Using Deep Learning and EMG to recognize non-audible speech

Rommel T. Fernandes
Seaver College of Science and Engineering
Loyola Marymount University
Los Angeles, CA, USA
Email: rferna16@lion.lmu.edu

***Abstract*—Many post-stroke victims deal with physiological problems such as speech impediments due to aphasia. With the advancement of Human-Computer Interface (HCI) research, this paper aims at non-audible speech recognition using Electromyography (EMG) and Deep Learning. We will first briefly introduce HCI systems such as Silent Speech Interfaces and go over how Deep Learning and Machine Learning can be used for speech recognition. Work in progress...**

***Index Terms*—Deep Learning, Electromyography, Silent Speech Interfaces, Human Computer Interfaces**

I. INTRODUCTION

Over the past few years, HCI has been an increasing field of study. Human Computer Interaction can be described as a feedback loop between human and computer. With the increased usage of sensors worn on humans, such as watches, heart rate monitors, and other smart sensors, researchers are trying to extract bio-signal information and classify typical human activities. One of the ways HCI is used is for Silent Speech Interfaces (SSI). SSI aims to use signal-extracting systems like electromyography (EMG) and electroencephalography (EEG) to convert signals of silent or non-audible speech and use a machine to classify the results. This feedback loop involves feature extraction, model training, and activity inference [1].

This paper is motivated by recent work that uses machine learning and electrocardiogram (ECG) to detect irregular heartbeats [2] and electroencephalography (EEG) [3] and EMG [4], to predict body movements.

For the purpose of this research, non-audible speech can be classified as the inability to verbalize words or sentences through the use of sound in an effective way. SSI systems are not new, what is new is the computing resources and type of algorithms used to classify speech in SSI systems. In the past, machine learning algorithms such as decision tree, support vector machine, Naïve Bayes, and hidden Markov models were used as classifiers for speech. Though powerful, they require extensive feature extraction from EMG signals. Typically, with traditional machine learning , only shallow features can be learned from those approaches, leading to undermined performance. Recently, we have witnessed the incremental development of Deep Learning, which alleviates the issue of feature engineering, since the models extract valuable information through several iterations. Therefore, using EMG signals to classify non-verbal speech does not have to a be a laborious task.

The paper is organized as follows. Section II discusses the related work to the problem. Section III discusses the experimental setup used to capture, analyze and model the data. Section IV will discuss our proposed solution. Section V will discuss the results from our analysis and compare that to other work. Finally we we will conclude this paper with a disucssion and conclusion.

II. RELATED WORK

The research conducted using EMG to predict speech for SSI systems has been going on for over two decades. Before machine learning became very popular, using EMG to predict speech involved heavy feature extraction of the data, along with discrete mathematical modeling. There have been well documented results that have used a combination of mathematical modeling and deep learning to to predict speech using EMG. Some of the research addresses syllable and single word based prediction [5], [6]. Other research has addressed using EMG to predict entire phrases [7], [8].

One of the earliest attempts to use EMG to predict speech was done in [6]. The goal was to predict isolated word recognition, which was performed on a vocabulary consisting of the ten English digits, *0-9*. Seven electrodes were positioned on the face to extract bio-signals from the subjects. Hidden Markov Models (HMM) with Gaussian Mixture Models (GMM) were used as classifiers. The study was able to get an average word accuracy of 97.3%.

In [9], newer machine learning models were introduced, such as Restricted Boltzmann Machine algorithms. Their corpus consisted of 25 sessions from 20 speakers comprising of 200 read English-language utterances such as phonemes, consonants, and vowels. In the results, the vowel features achieved relatively low ac curacies (around 40%).

The work performed by [7] carried over some of the primary research done by [9]. This research focused on using modern Deep Learning techniques such as Long Short Term Memory (LSTM) and comparing it with GMM. Their results showed that LSTM models performed better than that of GMM, with a mean Mel-Cepstral Distortion (MCD) score of 5.46 versus 5.69, where MCD is a measure of distance, and lower numbers represent better results.

In [8], they built a proof of concept SSI system that uses a one-dimensional Convolutional Neural Network (CNN) as a classifier. Seven electrodes where placed around the throat face. In their quantitative results, they got an average accuracy for all subjects of 92.01%. Their corpus included individual words, and short phrases.

Finally, work done in [2] uses latest methods to classify ECG signals by converting the bio-signals to scalograms and passing them through CNN models. This method is used to predict irregular heartbeats. Similar techniques can be adopted using this method for EMG bio-signals.

The number of electrodes and bipolar channels are far greater in [8], [9], [7], [6] than compared to our research, where we are only using three channels. This number was based on [6], which stated that EMG based speech processing requires the very least signals from the cheek area and the throat.

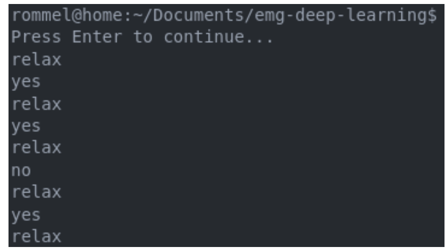
Our paper will attempt to research and contribute the following:

- Reproduce existing work for LSTM models used for speech recognition.
- Use similar techniques of wavelet tranforms and CNN, which is presented in [2] for ECG signals, and apply it to our EMG non-audible speech recognition models.

III. System Description

The system consists of two Shimmer3 EMG units (24 MHz CPU). Each EMG unit has the capability of recording two channels of data using Ag/AgCl bipolar electrodes, with a reference electrode connected to a bone-dense area. The bipolar electrodes are placed strategically based on work done in [5]. The areas where the EMG electrodes are placed are as follows. *Depressor anguli oris* (EMG1), *Zygomaticus major* (EMG2), and *Anterior belly of the digastric* (EMG3). The reasoning of only choosing three EMG channels, is to reduce discomfort by the user and abide by the minimum electrode placement documented by [6]. Each bipolar electrode of the muscle group is placed approximately 2 cm apart, based on the unit specifications in (Fig.1b). After proper placement of the electrodes on a subject, the EMG units are placed on the subjects upper torso and shoulder, using comfort straps. The EMG units transmits data via Bluetooth to a Linux (Ubuntu) Intel laptop, which captures the EMG recordings and timestamps. The EMG units utilize open source Python to transmit data to the laptop.

Capturing Data: After the subject is connected to the EMG units with the electrodes in place, the process of acquiring EMG data with annotated samples begins. In our experiments we are capturing two types of annotations for a our sample data. Our *first set* of annotations consist of the labels for the words *yes* and *no*. Our *second set* of annotations consit of the labels of the numeric digits *0-9*. The annotations are generated at random using a python script that prints out the label for the subject to read (Fig.1a). The label persists on the screen for two seconds, it is then followed by the word *relax*, which persists for two more seconds. The next label



(a) Annotated labels displayed on screen for subject to read

(b) Connections to speech-focused muscle groups for EMG data.

Fig. 1. Subjects reading rannotated labels on screen while connected to EMG units and electrodes.

in the annotations is displayed and repeated at random for a total of 50 labels per annotated set. The subject performs this task for the *first set* and *second set*. The associated EMG signals captured with the annotated labels will be used to train the various Deep Learning models which will be discussed in a later section. In total, 10 subjects are recruited to volunteer their data. For each set, the data will be split into 80/20 , which will be used for model training and validation respectively.

Cleaning Data: Once the data is captured from the subjects, post processing of the data can begin. In order to remove the noise from the signals, filters are added after acquiring the data. We assumed that we are capturing muscle movements below 4 Hz. We also want to eliminate the 60 Hz interference from the surroundings. First, we applied a low-pass filter centered at 4 Hz. The filters are ideal and designed around a window sinc function [10]. After applying a low pass filter, we add a high pass filter of 0.5 Hz, which removes the aliasing and the associated DC offset. The coinciding timestamps of the EMG data with the annotations are mapped together to create an input-output relationship (Fig.2).

IV. Proposed Solution

Once the data is cleaned and transformed it will be ready for modeling. We will use the two second window samples of when the subject repeated an annotated label on the screen, and drop the instances where the word *relax* exists. In our first experiment, we will create an LSTM model. The LSTM model

Fig. 2. Signals after cleaning and adding a low pass filter. First Column: EMG1, Second Column: EMG2, Third Column: EMG3, Fourth Column: Annotated Labels

will comprise of an Sigmoid activation with a learning rate of *1e-4*. To measure the Loss, we will be using birnary cross-entropy for the binary cases of *yes* and *no*, and categorical cross- entropy for the cases *0-9*. The LSTM model will consist of single LSTM layer followed by a Dense layer.

For our second experiment, we will use the method of CNN as a Deep Learning model. We will be converting our singals into wavelet transfomrms, which will generate scalograms that show resolution in the frequency and time-domains [11]. Since there are three EMG channels of data, we create three sclaeograms per annotated label. We will then place the three scalograms on top of each other for each label, and create one image. Each image will then be utilized as features to train the CNN model. All models will be trained on a google cloud Graphics Proceesing Unit (GPU).

V. Experimental Results

VI. Discussion

VII. Conclusion

References

[1] L. Masuch, "Deep Learning - The Past, Present and Future of Artificial Intelligen. . . ." [Online]. Available: https://www.slideshare.net/LuMa921/deep-learning-the-past-present-and-future-of-artificial-intelligence?ref=https://www.analyticsindiamag.com/popular-presentations-on-artificial-intelligence-and-machine-learning/

[2] "Classify Time Series Using Wavelet Analysis and Deep Learning - MATLAB & Simulink Example." [Online]. Available: https://www.mathworks.com/help/wavelet/examples/signal-classification-with-wavelet-analysis-and-convolutional-neural-networks.html

[3] A. Eltvik, "Deep Learning for the Classification of EEG Time-Frequency Representations," p. 122.

[4] A. A. Altamirano, "EMG Pattern Prediction for Upper Limb Movements Based on Wavelet and Hilbert-Huang Transform," p. 134.

[5] E. Lopez-Larraz, O. M. Mozos, J. M. Antelis, and J. Minguez, "Syllable-based speech recognition using EMG," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, Aug. 2010, pp. 4699–4702.

[6] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* San Juan, Puerto Rico: IEEE, 2005, pp. 331–336. [Online]. Available: http://ieeexplore.ieee.org/document/1566521/

[7] M. Janke and L. Diener, "EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2375–2385, Dec. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8114359/

[8] A. Kapur, S. Kapur, and P. Maes, "AlterEgo: A Personalized Wearable Silent Speech Interface," in *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI '18*. Tokyo, Japan: ACM Press, 2018, pp. 43–53. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3172944.3172977

[9] M. Wand and T. Schultz, "Pattern learning with deep neural networks in EMG-based speech recognition," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Chicago, IL: IEEE, Aug. 2014, pp. 4200–4203. [Online]. Available: http://ieeexplore.ieee.org/document/6944550/

[10] "How to Create a Simple Low-Pass Filter - TomRoelandts.com." [Online]. Available: https://tomroelandts.com/articles/how-to-create-a-simple-low-pass-filter

[11] admin, "A guide for using the Wavelet Transform in Machine Learning," Dec. 2018. [Online]. Available: http://ataspinar.com/2018/12/21/a-guide-for-using-the-wavelet-transform-in-machine-learning/