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 \textit{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.

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 \textit{FDLP}) approach.

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.

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

![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.](image)

<table>
<thead>
<tr>
<th>System</th>
<th>Dev-Worn Mic</th>
<th>[Dev / Eval]-Beamform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>48.0</td>
<td>81.3</td>
</tr>
<tr>
<td>System-A</td>
<td>44.1</td>
<td>75.8</td>
</tr>
<tr>
<td>System-B</td>
<td>45.5</td>
<td>77.4</td>
</tr>
<tr>
<td>Sys. Comb (A + B)</td>
<td>41.3</td>
<td>73.4 / 66.1</td>
</tr>
</tbody>
</table>

- 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


