LEAP Submission to CHiME-5 Challenge

Sriram Ganapathy, Purvi Agrawal

Learning and Extraction of Acoustic Patterns Lab (LEAP), Dept. of Electrical Engg.,
Indian Institute of Science, Bengaluru-560012, India.

(purvia, sriramg)@iisc.ac.in

Abstract

This report describes the LEAP system submitted to the CHiME-5 Automatic Speech Recognition (ASR) challenge (Track A-1 i.e, single-array track). The system submitted for the evaluation is a combination of two sub-systems, one based on conventional mel frequency features and second one based on the frequency domain linear prediction features. Both the systems use the convolutional neural network based acoustic model which advances the baseline system. The combination result improves the baseline system absolutely by 8% in terms of word error rate on the development data (beamformed baseline).

1. System Description

1.1. System-A

For this sub-system, the feature extraction is done using 40 dimensional mel-frequency filter bank energies which are extracted using 25ms windows with a shift of 10ms (denoted as *fbank*). The features are mean and variance normalized and are used in acoustic modeling. We use the same setup as described in the CHiME-5 baseline system [1] which uses both worn microphone and beamformed audio for model training.

The acoustic model used in this system is given in Fig. 1. The system consists of convolutional neural network front-end followed by time-delay neural network (TDNN) layers. The output of the TDNN layers are fed to long-short-term memory network (LSTM) which outputs the target senones. The model is implemented in Kaldi [2] and this is trained using the chain training framework [3].

1.2. System-B

For this sub-system, the acoustic model described in Fig. 1 is used as it is. However, the spectrogram is derived using the multi-variate auto-regressive (MAR) model. These features are based on frequency domain linear prediction (denoted as *fdlp*) approach [4]. The feature extraction module is shown in Fig. 2. These features are also 40 dimensional.

2. Results

The speech recognition results using baseline system (provided by [1]), System-A, System-B and combined system (system combination using lattice combination performed using Kaldi) are given in Table 1.

Table 1: *ASR results - word error rate (%) for various systems for single-array track.*

System	Dev-Worn Mic	Dev-Beamform
Baseline	48.0	81.3
System-A	44.1	75.8
System-B	45.5	77.4
Sys. Comb (A + B)	41.3	**73.4**

3. 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] S. Ganapathy and V. Peddinti, "3-d cnn models for far-field multichannel speech recognition," *ICASSP*, 2017.

[3] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[4] S. Ganapathy, "Multivariate autoregressive spectrogram modeling for noisy speech recognition," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1373–1377, 2017.

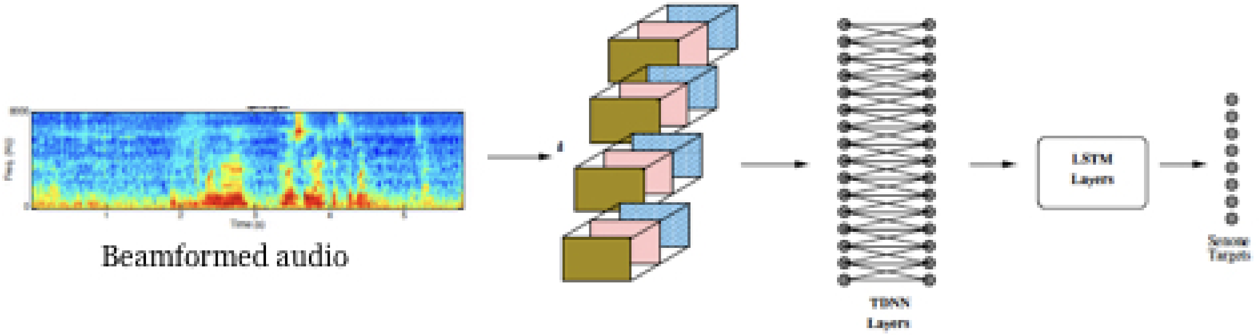

Figure 1: *The acoustic model used in the LEAP system consisting of CNN-TDNN-LSTM neural network. The model is trained with chain training framework in Kaldi.*

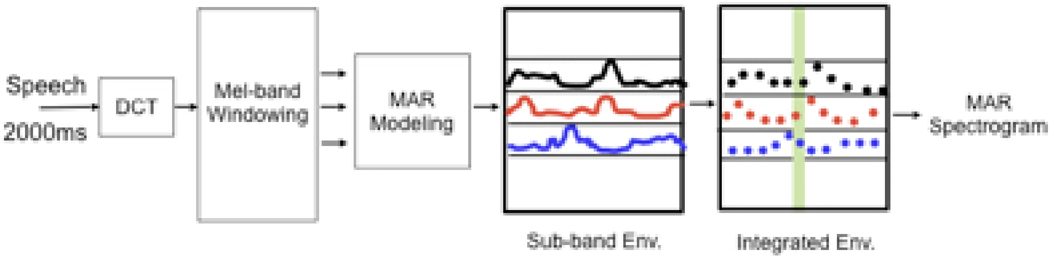

Figure 2: *The feature extraction module based on multi-variate autoregressive modeling [4].*