536 Final Project Proposal

Yue Cao, Zhengyang Yuan, Tianyao Zhang

March 30th, 2016

1 Team Members

- Yue Cao (NETID: yc775), Team Leader
- Zhengyang Yuan (NETID: zy113)
- Tianyao Zhang (NETID: tz133)

2 Convolutional Neural Network for Speech Recognition

Convolutional Neural Network (CNN) is a very versatile technique mainly for Computer Vision tasks. Recently, CNN has significantly improved image classification and object detection accuracy. But in the domain of Automatic Speech Recognition (ASR), although networks such as RNN and ANN are being incorporated into many speech recognition models, CNN did not play a significant part.

In our project, we are going to apply CNN to speech recognition task. Our project will first converting audio data into spectrogram, which is a efficient visualization of audio. Each word would become an "object" in this spectrogram. We then would build and train a revised CNN, with its maturity and dexterity in image recognition, to detect and recognize the "objects" in the "image" of speech. CNN has shown a strong ability for recognizing objects in challenging environments with different lighting and partial occlusions, thus, it could have high potential to improve the performance of ASR by distinguishing noise and handling frequency shifts, which are common but annoying in speech signal.

3 Data Set

We plan to use a subset of TIMIT data for our project, which is designed to provide data for acoustic-phonetic studies. It contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance. The TIMIT corpus transcriptions have been hand verified. Test and training subsets, balanced for phonetic and dialectal coverage, are specified.