(Tentative) Final Project: Automatic speech recognition for speakers with dysarthria

1. Dataset of choice

The TORGO Database: Acoustic and articulatory speech from speakers with dysarthria

http://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html

Rudzicz, F. (2012). Using articulatory likelihoods in the recognition of dysarthric speech. Speech Communication, 54(3), March, pages 430--444.

2. Methodology

a. data preprocessing:

- data on the articulatory positions (articulography) will not be used for this project due to its complexity
- I think for now I will only focus on making the model recognize single words command (i.e., numbers, directions, etc.)
- key information: recordings of speech (as wav files), orthographic transcription of speech (as txt files), phonemic transcription of speech with time alignment (as txt files)
- Preprocessing:
  1. filter for the words of interest
  2. extract the audio files for the part where the word is spoken (using the phoneme-time alignment files)
  3. associate the audio file with the corresponding words
  4. clean the audio data (e.g., standard duration, filter noise)
  5. get the spectrogram for the audio
  6. down-sample and normalize the spectrograms so that the model can train reasonably well and fast
  7. data augmentation techniques proposed in https://www.tensorflow.org/io/tutorials/audio could be implemented as well, depending on the results

b. ML model:

- Desired prediction/estimate: Estimate what (single) word is being spoken from a given audio recording of a patient with dysarthria
- Proposed ML model: Convolutional neural network. The reasoning is that the spectrogram is a 2D representation of an audio recording (component frequencies over time), where each “pixel” (if seen as an image) is the magnitude of that frequency at that moment in time. CNNs are performant at learning patterns in image-like data. CNNs are computationally intensive however, so it will be necessary to down-sample the input data for training times to be reasonable.
- Alternative models: More advanced models could be used for actual Natural Language Processing, such as Hidden Markov Chains (HMMs) or recurrent networks. If possible (given enough time and resources), these could be integrated at a later stage to recognize sentences based on phonemes and word probabilities (TORGO has the data for it). Maybe transfer learning can be used as well (from a pretrained model for typical speech) – the concern I have is that I only have 8 speakers with dysarthria, and they only recorded each word of interest once, so maybe the data I have will not be “strong” enough to influence the pre-trained model.

c. Evaluation Metrics:

- Confusion matrix should be interesting (as we are categorizing spoken words).

d. final conceptualization:

- Since I don’t have a dysarthria patient to do a live demo, I think I will have some audio clips that are not used. People can listen to the clips, and then the network will predict/recognize what it says.