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 experients.
- 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

Experimental results

Table: Various model architectures and Word Error Rates (WER) % results for the CHiME-6 development set.

Models (# layers)	S02 DINING	S02 Kitchen	S02 Living	S09 Dining	S09 Kitchen	S09 Living	Overall
Kaldi-Recipe-Results[1]	53.80	56.47	47.78	53.76	50.30	49.92	51.75
HMM-GMM	84.39	86.12	78.63	86.00	83.67	84.17	83.27
F-TDNN (15)	55.40	57.78	48.49	54.92	50.95	50.95	52.79
F-TDNN (18)	53.53	55.50	46.64	53.45	50.31	48.93	**51.04**
F-TDNN (18) LSTM (3)	56.80	59.79	49.91	57.68	52.65	53.42	54.65

[1] Shinji Watanabe et al. "Sixth CHiME Challenge - Evaluation Plan". In: (2020). URL: https://chimechallenge.github.io/chime6/.

Summary

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