IFT 6390: Project Proposal
A Deep Learning Approach to Speaker Accent Classification

Jonathan Bihmani-Burrows Khalil Bibi Arlie Coles Akila Jeeson Daniel
Y. Violet Guo Louis-François Préville-Ratelle

The task of speaker accent recognition from audio data has several potential applications, including improvement of automatic speech recognition (ASR) systems, speech synthesis where a particular accent is desired, and accent coaching and correction. While preliminary results in this domain have been achieved by SVM and GMM approaches, the application of deep learning models to this task has been less explored (Omar and Pelecanos 2010, Ge 2015). Hence, we are motivated to pursue automatic accent classification using deep neural networks on several English variants.

We propose working with (North) American vs. British English, particularly because these variants are the most well-sourced. (If we are able to find a dataset with enough clean audio from another variant of English, we will include it also in our classifiers.) For the (North) American English, we will work with the Librispeech corpus, a free corpus assembled from recordings of audiobooks (Panayotov et al. 2015). For the British English, we will work with a subset of Audio BNC, the spoken component of the British National Corpus (Coleman et al. 2012). This subset contains lectures, sermons, and other single-speaker audio, making it a good match for comparison against Librispeech.

For the feature-extraction stage of our approach, we would like to try three approaches: first, the standard Mel-Frequency Cepstral Coefficient filterbank featurization often used in ASR systems; second, the extraction of first, second, and third vowel formants of voiced phonemes in the audio (since formant frequencies have been shown to be indicative of accent, Hansen and Arslan 1995); and third, the raw spectrogram data of the audio.

Finally, for the classifiers themselves, we would like to start by training a baseline SVM. Then we would like to train a Recurrent Neural Network on the above features of the audio, since RNNs are capable of capturing sequential information potentially important to identification of accent in a string of speech. We would also like to train a Convolutional Neural Network, particularly on the raw spectrogram features, which can be thought of as an image, in order to better understand the importance of temporal information and of human-invented cues (MFCCs, formant frequencies) for accent classification in a deep learning context.

We also remark that interesting applications to build "on top" of our classifiers, if any time permits, could include online accent correction in the case of the RNN, which could identify precisely where a speaker's attempted accent falls short of the mark, or speech synthesis/neural accent transfer in the case of the CNN trained on spectrogram features, which could generate new audio in a desired accent.

References

Coleman, John et al. (2012). "Audio BNC: the audio edition of the Spoken British National Corpus". In: *Phonetics Laboratory, University of Oxford*.

Ge, Zhenhao (2015). "Improved accent classification combining phonetic vowels with acoustic features". In: *Image and Signal Processing (CISP), 2015 8th International Congress on*. IEEE, pp. 1204–1209.

Hansen, John HL and Levent M Arslan (1995). "Foreign accent classification using source generator based prosodic features". In: *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on.* Vol. 1. IEEE, pp. 836–839.

Omar, Mohamed Kamal and Jason Pelecanos (2010). "A novel approach to detecting non-native speakers and their native language". In: *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* IEEE, pp. 4398–4401.

Panayotov, Vassil et al. (2015). "Librispeech: an ASR corpus based on public domain audio books". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on.* IEEE, pp. 5206–5210.