Machine Learning Engineer Nanodegree

Capstone Proposal for Sub-Vocal Recognition

Brian Coe
March 18, 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.

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 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 primary challenge of this project is to construct a model capable of identifying intended phonemes given processed voltage data from the electromyograph. Fundamentally, the attempt is to infer a series of hidden symbols or states (the original words) which produced the observed voltage patterns. In the general use case, the user sub-vocalizes a series of words or single words, which the system records as a time series of electrical patterns or waveforms. The system must be capable of correctly identifying phonemes or words from nothing more than this time series voltage data. Accuracy of the system can therefore be determined on a per phoneme or word basis, given what 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 dermal electrodes [2] and high gain, high frequency amplifiers with analogue to digital conversion.

Datasets and Inputs

Input data consists of time-series voltage data from the dermal electromyograph. Additional data forming the output or labels consists of select samples of text from Austen's Sense and Sensibility [4]. Our input

data measures speech muscle activation intensity and timbre during sub-vocalization, allowing us to infer sub-vocalized words. Our text samples serve as targets for supervised learning which use common English words and phrases, and are available on the creative commons [4]. 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. 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]. The voltage data will serve as “observed” data or “input”, while the text is the “target” or “labels” per supervised learning convention. The processed voltage data will serve as feature sets to train a model with text as the target labels. 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 the problem area of speech recognition, Hidden Markov Models (HMM) [2] are generally used to infer states (actual phonemes or words) from observed inputs (voice recording). The same approach appears appropriate for the present work. A preprocessing pipeline will filter raw voltage data, segment it into appropriately sized temporal windows, and perform a kind of FFT on them. These FFT windows will serve as our observations for HMM. Intermediate models will be used to map phonemes to observed data, yielding the “emission probability” for each phoneme. Data on transition and start probabilities for different phonemes in English will be required, which are available in the open source domain along with HMM implementations such as HMMLearn [7]. With known transition, start, and emission probabilities, observed data, and known phonemes, a model can be trained to find the maximum a-posteriori likelihood phonemes given previously unseen observations. The solution can be measured by its accuracy in reconstructing intended phonemes from sub-vocalization voltage 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

Speech recognition is currently a widespread consumer technology. Google on Android offers an industry standard for constructing text phonemes from recorded speech in near real time. Such an implementation would appear to offer the closest possible next-best incarnation of a solution to the problem of SVR, as the only major difference is the former's use of vocalized speech acoustic data, while the latter requires non-vocalized myoelectric voltage data. In both contexts, the problem of reconstructing phonemes from noisy data produced in a natural environment is present, and is generally solved using (a) FFT, (b) HMM, (c) models for handling emission probability, and (d) error correction techniques like spell check, Markovian word prediction, etc. For each passage of text used for validation against the benchmark, both the test model and the benchmark will have three opportunities to correctly construct phonemes given their respective type of data.

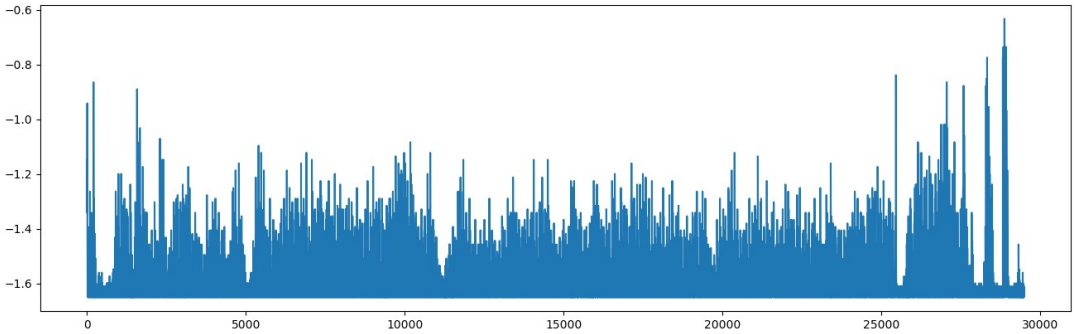
Evaluation Metrics

Quantifying performance of the solution model and comparison to the benchmark can be achieved by measuring the accuracy of each model's sequence of maximum likelihood words compared to the original sub-vocalized or read-aloud passage. The metric should be chosen to penalize incorrect words, especially superfluous words, but normalized such that values can go no lower than 0%. Therefore, it is possible to calculate the maximum of 0 and (P-I)/P, where P is the number of words in the original passage, and I is the number of incorrect words in the reconstruction. Additionally, the average can be found of three tries for each passage per model, yielding a single accuracy for each model for each group of three tries on a single passage. Since this

method accounts for differing passage size, the average of these values across all passages for each model will yield an overall model accuracy in each case.

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 [8].



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.

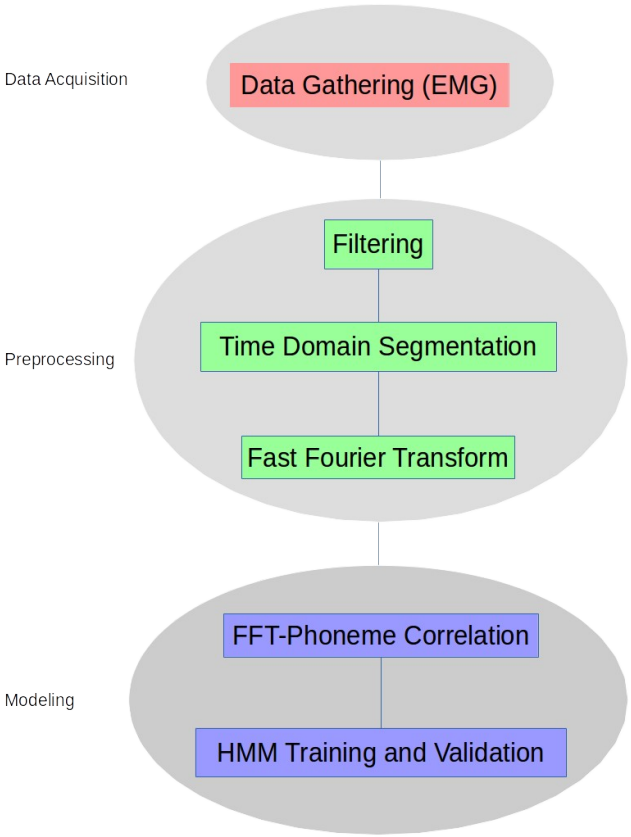


Illustration 1: Process diagram for project, showing general pattern of acquiring, processing, and utilizing data

As long as differences between passages appear greater than differences between trials for a single passage, the next step is to proceed to training emission probability models for HMM.

Once all required probabilities for HMM are found, 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), during HMM (using more intensive feature extraction models like neural nets for modeling emission probabilities), or post-HMM (Markovian modeling of word order, n-grams, etc.). If the SVR model approaches the accuracy of the prior art [2], within 10% percentage points, to a word error of 40% or lower, the present work will be sufficiently successful.

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] "hmmlearn/hmmlearn", *GitHub*, 2017. [Online]. Available: https://github.com/hmmlearn/hmmlearn. [Accessed: 18- Mar- 2017].

[8] “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].