The AnTech System for CHiME-5 Challenge

WangTao, LiXiufeng, WangLin

wangtaody@qq.com, lixiufeng1213@163.com, wanglin198717@163.com

Abstract

This paper submit the result for CHiME5 challenge, aiming at single-array audio based on the traditional audio speech recognition system. A GMM+DNN model was trained and rnn-lm rescoring was used. The final WER of array and worn of Dev dataset are 78.29%,41.11% respectively.

Keywords: ASR, rmmlm, beam forming

1. Background

Audio speech recognition(ASR) has been more and more useful in daily work and life. Word error rate(WER) may be less than 5% in quite environment while in daily environment full of noise or long distance between speaker and mircophone the WER can reach 100%. To improve the performance of the ASR system in daily life, CHiME5 provides data recorded in daily life and baseline script based on kaldi.

An ASR system can be separated into 3 parts, front-end for audio pre-processing, model training including acoustic model and language model, and decoding for getting transcription. Front-end plays an important role in decreasing WER, a lot of work have been done, like, noise reducing and speech enhancing. In ChiME-5, we can use single-array audio for single-array track or multiple-array audio for multi-array track.

This paper contributes to the single-array track. ChiME-5 data sets and baseline script were referenced[1,2].

2. contributions

2.1 Training Data flow

Data provided was showed in table 1.

Table 1 Data sets of CHiME-5

Dataset	Parties	Speakers	Hours	Utterances
Train	16	32	40:33	79,980
Dev	2	8	4:27	7,440
Eval	2	8	5:12	11,028

In Training and Dev dataset, there are array data recorded by 6 channel device and worn data recorded by 2 channel device. In Eval dataset, only multi-array of 6 channel audio was gave. In training stage, data flow was showed in figure 1.

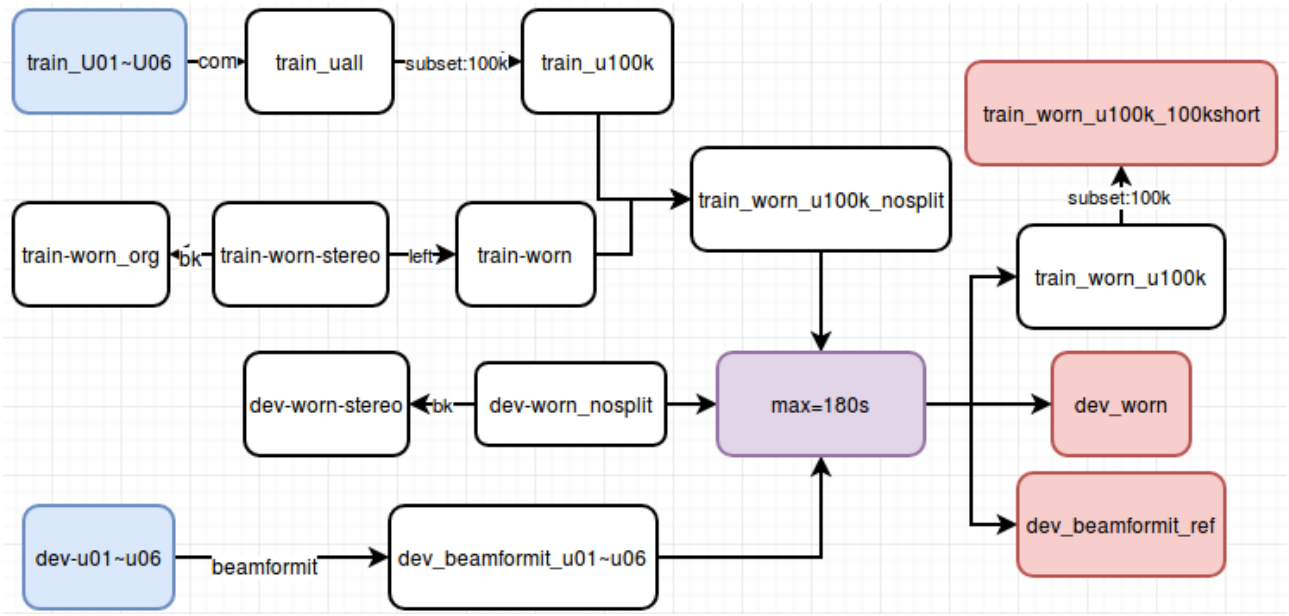

figure1　Data flow of the ASR system

In Dev dataset, array data was enhanced by beam forming using open source beamform toolkit[3,4]. For worn data, only left channel data was used.

2.2 Training and decoding

In this page, an ASR system based on GMM+DNN was provided as figure 2.

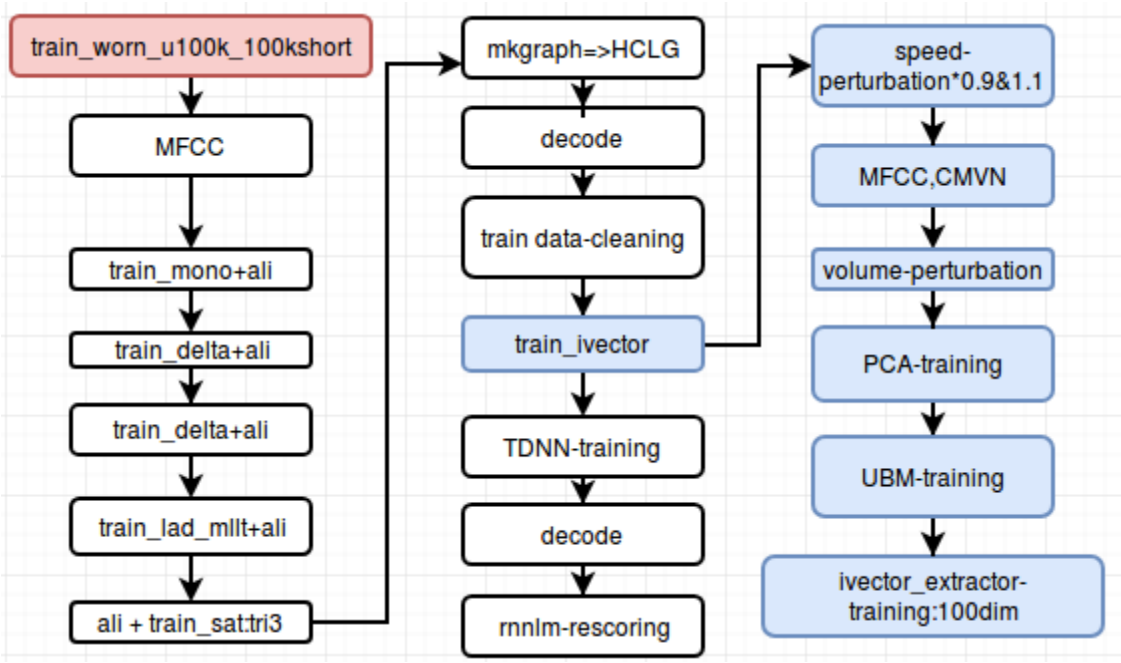

figure 2　Training steps

In training stage, main steps were : train-mono, train-delta, train-lad-mllt, train-sat and TDNN. The stages before TDNN were in GMM stage in witch MFCC of 39-dim and energy of 1-dim were used as the system's input. In TDNN stage, iVector and MFCC were used as the input.

RNN-LM rescoring was used in decoding stage. There are 3 steps to do this :

Step-1: Using tensorflow to train an RNN-LM with transcription of Training dataset and Dev dataset;

Step-2: Using script s5/steps/tfrnnlm/lmrescore_rnnlm_lat.sh to rescore the lattice get from normal decode stage and get the transcription of the dataset.

Step-3: Get WER of Dev dataset.

3. Experimental evaluation

As the reference transcription of Eval data was not given, only the WER on Dev dataset was present in table 2 .

Table 2 Results got in different stages

		dataSet	GMM	TDNN	Rnnlm-scrore
Single Track	baseline	worn	76.4	47.91	
		beam	93.56	81.28	
	AnTech	worn	71.66	49.27	41.11
		beam	90.68	82.44	78.29

For Dev dataset, scores per room and session are showed in table 3 with LM weight 9 and insertion penalty weight 0.0.

Table 3 Results for session

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	85.87%	75.74%	75.31%	78.99%
		S09	78.70%	76.88%	75.06%	76.88%
	Eval	S02	```	```	```	```
		S09	```	```	```	```

Lattices of the Dev and Eval dataset can be found in at :

https://pan.baidu.com/s/1BeXjPd-8iYy6heYGy3xZDw

4. References

[1] http://interspeech2018.org/author-resources.html

[2] http://spandh.dcs.shef.ac.uk/chime_challenge/instructions.html

[3]Acoustic beamforming for speaker diarization of meetings", Xavier Anguera, Chuck Wooters and Javier Hernando, IEEE Transactions on Audio, Speech and Language Processing, September 2007, volume 15, number 7, pp.2011-2023

[4]Robust Speaker Diarization for Meetings", Xavier Anguera, PhD Thesis, UPC Barcelona, 2006