LEAP Submission to CHiME-5 Challenge

Sriram Ganapathy, Purvi Agrawal
Learning and Extraction of Acoustic Patterns (LEAP) Lab,
Department of Electrical Engineering, Indian Institute of Science (IISc)
(purvia, sriramg)@iisc.ac.in

Contribution

This work describes the LEAP system submitted to the CHiME-5 Automatic Speech Recognition (ASR) challenge (Track A-1 i.e, single-array track).

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.

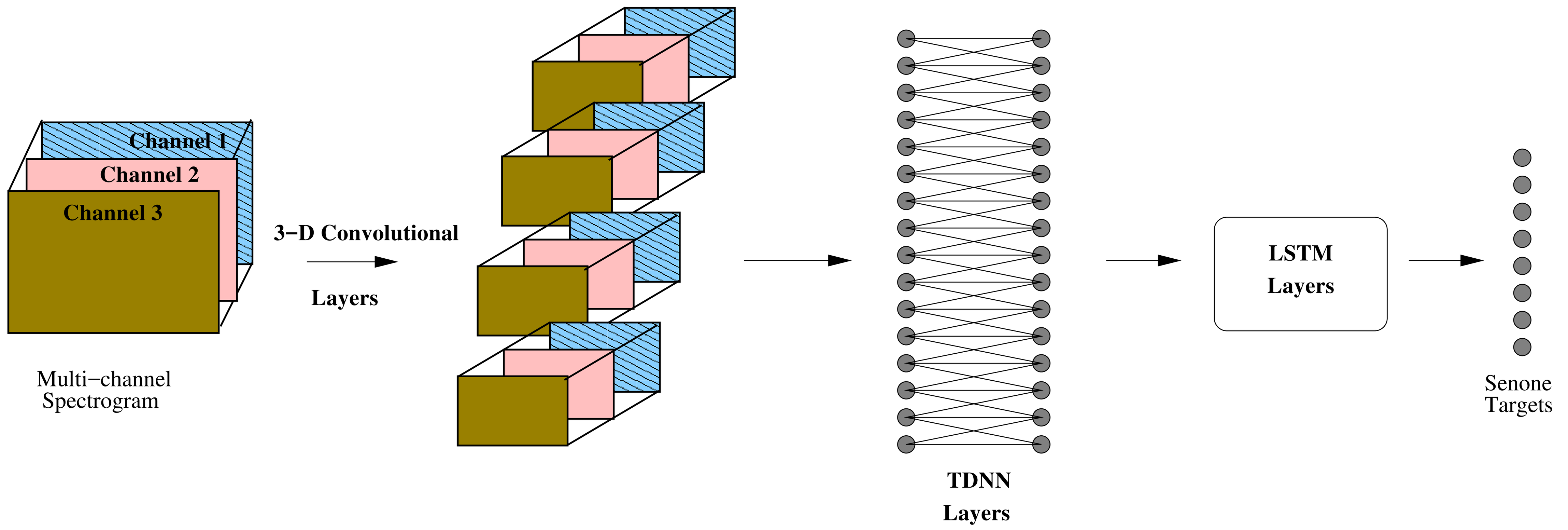

Figure 1: The acoustic model used in the LEAP system consisting of CNN-TDNN-LSTM neural network. The model is trained with chain training framework in Kaldi.

- 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 [4].
- These features are based on frequency domain linear prediction (denoted as *FDLP*) approach.
- The feature extraction module is shown in Fig. 2. These features are also 40 dimensional.

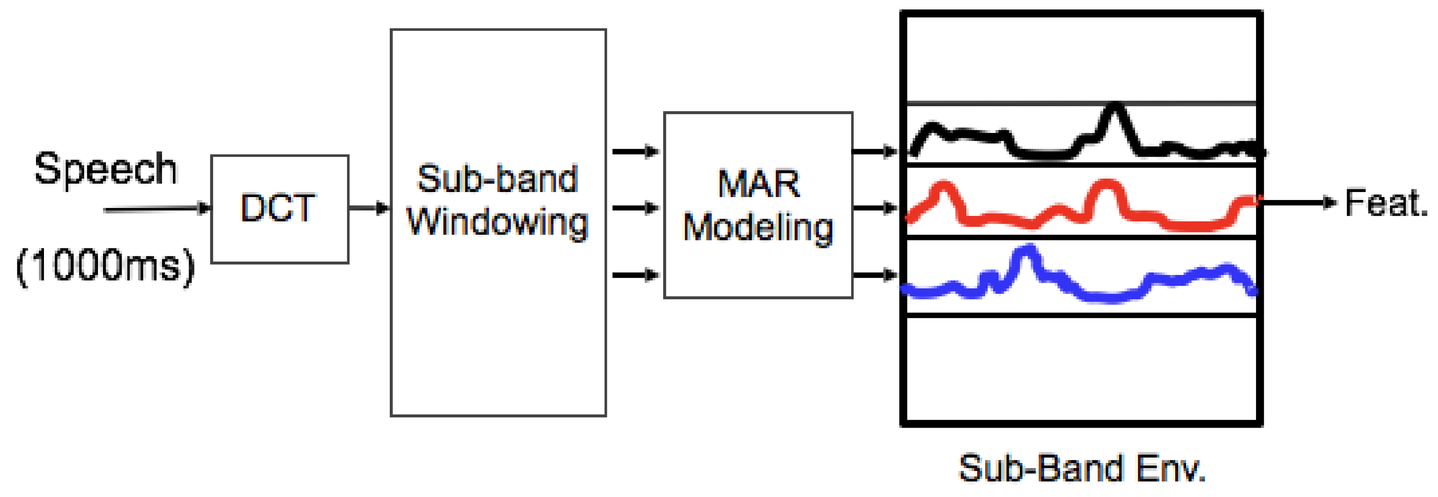

Figure 2: The feature extraction module based on multi-variate autoregressive modeling [4].

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 / Eval]-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** / **66.1**

- The system 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.
- The combination result improves the baseline system absolutely by 8% in terms of word error rate on the development data (beamformed baseline) and absolute 15 % on the evaluation data.

References

[1] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan 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, September 2018.

[2] Sriram Ganapathy and Vijayaditya Peddinti. 3-d cnn models for far-field multi-channel speech recognition. *ICASSP*, 2017.

[3] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755, 2016.

[4] Sriram Ganapathy. Multivariate autoregressive spectrogram modeling for noisy speech recognition. *IEEE Signal Processing Letters*, 24(9):1373–1377, 2017.