Speaker ID with the TIMIT Dataset

Team members: Jessuca Centers, Xinlin Chen, and Michael Martinez

December 6, 2019

Abstract

The TIMIT database contains clean audio from 630 speakers (192 females, 438 males) categorized into 8 American dialects. These speakers each read ten phonetically rich sentences. The goal of this project was to evaluate various neural network (NN) architectures and feature spaces for the purpose of speech-independent speaker identification on this dataset. We evaluates three feature spaces, which we describe as our three methods. The first method was to extract common speech features such as mel-frequency cepstral coefficients (MFCCs), delta, and deltaDelta coefficients gathered from the audio clips and use those as the input to a neural network. The second method required generating the audio spectrograms and using those 2-D representations of the speech as input to a neural network. The third method takes an end-to-end deep neural network approach and uses the raw audio data as the input. The use of an autoencoder was also explored. For simplicity, the number of speakers considered was reduced to 20, which were randomly selected. Based on our experiments, the best feature space to use in a speaker ID neural network was [insert result] as it provided [insert result] accuracy for the 20 speaker subset.

1 Introduction

1.1 Motivation and importance of the problem

Speaker identification (ID) is critical to authentication and surveillance applications. In many scenarios, including over-the-phone listening, speaker identification is limited to only the acoustic-based identification approach. Scenarios in which other biometric measurements would be used to perform speaker identifcation can easily be argued. For the sake of this project, it is assumed that the acoustic data of a person speaking is the only biometric measurement available and/or favorable for a speaker ID classifier. Initial speaker ID classifiers did not use neural networks (NNs), however, the performance of NN based approaches have surpassed traditional approaches according to the technical literature. For that reason, the goals of this project were to evaluate the performance of different:

1. input feature spaces to a speaker ID neural network
2. speaker ID neural network architectures for specific input feature spaces

1.2 Description of the TIMIT Database

2 Related Works

As mentioned in the abstract, three feature spaces were considered for a speaker ID neural network classifier. The literature used for each of these methods are described below.

2.1 Method 1: Related Works

2.2 Method 2: Related Works

2.3 Method 3: Related Works

3 Details of the Project

For all methods that were experimentally evaluated for this project, a few data organization techniques were kept common.

3.1 Data Splitting

3.2 Scaling for the Number of Speakers

3.3 Contribution of each Member of the Team

This section includes each team member's descriptions of their contribution to this project.

3.4 Jessica Centers

3.5 Xinlin Chen

3.6 Michael Martinez

4 Experimental Results

4.1 Method 1: Results

4.2 Method 2: Results

4.3 Method 3: Results

5 Concluding Remarks

References