Machine Learning Engineer Nanodegree

Capstone Proposal for Sub-Vocal Recognition

Brian Coe
April 20, 2017

Domain Background

Sub-vocal recognition (SVR) is a method of interfacing with computers that uses external readings of nerves involved in production of speech to synthesize text or voice without actively voicing the speech. The primary nerve involved is the glossopharyngeal nerve, innervating the pharynx and basal lingual muscles. Electrodes are usually placed on the outside of the neck, and the user engages in voiceless speech without opening their mouth. Electrical signals are read from the electrodes, and transformed into text corresponding to the phonemes the user sub-vocalized. NASA's project in 2004 demonstrated both the feasibility and difficulty in this type of computer interface.[1] Their system sometimes achieved as high as 92% accuracy in recognizing phonemes during sub-vocalization, but would sometimes drop to less than 50% accuracy even for the same user on a different day. The main obstacles involved in SVR are (1) achieving high signal resolution (voltage and temporal), (2) noise- and placement-tolerant pattern recognition, especially feature extraction, and (3) anatomical access to appropriate nerves or muscles [2]. Similar obstacles are found in speech recognition, but SVR uses voltage data rather than acoustic data with additional complexity arising from the specifics of data gathering. SVR is, however, more tolerant to environmental acoustic noise, and can be used when voiced speech is impossible or undesirable.

Advances in hardware and software mitigate the obstacles NASA faced in 2004. First, high signal resolution is made possible by the widespread availability of high sampling frequency (1kHz+) 8-bit+ differential voltage analogue amplifiers [3]. Second, filtering and signal processing are usually accomplished in speech recognition with the fast Fourier transform (FFT) and related techniques for recognizing phonemes [2]; therefore, these techniques should prove useful for sub-vocal recognition in processing voltage data. Third, the muscles innervated by the aforementioned nerves of interest are sufficiently large, near the dermal surface, and unobstructed by bone or adipose to register on a simple electromyograph (EMG) with surface-contact electrodes [2].

Successfully solving the remaining challenge of real-time speech recognition from electromyographic data during sub-vocalization would constitute something akin to artificial telepathy, thus allowing communication between individuals without voiced speech. If the technique can be made sufficiently accurate and unobtrusive, then the impact is potentially widespread and profound.

Problem Statement

The focus of this project is to construct a model capable of classifying sub-vocalized phonemes given voltage data from the EMG and lists of phonemes generated from passages of text. Our text samples serve as targets for supervised learning which use common English words and phrases, and are available on the creative commons [4]. The model will have to learn how phonemes appear in the EMG data. Once a model has been trained to classify each 50 millisecond segment of EMG data as containing a specific phoneme, it should be able to identify phonemes in new, unseen EMG data. Accuracy of the system can be determined through F Score, a standard in this field [2], given what phonemes the user actually sub-vocalized compared to what the system was able to infer from voltage data. Some model parameters may need to be tuned or trained on a per-user basis, much as an acoustic speech model is tuned to a particular user, but virtually everyone with functioning speech muscles should be capable of producing the necessary inputs. These inputs can be captured with industry standard AgCl or conductive polymer dermal electrodes [2] and high gain, high frequency amplifiers with analogue to digital conversion.

Real-time identification of phonemes from EMG data, as proposed in this project, would serve as a foundation to full sub-vocal speech recognition. The remaining challenges of large-vocabulary speech recognition would be addressed in subsequent project iterations. The present work will focus exclusively on reconstructing phonemes from EMG data.

Datasets and Inputs

Input data consists of time-series voltage data from EMG. This input data is in the form of sequential voltage readings. It will be read from CSV, put into a dataframe, Fourier transformed and segmented by fixed timing increments (50ms each). Additional data for labels consists of select samples of text from Austen's Sense and Sensibility [4]. The outputs are encoded using NLTK and Sci-kitLearn first into phonemes and then into sparse arrays whose columns represent articulatory features (AF), the unique phonological aspects of each phoneme such as plosive, dental, etc. Our input data measures speech muscle activation intensity and timbre during sub-vocalization. Obtaining the data will involve a raspberry pi computer with a PCF8591 analog board [5] for recording the analog electrical signals produced by speech muscles. The analog board will be configured to record from two mono-polar dermal electrodes in differential voltage mode, obtaining the voltage difference across a small ~4cm area of the neck near the carotid artery. A separate CSV file is recorded for each sample, beginning with approximately two seconds of silence followed by deliberate reading [2] (without full vocalization) of a sample. This is repeated a total of three times, followed by approximately two seconds of silence before closing the write-to-file pipeline.

The data have been generated by me, using my own biosignatures, and will inevitably include heartbeat and other myoelectric data. These are available in the MLND-Subvocal repository on my personal github [6], and have dated labels for Saturday March 4th. Each raw EMG file is about 500 kB in size, 11.7 MB in total for all 22 files. The 22 target sentences for all of these files, in order, is stored in “austen_subvocal.csv” in the same repository as the sub-vocal data, and requires a total of 2.2kB. The EMG data will serve as “input,” while the text is the “output” per supervised learning convention. The EMG data will have features extracted, and those features will be used to train a model to recognize the target phoneme outputs. For recognizing subvocalizations, this dataset appears appropriate in content, and does not present any compromise of generally established professional ethics, nor health and safety, for myself or others.

Solution Statement

In prior art for SVR, special attention was given to engineering AF's from EMG data, allowing classification of short snippets of EMG data as containing phonemes, based on each phoneme's unique phonological signature in the EMG data. [2] An approach along similar lines appears appropriate for the present work. A preprocessing pipeline will filter raw voltage data, segment it into appropriately sized temporal windows, and perform the Fast Fourier Transform (FFT) on them. The pipeline will be realized using sklearn for the pipeline itself, and numpy for the filtering and FFT. These FFT windows will serve as our observations for training classifiers which identify articulatory features and allow identification of phonemes in a data point, which will make them articulatory feature classifiers (AFC's).

A Multi-Layer Perceptron Classifier (MLPC) will be used to realize the articulatory feature classifier. MLPC can handle dense floating point arrays, produce multiple output classes, and can approximate functions of arbitrary complexity. These traits are requirements for any model to learn this system of inputs and outputs. The MLPC will be trained on frequency domain voltage data to identify and classify segments of data as containing specific articulatory features and thus a specific phoneme. A model composed of articulatory feature classifiers can be trained to find the maximum a-posteriori likelihood phonemes given previously unseen EMG observations. The solution can be measured by its accuracy in reconstructing intended phonemes from EMG data. The model can be verified using sub-vocal data from other speakers, new text passages, etc., by anyone with appropriate equipment and the ability to sub-vocalize.

Benchmark Model

Our benchmark model will be a MLPC capable of reproducing the phoneme recognition accuracy of previous inquiries, in the 0.40-0.50 mean F score range across all phonemes. This benchmark is chosen for measuring the performance of our AFC. The AFC model is a classification model, so given a number of fixed input windows from EMG data, it will output a corresponding number of predicted phonemes. We can consider false positives, false negatives, and true positives/negatives because of our use of classification. The specific value chosen is comparable to the scores obtained using single-channel inputs in the prior art [2], especially when the single channel is in the lateral pharyngeal region as in the present study.

Evaluation Metrics

The evaluation metric relevant for the chosen benchmark is F Score. For the present study, our focus on articulatory feature classification to find phonemes (rather than whole words) makes “word error rate” unattractive as a metric. F Score allows us to measure the ability of our classifier set to correctly identify phonemes. F Score with alpha 0.5 (equal weight on precision and recall) is 2PR / (P+R), where precision is P = Ctp/(Ctp+Cfp). Recall is R = Ctp / (Ctp + Cfn), Ctp is the true positive count, Cfp is the false positive count, Cfn is the false negative count. This will allow us to gauge the balance between precision and recall of our model, as well as its ability in both compared to the benchmark and prior art.

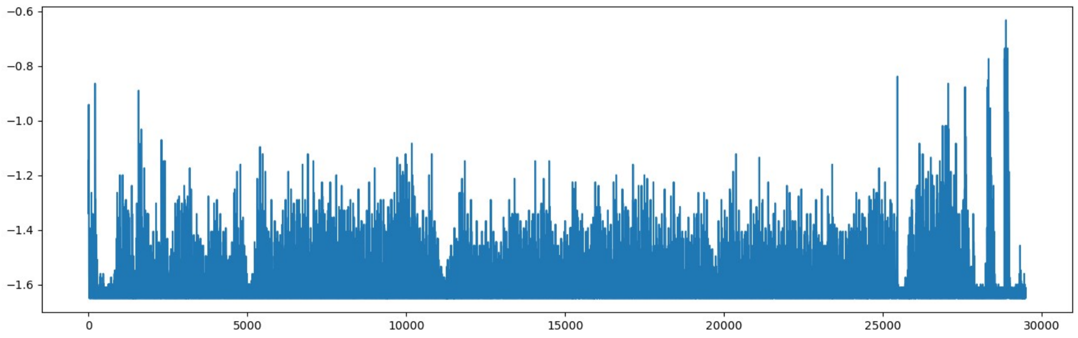
Project Design

Some sub-vocal data has already been acquired (about 22 passages, 3 trials each) but it is likely necessary to generate (much) more data, using a variation on the previously-employed technique involving a raspberry pi [5] and dermal electrodes [7]. First, the EMG data will need to be processed and segmented. Next, we will have to label each EMG segment as containing a specific phoneme. Then, we convert this phoneme into an array corresponding to AF's unique to that phoneme using NLTK. Our result as this stage is a labeled dataset that, for each data point, has an FFT array based on a 50ms segment of EMG data as 1D array of inputs, and AF's (also as a 1D array) as outputs.

We then divide this whole dataset into training and validation sets, and begin training the multilayer perceptron classifier as our articulatory feature classifer. The only goal of the MLPC is to learn to classify phonemes, by their articulatory features, based on processed EMG data snippets. If training on the manually labeled datapoints yields poor results, an F score below that of our desired benchmark, there are options for more precise mapping of EMG data and phoneme content.

If early results prove unsatisfactory, it may be necessary to bootstrap our training of an AFC using an audio channel with vocalized data that is synced (but offset) to our EMG data. The audio channel would be used as an anchor to identify where specific phonemes occur. Using a fixed offset or one that varies with AF type, we can then train the AFC model to learn an EMG pattern that most likely corresponds to a specific phoneme. We can then iterate on the resulting AFC with less precisely 'labeled' data, the original EMG data, using incremental learning classification techniques. After the initial boostrapping stage involving audio, we could move to using only EMG windows, finding windows that closely resemble what has already been learned for a specific phoneme, and adjusting model parameters accordingly. We would then run validation tests on this bootstrapped model.

Using a bootstrap approach as outlined above would also allow us to use a more precise means of identifying actual articulatory features. Speech recognition could be used on the audio component to directly identify the phonemes as they are actually pronounced, allowing us to use more specific atriculatory features that accounts for the user's accent or dialect. The phonemes and, ultimately, articulatory features extracted from the audio would then be used as the training outputs for identifying EMG features for a specific phoneme. A global or phoneme-specific offset between vocalized speech and EMG feature learning would help compensate for physiological delays between muscle activation and speech as well as differences in recording latency between the two recording techniques used.

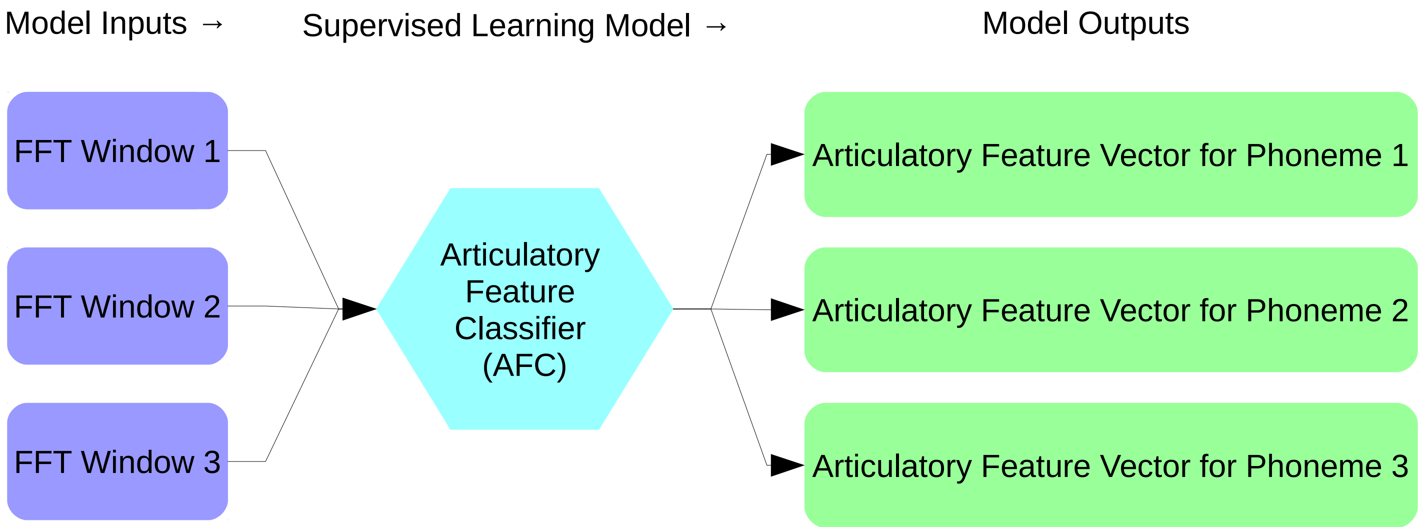


Graph 1: Raw voltage data from sub-vocal reading of a passage. Opens with two seconds of silence, followed by reading, this repeated two more times, before ending with two seconds of silence and closing. Timing is approximate. Data does appear to show some recurring motifs, even at this unprocessed level.

Preliminary analysis of the data will proceed by removing background noise from AC power, heartbeat, etc., sampling FFT windows in regions of data where sub-vocalization is likely to have occurred, and comparing these between different passages as well as for the same passages during different trials using statistical similarity techniques.

As long as differences between passages appear greater than differences between trials for a single passage, the next step is to compare feature scales between different FFT windows, and scale them accordingly using sklearn's standard scaler to remove the mean and scale to unit variance to aid in supervised learning. Then, an attempt will be made at training of AFC models to identify phonemes.

Once a basic AFC model has been trained, it is possible to perform validation of our model, and determine whether to use accuracy improvement techniques during data capture (more electrodes, higher voltage resolution ADC), before model training (using other feature scaling techniques, dimensionality reduction, other preprocessing techniques like independent component analysis), or during AFC training (using more intensive feature extraction models like neural nets for modeling articulatory features).


Model Inputs →
Supervised Learning Model →
Model Outputs
FFT Window 1
FFT Window 2
FFT Window 3
Articulatory Feature Classifier (AFC)
Articulatory Feature Vector for Phoneme 1
Articulatory Feature Vector for Phoneme 2
Articulatory Feature Vector for Phoneme 3

Drawing 1: An Articulatory Feature Classifier (AFC) is trained to identify articulatory features, the phonological subcomponents of phonemes, from transformed windows of EMG data. Each phoneme has a unique articulatory feature vector, and the model will learn what EMG features correspond to each articulatory feature. The AFC allows us to use EMG features, estimate articulatory features, and ultimately recognize phonemes which correspond to the estimated articulatory feature set.

References

[1] "NASA -", Nasa.gov, 2004. [Online]. Available: https://www.nasa.gov/centers/ames/news/releases/2004/04_18AR.html. [Accessed: 18- Mar- 2017].

[2] S. Jou and T. Schultz, "EARS: Electromyographical Automatic Recognition of Speech.", BIOSIGNALS, vol. 1, pp. 3-12, 2008. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.154.6348. [Accessed: 18- Mar- 2017].

[3] T. DiCola, "Overview | Raspberry Pi Analog to Digital Converters | Adafruit Learning System", Learn.adafruit.com, 2016. [Online]. Available: https://learn.adafruit.com/raspberry-pi-analog-to-digital-converters/overview. [Accessed: 18- Mar- 2017].

[4] "nltk/nltk_contrib", GitHub, 2009. [Online]. Available: https://github.com/nltk/nltk_contrib/blob/master/nltk_contrib/hadoop/tf_idf/austen-sense.txt. [Accessed: 18- Mar- 2017].

[5] "Quick2Wire I2C Analogue Board Kit [Q2W-ANALOG] - �13.80 : SK Pang Electronics, Arduino, Sparkfun, GPS, GSM", *Skpang.co.uk*, 2017. [Online]. Available: http://skpang.co.uk/catalog/quick2wire-i2c-analogue-board-kit-p-1191.html. [Accessed: 18- Mar- 2017].

[6] B. Coe, "bwc126/MLND-Subvocal", *GitHub*, 2017. [Online]. Available: https://github.com/bwc126/MLND-Subvocal. [Accessed: 18- Mar- 2017].

[7] “1.25" Round Tan Cloth Electrodes (TYCO Gel)”, TENSpros, 2017. [Online]. Available: https://www.tenspros.com/125-Round-Tan-Cloth-Electrodes-TYCO-Gel_p_46.html. [Accessed: 18- Mar- 2017].