Seong-Eun Cho
Kyle Roth

IPA transcription

Transcription of spoken language is a difficult process. Field linguists spend their lives training their ear to catch small differences in sound, recording them in the International Phonetic Alphabet (IPA). Recorded audio has helped little. Automatic phonetic transcription has been attempted by [1], [2], and many others, but was limited to one language and didn't use neural networks. [3] used temporal flow networks to achieve 80% accuracy, but on a dataset of only 37 words spoken over 2.5 hours. Their approach used an array of networks estimate phonetic features and then predict phonemes. We feel that these approaches were limited by the lack of data and their specificity to one language. We will attempt this task using unsupervised learning, allowing us to learn on varied, extensive, unlabeled data.

Data sources and preparation. To accomplish this task, we need speech data that is reasonably clean and from a variety of speakers and languages. General conference addresses fit that description best. Other options include VoxCeleb and YouTube-8M. YouTube-8M contains lots of non-speech data, so we will probably not use it unless necessary. Useful features extracted from the data for the neural network could include the frequency spectrum calculated using the FFT, the double FFT, or the cepstrum. To calculate accuracy and precision, we may ask the linguistics department for handmade phonetic transcriptions from field work which we can use as a test set.

Possible method 1. One possible approach involves a network that takes a sequence of features for one utterance, and outputs a sequence of timestamps that approximate phoneme boundaries. Normally, this approach would need to be supervised by a phonetic transcription of the text, but we can design a loss function by taking the segments between the approximate boundaries, mapping them to phoneme embeddings, and comparing them to their nearest neighbors in phoneme space. The loss can be defined as the distance to all the segments' nearest phoneme neighbors in phoneme space.

Possible method 2. Another possible method is to extend on the idea of a Generative Adversarial Network (GAN), but instead of having a generative network, we have a segmentation network which could take as input an audio data and produces some number of phoneme sequence. Each of these phoneme will then be used as inputs of an adversarial network which could determine whether or not the segmented audio sounds like a phoneme or not. The idea is that the segmentation network will learn to better segment the audio while the adversarial network gets better at identifying false phoneme boundaries. The problem with this method is that we need a baseline data of phoneme samples in order to create a phoneme distribution which we don't have.

Work outside the scope of this project. Because this project requires more work than a usual data project, we propose dividing some of this work into projects for other classes. Seong will do language identification on our dataset for his project in C S 501R, which will help him understand the shared phonetic characteristics between languages. Kyle will experiment with creating phoneme embeddings for his 501R project, giving us the algorithm for the loss function in the methods described above. Throughout the process, we may periodically ask for linguistic insights from Dr. Lonsdale of the linguistics department, and Kyle's coworkers at Cobalt Speech and Language. Seong will focus on scraping, cleaning, and analyzing General Conference data in multiple languages, and Kyle will focus on ways to identify phoneme boundaries and embed phonemes. Next semester we will attempt the methods mentioned above and try to solve the problem in its entirety.

[1] Van Bael, et al. "Automatic phonetic transcription of large speech corpora." Computer Speech & Language, Volume 21, Issue 4. 2007 Oct. Pg 652-668.

[2] Schiel, Florian. "Automatic phonetic transcription of non-prompted speech." Department of Phonetics, University of Munich, Germany. https://epub.ub.uni-muenchen.de/13682/1/schiel_13682.pdf.

[3] Chang, et al. "Automatic phonetic transcription of spontaneous speech (American English)." International Computer Science Institute, UC Berkeley. ICSLP 2000, Beijing, China.