LEAP Submission to CHiME-6 ASR Challenge
ICASSP 2020

Aniudh Sreeram, Anurenjan Purushothaman,
Rohit Kumar and Sriram Ganapathy

LEAP lab, Electrical Engineering Department,
Indian Institute of Science, Bangalore.
Outline

▶ Introduction
▶ Data set description
▶ Baseline setup
▶ Proposed method and model description
▶ Experimental results
▶ Summary
Introduction
Introduction

- Automatic Speech Recognition (ASR) find widespread applications in human-machine interface, virtual assistants, smart speakers etc...
- Recognition of speech in noisy and reverberant conditions continues to be a challenging task.
- In this paper, submitted by LEAP lab, to the CHiME-6 ASR challenge, Track 1 tries to reduce the effects of noise and reverberation on the ASR by using extensive data augmentation coupled with Factorized Time Delayed Neural Networks (FTDNN) based acoustic model.
- We also discuss about the combination of FTDNN and Long Short Term Memory (LSTM) at the acoustic model level.
Data set description
Data set description

▶ The CHiME-6 data set consists of distant microphone conversational speech elicited using a dinner party scenario.
▶ The transcriptions and signals are generated from CHiME-5 data set.
▶ The training data consists of 79,980 utterances, spoken by 32 speakers, yielding 40.33 hours of data.
▶ The dev data consists of 7,440 utterances, spoken by 8 speakers, yielding 4.27 hours of data.
▶ The eval data consists of 11,028 utterances, spoken by 8 speakers, yielding 5.12 hours of data.
Baseline setup
Baseline setup

- The multichannel data is augmented with noise and artificial reverberation using 5 small and medium rooms.
- Mel Frequency Cepstral Coefficients (MFCC) based features is extracted and a mono-phone model is trained.
- Using the alignments from the mono-phone model a tri-phone model is trained.
- Another tri-phone model is trained with the addition of linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT).
- Further, speaker adaptive training (SAT) is also included, which is the final HMM-GMM model.
Baseline setup

- 3-way speed perturbation is performed on the data, yielding 15 copies of the original training data.
- An acoustic model with 15 layer Factorized Time Delayed Neural Network (F-TDNN) with lattice free MMI cost function is trained, using the alignments from the HMM-GMM model.
- I-vector is also used to provide speaker specific information to the acoustic model.
Proposed method and model description
Proposed method and model description

- Experiments were performed with data augmented with noise, reverberation and 3-way speed perturbation.
- MFCC features is extracted for the augmented data and various acoustic model architectures were used for the experiments.
- The F-TDNN layers were increased to 18 layers and additionally 3 layers of LSTM is employed. However, this architecture did not provide better Word Error Rate (WER).
- The submitted model contains 18 layers of F-TDNN, which improved the WER by 2% over the baseline.
Experimental results
Table: Various model architectures and Word Error Rates (WER) % results for the CHiME-6 development set.

<table>
<thead>
<tr>
<th>Models (# layers)</th>
<th>S02 DINING</th>
<th>S02 Kitchen</th>
<th>S02 Living</th>
<th>S09 Dining</th>
<th>S09 Kitchen</th>
<th>S09 Living</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaldi-Recipe-Results¹</td>
<td>53.80</td>
<td>56.47</td>
<td>47.78</td>
<td>53.76</td>
<td>50.30</td>
<td>49.92</td>
<td>51.75</td>
</tr>
<tr>
<td>HMM-GMM</td>
<td>84.39</td>
<td>86.12</td>
<td>78.63</td>
<td>86.00</td>
<td>83.67</td>
<td>84.17</td>
<td>83.27</td>
</tr>
<tr>
<td>F-TDNN (15)</td>
<td>55.40</td>
<td>57.78</td>
<td>48.49</td>
<td>54.92</td>
<td>50.95</td>
<td>50.95</td>
<td>52.79</td>
</tr>
<tr>
<td>F-TDNN (18)</td>
<td>53.53</td>
<td>55.50</td>
<td>46.64</td>
<td>53.45</td>
<td>50.31</td>
<td>48.93</td>
<td>51.04</td>
</tr>
<tr>
<td>F-TDNN (18) LSTM (3)</td>
<td>56.80</td>
<td>59.79</td>
<td>49.91</td>
<td>57.68</td>
<td>52.65</td>
<td>53.42</td>
<td>54.65</td>
</tr>
</tbody>
</table>

Summary
Data augmentation with noise, room reverberations and 3-way speed perturbation along with F-TDNN based acoustic model improves the WER from the HMM-GMM model.

Addition of extra 3 layers of F-TDNN provides 2% improvement over the kadli baseline recipe.

However, addition of LSTMs over the F-TDNN did not improve the WER.
Thank you