Intelligent Voice Identification with Neural Networks

CS 230 Winter 2018 Final Project
Category: Speech Recognition

Lloyd Maza	Michael Thompson	David Troner
lmaza	mt1	dtroner

Project Proposal

This project aims to expand upon typical speech recognition to voice recognition, where the person who is speaking can be identified independent of what is being said. We intend to tackle this problem using a standard multilayer perceptron trained to identify specific people based on a prior recording of their voice. Ultimately we believe this utilization of neural networks for speaker identification include practical implementations such as identity verification for over-the-phone bank transactions or more commonly speaker diarization for speech-to-text applications.

Given that this problem is at its core a multiclass classification problem, we propose the use of a deep neural network with an output layer activated by the softmax function as a baseline to perform the voice recognition task. Before this can be done, the speech data also will need to undergo considerable pre-processing and feature extraction. Last September, Ge et al. demonstrated an effective voice recognition system for identifying speakers in a large group of people (>300) with a shallow neural network [1]. We intend to narrow the scope of our project compared to theirs, using a smaller group of people (roughly 10-20 speakers) to make the data collection process more tractable, while using their approach as a foundation for our model, including their data pre-processing techniques. A major component of this project will be the exploration of alternative architectures and other hyperparameters to improve performance for the reduced class problem.

Key anticipated challenges include the limitation of available voice data for many distinct voices along with the fact that the neural network will have to be trained on each unique voice that the user wishes to identify. Additionally, the network will likely have to be rather deep to capture subtle distinctions in voices, and a high level of hyperparameter tuning will likely be required for high performance.

In light of this, a proposed minimum viable product would be a network that correctly identifies 9 out of 10 people from training on 60 seconds of those individual's voices. The goal would be to extend this to higher test performance along with less than 60 seconds of training audio per unique voice through tuning of hyperparameters and careful selection of neural network architecture in addition to regularization and other techniques from the class. By the end of the course, we hope to demonstrate the efficacy of our system by training the network on clips from a movie, then having it identify speakers from different clips in the same movie.

References

[1] Z. Ge, A. Iyer, S. Cheluvaraja, R. Sundaram, A. Ganapathiraju, "Neural Network Based Speaker Classification and Verification Systems with Enhanced Features," in *Intelligent Systems Conference, London, UK, September 7-8, 2017.*