized on all training data. During Training, we also used SpecAugment to improve the robustness of our speech recognitions system.In addition, 24 and 12 micorphones GSS has been employed to augment training data. Two types of data are prepared, worn with array and guided speech separated training data, and pure guided speech separated training data.

The models are trained in two ways: one is pure GSS training data, and the other one combines worn, augmented worn and GSS data to further improve the training diversity. Table 6 illustrates the results of these two training methods. However, beamforming data makes degradation for results. It is inferred that beamforming method makes use of multi mic arrays which is not compatible with the test data generated by separation method.

Model	Data Combinations	Eval (WER)
Model A	Augmented worn + array + GSS	45.53%
	Pure GSS	47.06%
	Augmented worn + array + beamforming + GSS	51.04%
Model B	Augmented worn + array + GSS	45.87%
	Pure GSS	46.84%
	Augmented worn + array + beamforming + GSS	50.82%
Model C	Augmented worn + array + GSS	45.59%
	Pure GSS	46.69%
Model D	Augmented worn + array + GSS	45.11%
	Pure GSS	46.47%

Table 6: Data combination brings more diversity in training process

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.

In hybrid ASR system, Phonetic alignment is critical to training. Since chain model performs best result and the robustness of neural network model. However, chain model output frame labels with sub-sampling of 3. To align the training data with it, a chain model without sub-sampling is used designed. 1.21% absolute improvement is received.

Ctiterion	Type
LF-MMI	1
LF-bMMI	2
Hybrid	3

Data	Type
Augmented worn + array + GSS	A
Pure GSS	B

Table 7 and Table 8:Training criterion and training data setup

The models are trained in two ways: one is pure GSS training data, and the other one combines worn, augmented worn and GSS data to further improve the training diversity, figure 9 illustrates the results of these two training methods. However, beamforming data makes degradation for results. It is inferred that beamforming method makes use of multi mic arrays which is not compatible with the test data generated by separation method.
In the final, we ensemble some of the best models using MBR(Minimum Bayesian Risk) decoding. The experiments indicate that not combining the best trained single models will not give the best ensemble result, but the models with rich training and data diversity do.

MODEL	Criterion	DATA	Eval (WER)
4 AMs	1	A	42.15%
4 AMs	1	B	43.54%
4 AMs	2	A	42.67%
4 AMs	3	A	41.34%
6 AMs	1	A+B	40.89%
6 AMs	3	A	41.15%
6 AMs	3	A+B	40.67%
8 AMs	3	A+B	40.24%

Table 9: Different ensemble combination and their results

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 system layout of E2E can be reviewed in figure 2.

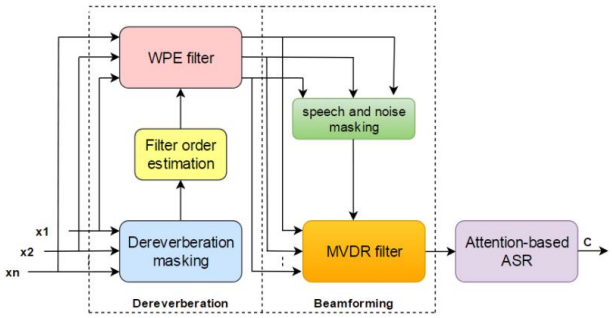

Figure 2: *E2E System Overview*

- **E2E data augmentation:**The data augmentaion is similiar as rank A. For better accuracy of the end-to-end speech recognition systems, we exploited three front-end speech enhancement algo- rithms on the development set and evaluation set of the CHiME- 6 corpus: delay-and-sum beamforming, NN-based MVDR beamforming and hybrid speech enhancement.
- **Recognition details:** For the multi-channel end-to-end ASR system(ME2E), BLSTMP Neural Networks were used for the dereverberation subnetwork and beamforming subnetwork to estimate the as- sociated masks. Transformer architecture was used as the attention-based ASR module of ME2E. The training steps of ME2E were divided into three stages. First, we used the worn data to train a Transformer ASR model. Second, worn data and array data were used to train the dereverberation subnetwork and beamforming subnetwork based on MSE and SI-SNR objectives using the worn data as the training label. Finally, we trained the Transformer ASR model and the two speech en- hancement networks jointly with worn data and array data.

The result of E2E system is shown in table 11.

3. Track2

The challenge speech contains multiple speakers, so we need to know the start and end time of each speaker, otherwise the performance of ASR will deteriorate sharply. This is exactly the goal of diarization. In this section, we will introduce the acoustic features, embedding extractors, and clustering algorithms used in diarization for track2.

3.1. Acoustic features

In diarization we mainly use three different configurations of acoustic features.

3.1.1. *Mel frequency cepstral coefficient features*

For the first mel frequency cepstral coefficient (MFCC) feature extraction configuration, all audios are converted to the cepstral features of 23-dimensional MFCC with a frame-length of 25ms and a frame shift of 10ms. The second configuration differs only in the choice of dimensions, using 40 dimensions.

3.1.2. Filterbank features

The filter bank (FB) feature vectors include 40 dimensional FBs from the raw signal with a 25ms frame-length.

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 [9] DNN is trained with the VoxCeleb [10, 11] data and PLDA model is trained with the CHiME-6 data. We use the same data to train the second model, with a different architecture and feature. In order to make speaker embedding have better distinction between classes, we chose the factorized time delay deep neural network (F-TDNN) [12] 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. The architecture configuration is shown in Table 10. The embedding are extracted as the feature of different speakers used for clustering algorithms.

Layer	Layer Type	Context factor1	Context factor2	Skip conn. from layer	Size	Inter size
1	TDNN-ReLu	t-2:t+2			512	
2	F-TDNN-ReLu	t-1,t	t,t+1		640	180
3	F-TDNN-ReLu	t	t		640	180
4	F-TDNN-ReLu	t-2,t	t,t+2		640	180
5	F-TDNN-ReLu	t	t	3	640	180
6	F-TDNN-ReLu	t-3,t	t,t+3		640	180
7	F-TDNN-ReLu	t	t	2,4,6	640	180
8	F-TDNN-ReLu	t-3,t	t,t+3		640	180
9	F-TDNN-ReLu	t	t		640	180
10	F-TDNN-ReLu	t-3,t	t,t+3	5,7,9	640	180
11	F-TDNN-ReLu	t	t	6,8,10	640	180
12	Dense-ReLU	t	t		1024	
13	Pooling(mean+stddev)	full-seq			2048	
14	Dense-ReLU				512	
15	Dense-ReLU				512	
16	Dense-Softmax				N.spks	

Table 10: F-tdnn architecture

3.3. Clustering algorithms

In addition to the agglomerative hierarchical clustering (AHC) [13] used by the official baseline, we also explored the spectral clustering used in [14]. The number of speakers in each sentence in ChiME6 is confirmed to be 4, so the threshold is also determined accordingly.

3.4. VB Refinement

After segment level clustering, because embedding segments are too quantized, we use Variational-Bayesian(VB) refinement [15] to refine the mark boundaries. The parameters are relearned with VoxCeleb data using 23-MFCC. All the diarization results are shown in Table 12.

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 11: Track 1 submitted 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 12: Track 2 submitted 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. Our team started Chime-6 at Feburary, we think more experiments can be done as well as better improvements can be achieve if we start it earlier.

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] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *CoRR*, vol. abs/1502.03784, 2015. [Online]. Available: http://arxiv.org/abs/1502.03784

[4] 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.

[5] 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.

[6] 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

[7] 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

[8] 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.

[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[10] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[12] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[13] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

[14] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.

[15] G. Sell and D. Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4794–4798.