Speaker Determination with Neural Networks

Sparsh Agarwal, Wayne Chew, Newton Wolfe, Keshav Sharma
NetID: agarwal39, mchew2, npwolfe, ksharma28

Motivation

It has long been science fiction for every device to be able to interact with humans via voice control, and with the advent of cloud computing and large-scale machine learning, voice based interfaces are increasingly able to understand and act upon human speech. However, many of these interfaces are not specific to certain speakers, even in situations where that would be desirable. For example, when a person asks an Amazon Echo or Google Home device what their day will look like, it's important for the device to be able to understand which of multiple users is speaking, and reply with that context as a human would. Our task is thus to distinguish multiple voices reliably based on the sounds of the voices, by training on recorded voice samples.

Dataset

We will be training primarily on audiobooks sourced from LibriVox, a repository of public-domain audiobooks. Audiobooks are convenient because they have generally similar audio recording parameters (reasonable quality microphones, mastered to similar quality) and they are tagged with speakers; additionally, there's an enormous corpus on which we can learn. The audiobooks are divided into chapters with speaker annotations on each chapter; we will then break each chapter into 50 millisecond fragments, then apply a Fourier transform (1) to provide the audio in a format that is easier for a neural network to process. (2) The use of recurrent neural networks, as in Baidu's Deep Speech architecture (Hannun et. al. 2014) allows for time-series data to be processed more efficiently than looking at small chunks of time; we theorize that this time-series data will encode much more unique information than single chunks of audio.

Computational Cycles

CPU and GPU time will be acquired via the student plans at the class' preferred cloud vendor, like Amazon or Google. By restricting our scope from doing voice recognition to voice distinction, we hope to cut down the otherwise enormous training times, even with good GPU acceleration; in Amodei et. al. (2015) Baidu discussed that training an end-to-end deep neural network for voice recognition took 12 exaFLOPS, which is clearly infeasible for our education project.

Software Usage

We plan to use TensorFlow to build and run our neural networks. This software was chosen due to its standing as a de facto standard and the multitude of examples using TensorFlow to do audio processing, which may be helpful for debugging.

Experimental Methodology

We are planning to start with a simple fully connected neural network and convolutional neural network architecture that we used for previous lab projects. Once we are sure that the voices can be predicted using these methods (do not need high accuracy), we will move on to using a recurrent neural network and maybe a Long Short Term Memory neural network. We will also try out existing models or libraries and compare it with our neural network.

References

1) https://betterexplained.com/articles/an-interactive-guide-to-the-fourier-transform/
2) https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a
3) http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/
4) Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

5) Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., ... & Elsen, E. (2015). Deep speech 2: End-to-end speech recognition in english and mandarin. arXiv preprint arXiv:1512.02595.
6) http://andrew.gibiansky.com/blog/machine-learning/speech-recognition-neural-networks/