The SHNU System for the CHiME-5 Challenge

Yanhua Long, Renke He

Laboratory of Natural Human-Computer Interaction, Shanghai Normal University, Shanghai, China

yanhua@shnu.edu.cn, heryan23@163.com

Abstract

This paper is a description of our system submitted to the 5th CHiME challenge. In this challenge, we only focus on the *A-single-array* task. Several improvements over the conventional ASR baseline have been used, including the speech processing front-end, the automatic training data augmentation of the official training data, and the acoustic model combination with different structures. The final overall WERs of both the development and evaluation sets only using the reference Kinect array are around 64%.

1. Background

The series of CHiME challenge aims to advance robust automatic speech recognition (ASR) technologies. The 5th CHiME Challenge (CHiME-5), which considers the task of distant multi-microphone conversational ASR in real home environments. CHiME-5 features two tracks depending on the number of microphone arrays available for testing: a single-array track (*A-single-array*) and a multiple-array track. For each track, distinct rankings will be produced for systems focusing on robustness with respect to distant-microphone capture vs. systems attempting to address all aspects of the task including conversational language modeling [1].

It is the first time that we are participating in the CHiME challenge. For CHiME-5, we only focus on the *A-single-array* track, in which only one reference array is used to recognize a given evaluation utterance, and all our sub-systems are based on the conventional acoustic modeling and official language modeling: the outputs of the acoustic model remain frame-level tied phonetic (senone) targets, and the lexicon and language model are the same ones used in the conventional ASR baseline [1]. Compared with the baseline, our contribution includes: 1) New log mel filterbank (FBANK) features instead of MFCCs are used to build the acoustic models; 2) Training data augmentation; 3) Different acoustic model architectures; 4) System fusion using minimum Bayes risk decoding.

2. Contributions

The overall system framework is shown in Fig.1. We perform the 4-channel singe-array beamforming used in the Kaldi baseline [1] to all the training, development and evaluation datasets. Three types of acoustic models with different structures are built on the same augmented training data. In the testing stage, multiple systems with different acoustic models are fused at the lattice level, using the minimum Bayes risk decoding. We don't change the language model and lexicon, they are the same ones as in the conventional ASR baseline.

2.1. Speech enhancement

Two speech enhancement approaches are tried in our systems. The first one is the same approach used in the ASR baseline,

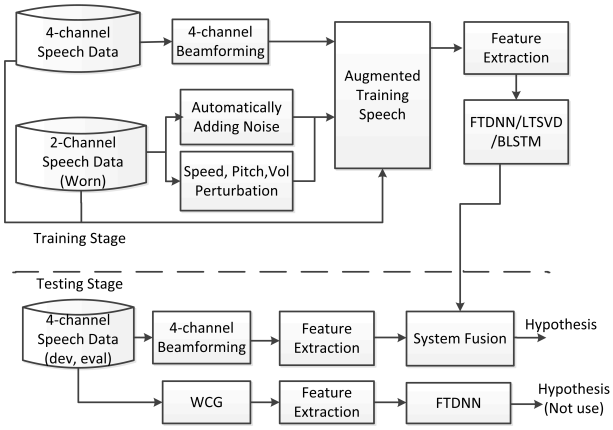

Figure 1: *System framework.*

which is a weighted delay-and-sum beamforming approach [2]. This beamforming is used to enhance both the development and evaluation speech. Moreover, it is also applied to the 4-channel training speech to generate its enhanced version to match the testing 4-channel case.

The second one is also similar to the baseline beamforming, as shown in Fig.2, it enhances speech in three steps: first we use the weighted prediction error (WPE) method to improve the robustness of source separation against reverberation [3], then a mask estimation method based on a complex angular central gaussian mixture model (CACGMM) is used to separate the sources [4], at last the generalized sidelobe cancelation (GSC) beamforming weights are estimated based on the soft mask. We call this approach as WCG beamforming.

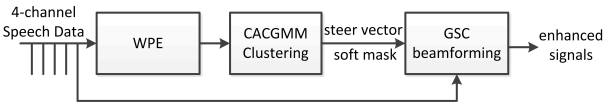

Figure 2: *WCG beamforming framework.*

2.2. Training data augmentation

Data augmentation is a common way to enlarge the training data coverage. It is adopted to improve the robustness of the acoustic models. In our system, we perform several straightforward ways which directly process the raw audio signal collected only from the binaural microphones (both left and right channels), to augment the training data. First, we change the speed of the audio signal, producing 3 versions of the original signal with speed factors of 0.9, 1.0 and 1.1. Then, the pitch and volume of the audio is randomly modified to simulate the effect of pitch variation and different recording volumes. Moreover, in order

10.21437/CHiME.2018-14

to simulate the noise types of parties, we automatically select those segments in the training data which have been forced-aligned as noises. These noise segments are then taken as the noise sources to be added to the raw audio signal at a random SNR. Finally, we combine the beamformer enhanced version of the training data together with these augmented versions, the channel 1 of all Kinect microphone data and the development set of its binaural microphone data as the final training dataset to train all of the acoustic models with variety structures.

2.3. Feature extraction

Instead of using the MFCCs as in the official conventional ASR baseline, we use a 80-dimensional FBANK features to train all of our deep neural networks.

2.4. Acoustic models

Based on the hidden Markov model-Gaussian mixture model (HMM-GMM) system in the conventional ASR baseline, we use Kaldi toolkit [5] to train three types of neural networks. They are the FTDNN: a factored form of time delay neural network (TDNN) which is structurally the same as a sub-sampled TDNN [6] whose layers have been compressed via Singular Value Decomposition (SVD), but is trained from a random start with one of the two factors of each matrix constrained to be semi-orthogonal [7], it has 11 layers, the hidden-layer dimension is 1536, and the linear bottleneck size is 384; BLSTM: the bidirectional long short term memory model (BLSTM) with 3 layers, each hidden layer consists of 1024 memory cells together with a 256-node projection and non-projection layer [8]. LTSVD: it is similar to the LTDNN structure in [9], it is a 768 dimensional system with a mixture architecture of LSTMPs and sub-sampled TDNNs, using 3 fast-LSTMP layers interleaved with 7 spliced TDNN layers, but the 3 fast-LSTMP layers have been compressed via 256-dimensional SVD; All of these acoustic models use the lattice-free maximum mutual information (LF-MMI) training criterion [10]. We use the same context-dependency tree of the baseline TDNN system for all the other AM structures.

For the FTDNN and LTSVD models, the skip connections idea is applied. This is somewhat related to the highway connections [11] The basic idea is that some layers receive as input, not just the output of the previous layer but also selected other prior layers which are appended to the previous one. Details of the skip connection structures can be found in the Kaldi github repository of egs/swbd/s5c/local/chain/tuning/run_tdnn_7p.sh and egs/swbd/s5c/local/chain/tuning/run_tdnn_lstm_1n.sh.

3. Experimental evaluation

3.1. Speech enhancement

In our experiments, we compared the official beamforming approach with our proposed WCG one. We use the official TDNN baseline system to examine these speech enhancement approaches. Results are shown in Table 1. We find that the official one is a bit better than ours. So we choose to directly use the official beamforming to build our final system.

3.2. Feature extraction

Table 2 reports the results of the difference between MFCC and FBANK features on the development set for A-single-array track, using the official TDNN system. It can be seen that, the 80-dimensional FBANK features outperform the MFCCs by a relative 3% WER reduction. We also tested the effectiveness of ivectors, our results in Table 2 tell us that, no improvements can be obtained by using ivectors. Therefore, we choose to use the FBANKs without ivectors to build the FTDNN, LTSVD and BLSTM acoustic models.

Table 1: *Overall WER (%) obtained with the proposed WCG and official beamformers on the development set for A-single-array track using the official TDNN system.*

Speech enhancement	WER
WCG	82.60
official beamforming	80.80

Table 2: *Overall WER (%) obtained with the MFCCs and FBANKs on the development set for A-single-array track using the official TDNN system.*

Features	WER
MFCC+ivector	81.01
MFCC	80.80
FBANK	78.41

3.3. Training data augmentation

Table 3: *WER comparison with different data augmentation techniques on the development set for A-single-array track using FTDNN system.*

Data	WER
train_worn_u100k_cleaned_sp	75.59
train_worn_uall_cleaned_sp	72.46
+dev_worn	70.60
+vol_pitch_add_noise	68.16
+uall_bf	70.06

We use the FTDNN system to evaluate the effectiveness of the training data augmentation approaches. The results on the development set are shown in Table 3. We take the training dataset (in *data/train_worn_u100k_cleaned_sp*) used for TDNN training in the official conventional ASR system as the baseline. Then we combine those utterances extracted from the channel 1 of all Kinect microphone array, and both left and right channels of the binaural microphone data. The *train_worn_uall_cleaned_sp* refers to its cleaned version with speed perturbation, it is around 646 hours training data in total. We obtain absolute 3.13% WER reduction when we use more data from the Kinect microphone array. As it is allowed to use the development audio and its annotation for training acoustic models, so we add the development set of its binaural microphone data to the 646 hours training data, further absolute 1.86% WER reduction is obtained. From the last second line of Table 3, we find only absolute 2.44% performance improvement by using the volume, pitch and noise augmentation. However, we don't achieve any benefit by adding the beamformer enhanced version (*uall_bf*) of all training data from Kinect microphone array to train the FTDNN model. Therefore, we take the

training dataset which obtained the WER=68.16% of FTDNN system to train all the other two sub-systems.

3.4. Acoustic models

Three sub-systems are built and fused at the lattice level using the minimum Bayes risk decoding citerion. We can see that, the FTDNN is the best single system, and system fusion gives absolute 3.71% WER reduction. The submitted results of both the development and evaluation test sets are the outputs from the fusion of three sub-systems.

Table 4: *Overall WER (%) for the sub-systems tested on the development test set for the A-single-array track.*

Track	System	WER
Single	Official TDNN	80.80
	FTDNN	68.16
	LTSVD	70.14
	BLSTM	72.58
	FTDNN+ LTSVD+ BLSTM	64.45

Table 5 shows the results for the best system (system fusion). It gives the result per session and location together with the overall WER. It is not surprising that adding the development data to the training dataset to build the acoustic models can improve the performance when it is evaluated on the development set itself. However, we are worried about that, the acoustic model parameters tuned on the development set may be difficult to be generalized to the evaluation set, since these models may overfit to the development dataset. It is surprising that we got almost the same performance for both the development and evaluation sets. And we also tested the evaluation set on the system that trained without adding the development data. We found that, the overfit problem was not serious as we worried about.

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

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	74.83	64.35	61.72	64.44
		S09	62.74	59.62	58.53	
	Eval	S01	70.33	58.46	75.91	64.45
		S21	67.84	58.12	60.83	

4. Conclusion

This paper is a simple description of the SHNU system that submitted to the CHiME-5 challenge. Different beamforming, data augmentation approaches, variants of acoustic model architectures, and system combination were investigated to improve the final system. Unfortunately, our WCG beamforming did not improve the performance, but the rest of the techniques significantly improved the performance from the baseline. Our future work will focus on exploring new and effective front-end technique to do the noisy speech separation and enhancement.

5. Acknowledgments

We would like to thank Daniel Povey and his team for their great work on Kaldi. And we also would like to thank Yuxing Cao for his contributions of speech signal processing. This work is funded by the Project 61701306 supported by National Natural Science Foundation of China.

6. References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2011–2023, 2007.

[3] Y. Takuya and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 10, no. 20, pp. 2707–2720, 2012.

[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 *Proceedings of the 24th European Signal Processing Conference (EUSIPCO 2016)*, Hilton Budapest, Hungary, Aug. 2016.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, and et.al, "The kaldi speech recognition toolkit," in *Proceedings of the IEEE 2011 workshop on automatic speech recognition and understanding*, Hawaii, USA, Dec. 2011.

[6] P. Vijayaditya, D. Povey, , and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Dresden, Germany, Sep. 2015.

[7] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018.

[8] G. Pundak and T. N. Sainath, "Lower frame rate neural network acoustic models," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, San Francisco, USA, Sep. 2016.

[9] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and lstms," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2017.

[10] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, and et.al., "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, San Francisco, USA, Sep. 2016.

[11] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory rnns for distant speech recognition," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, Sep. 2016.