Metis Week 9 - Project Kojak Proposal

Speaker Recognition in Noisy Environments

Steven Bierer, November 20, 2018

Data Resources and Tools

For the final project, I will analyze a type of data for which I have some familiarity, but which I have not explored deeply: speech. The ability to automatically recognize the source of speech (i.e. the identity of a human speaker) has growing commercial interest as more homes have speech-enabled computer systems like the Amazon Echo. For instance, such devices may detect the *content* of speech (i.e. the speech recognition problem) better if they can detect *which* of many household users is issuing a command, and to begin training on a voice if is from a previously unknown source. Speaker recognition also has important applications in security. However, noisy environments can make the training and ongoing performance of such systems very difficult. Fortunately, machine learning algorithms has the potential to make speaker recognition both robust and fast.

There are many large data sets with annotated human audio recordings, but I will first explore the 4th "CHiME Speech Separation and Recognition Challenge" (http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/data.html). The data consists of thousands of audio recordings made in quiet and realistic noisy environments (e.g. on a bus, in a cafe); the recordings represent hundreds of scripted utterances made by several dozen speakers. Recordings from 1 to 6 channels (i.e. with microphone arrays) are available, however I plan to focus mainly on the most challenging single-channel recordings.

Analysis will be performed in Python using MondodB to store the large amounts of data on a remote server. There exist numerous tools for handling audio files and analyzing speech, such as Vgg and pyAudioAnalysis, so I will use these where appropriate.

Analysis Approach

I plan to take the following steps with the data, though depending on outcomes this workflow will likely be altered. The overall strategy is to ramp up the level of difficulty in extracting speaker identification, first from noise-free "clean" recording environments, then from noisy environments.

1) Import the data into a MongoDB collection on a remote AWS server.
2) Connect to the remote server via ssh using Jupyter Notebook.
3) Load in speech waveforms, one at a time, and perform short-term Fourier transforms and other spectral-temporal analyses.
4) From a subset of noise-free recordings of single speakers, train a nonnegative matrix factorization or neural network model. This will be repeated for 5-6 total speakers.
5) Test speaker identification on clean recordings using new utterances from the same 5-6 speakers PLUS some speakers that were not trained on (for control).
6) Test speaker identification on noisy recordings, using source separation methods to "denoise" the signal.
7) Train the model on noisy data, to see if this reduces performance on testing in either clean or noisy environments.
8) Finally, it would be interesting to see how quickly the system can detect that a speaker is not one from the training corpus. Is it required to play several utterances in full, or can the system determine if the speaker is unfamiliar after a few milliseconds?

References

Sun DL et al. Universal speech models for speaker identification single channel source separation. https://ieeexplore.ieee.org/document/6637625

Zegers J and Van hamme H. Joint Sound Source Separation and Speaker Recognition. https://www.researchgate.net/publication/307889564_Joint_Sound_Source_Separation_and_Speaker_Recognition

Wikipedia - Speaker Recognition. https://en.wikipedia.org/wiki/Speaker_recognition

Models for AudioSet (Vgg). https://github.com/tensorflow/models/tree/master/research/audioset