Speaker-Targeted Speech Recognition

Henry Yeh, Gordan Gao

October 6, 2021

1 Introduction

Robust speech recognition in rowdy environments raises challenging tasks in Automatic Speech Recognition (ASR). While current ASR systems have competitive performance in a low noise environment, their performance stability when background noise or background speech is present. Currently, a potential solution is to build a pipeline of speaker embedding models to recognize the speech of a target speaker from a mixture of speech signals. In order to improve the efficiency of the speaker embedding models that have been developed, this work will focus on studying current development of speaker embedded models and ways to improve such models on top of the current solutions to achieve improved system robustness.

2 Literature research

For speaker identification and verification task, there are several speaker embedding models. These model train a representation of speakers through the process of stochastic characteristics or classifing the speech utterance signals to targeted speaker. Conventionally, Gaussian Mixture Model-Universal Backgroud Model (GMM-UBM) based i-vector (Denhak et al., 2010) [**?**] use lots of signals from different people to train a prior GMM model then fine-tune to target speaker. However, the aucoustic features based on gaussian assumption has some limitations. To conquer this problem, Deep Neural Network (DNN) based speaker embedding model such as x-vector (Synder et al., 2018) [**?**] are proposed to replace the GMM with neural network hidden to extract the speaker acoustic features. With the powerful speaker embedding feature, the performance of speech recognition systems can be further improved (Garimella et al., 2015) [**?**]. Furthermore, the DNN aucoustic model can be used in noisy and reverbent environment to enhance the robustness of speech recognition.

3 Method

With the knowledge that DNN aucoustic speaker embedding can be used in speaker identification and ASR system, we want to enhance the capibility of speaker representation in noisy environment. Thus, we are going to adapt Denoising Autoencoder (DAE) to improve x-vector for speech recognition. We can do speech enhancement by using DAE before training the x-vector, or just apply DAE on x-vector to eliminate the channel or environment effects.

4 Dataset

WSJ0

- Sample Type: 1-channel pcm compressed
- Sample Rate: 16000
- Data Source(s): microphone speech
- we will use wsj0-2mix (the mixed up version of wsj0) for speaker targeting ASR

Switchboard-1 Release 2 (optional)

- Sample Type: 2-channel ulaw
- Sample Rate: 8000
- Data Source(s): telephone conversations
- about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female)

5 Evaluation

- WER on WJS0 without our speaker-targeting embedding (x-vector, i-vector)
- WER on WJS0 with our speaker-targeting embedding
- WER on WJS0-2mix without our speaker-targeting embedding
- WER on WJS0-2mix with our speaker-targeting embedding

6 Timeline

Sep. 30 - Oct. 15: Literature research, methodology clarification Oct. 15 - Oct. 22: Data collection, processing, and trainig tools learning Oct. 22 - Nov. 5: Implementing proposed algorithm, building base line (Mid-term report) Nov. 8 - Nov. 30: Experiment Nov.30 - Dec. 3: Preparing poster and writing report