Article Subject (Silent Speech Interfaces)

*Using Deep Learning and EMG to predict non-audible speech: Project Proposal*

Rommel T. Fernandes1\*

1 Seaver College of Science and Engineering, Loyola Marymount University, Los Angeles, CA 90045, USA

\* Student Member, IEEE

**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 to create a project plan based on Silent Speech Interfaces that use Deep Learning and Machine Learning to predict non-audible speech. We will first briefly introduce HCI systems such as Silent Speech Interfaces, and go over how Deep Learning and Machine Learning can be used to predict speech. Next, we will go over our project plan that will investigate Deep Learning and Silent Speech Interfaces and go over the research problem, which will identify the main goal and objectives. Finally, we explain our research plan which will dive into the sub goals required to complete the main objective of building a silent-speech recognition system using Electromyography (EMG) and Deep Learning.

**Index Terms—**Human-Computer Interface, Deep Learning, EMG, Silent Speech Interfaces

# [[1]](#footnote-2)Introduction

Over the past few years Human Computer Interaction has been an increasing field of study. Human Computer Interaction can be described a as feedback loop between a human and computing interface. With the increased usage of sensors worn on humans, such as watches, heart rate monitors, and other smart sensors, researchers are trying to extract the 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].

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; they have been proposed in the past. What’s 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 feature extraction, 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 rest of this paper will discuss the research problem we are trying to achieve. We will define the main goal and a set of sub-goals such that the main goal is achieved when all the sub-goals are met. The paper will then transition into the research plan which will discuss the approach that will be followed to solve the proposed research problem. We will list the necessary resources for the execution of the research plan such as the mathematics, simulation and modeling of the project. Finally, with this project proposal, we can make conclusions to specific methods one can use for furthering research possibilities in Silent Speech Interfaces.

# Research Problem

There are several situations where non-audible forms of speech translation play an important role in society. In many major hospitals you can find ER physicians dictating their speech out loud to the computer, so that they don’t have type. Using speech to text software, the Physician reads his or her findings out loud, and the computer translates them to text. This type of audible speech can disturb bystanders and may also be overheard by eavesdroppers [2]. Another situation where translation of non-audible speech plays an important role as mentioned before is speech rehabilitation. Approximately 7.5 million people in the United states have trouble using their voices according to a survey by the National Institute of deafness and other Communication Disorders [3]. A silent speech interface will allow people who have trouble communicating due to health-related issues, eventually have to possibility to do so. This paper will attempt to research the use of EMG signals to predict non-audible speech using Deep Leaning models.

We plan on using EMG to measure the activation potentials of facial muscles using surface electrodes. This will provide us with information about muscle movement during speech production [4], [5]. One of the hypotheses we want to prove is that people who speak through non-audible ways, either through choice or a health condition, move their facial muscles the same way a person who speaks with audible sounds. Researchers value EMG for its high potential in terms of non-invasiveness, cost, silent-usage and other factors [2]. The expected solution is to build a silent-speech recognition system based on EMG signals recorded through the facial muscles that would deliver results through real-time processing.

With technology limitations, it becomes a daunting task train and silent speech recognition system on a large set of words. Many researchers create a small sample of data to train on. *Lopez-Larraz et. Al* [6] trained their model on a sample of 30 syllables based on the Spanish language. In our paper we will initially plan to train on binary words such as ‘yes’ and ‘no’. If results are promising, we will plan to train on additional data such as the numerical values, 0 through 9. A research project like this becomes very challenging because it relies heavily on data collection. There are no known public data sets available for EMG speech recognition. Therefore, collecting data and training our own model will be a foreseeable task.

This overall goal will be achieved through a series of sub-goals which are identified below:

* Complete necessary paperwork for **IRB submittal**
* **Build a data capturing platform** for EMG signals using Python and Shimmer3 API
* **Use signal processing techniques** to clean the EMG signals as necessary
* **Extract necessary features** from signals that might improve a model
* **Train a variety of deep learning models** with subjects on a specific word bank using the captured data
* **Test model** with unseen data to validate performance
* **Report results**

The next section will detail the research plan based on the listed points above. Each sub-goal’s objective is to provide a solution such that the main goal will be met.

# Research Plan

In the following section we plan to detail the procedures and course of action required to complete the overall objective.

***IRB Submittal*.** The Institutional Review Board (IRB) is a standard process, which aims to protect the rights and welfare of human research subjects who are recruited to participate in research activities conducted under the institution of interest. Since our project plans to recruit up to ten volunteers to extract muscle facial EMG data, our project will need to go through an IRB review process. This involves submittal of paperwork that goes over basic questions such as: research background, procedures, risks, interactions, etc. Once this is submitted the IRB will approve or deny the application, in order to conduct research with human subjects. If the application is denied, it can be resubmitted to clarify points that the IRB may have questions about. The plan to complete the application and approval processing will be by the end of February, depending on the backlog of the IRB.

***Build a data capturing platform.*** This might be the most labor intensive portion of the project. In order to train the silent-speech recognition model, data will have to be collected from individuals wearing surface EMG devices Shimmer3) while connected to the computer. Proper instructions on how we will collect data from subjects are as followed. A good electrode to skin contact is essential for accurate surface EMG readings. The skin will be cleaned with alcohol or sanitizer to remove any inhibitory particles. Up to 12 surface electrodes (6-channels) will be placed at appropriate locations on the subject’s facial area. These electrodes are disposable and are obtained from Covidien Kendall which measures around 1” (24mm) in diameter. New electrodes will be used for every subject. The electrodes are then connected to the leads provided with the Shimmer3 EMG unit, which then is connected to the Shimmer3 unit to measure and record electrical activity associated with muscle contractions in the face. The subject will be asked to repeat words or phrases displayed on the computer silently and out loud while data is being recorded from the EMG. The process of recording data should be no longer than 5-10 minutes.

The electrodes will be placed at various locations on the facial muscles according to their distinctive movements during the pronunciation and articulation of speech [6]. There exists no standard selection of the most appropriate muscles for EMG –based speech recognition as shown in Figure 1, therefore selection of the electrodes is very heuristic in practice [6].

A GUI based application will also be built during this process using python. It will display the words the subject has to repeat while they are connected to the EMG device. This GUI based system will display a new value every two seconds. Data will be read from the Shimmer3 EMG unit at a rate between 512 -1024 Hz. This will provide enough samples for a two second period of the subject repeating the words on the screen. The format of the data will be comma separated files with header information such as: timestamp, ch1 (millivolts), ch2 (millivolts) and Target. The Target will consist of the word the subject is asked to repeat, such as: ‘Yes’, ‘No’, ‘One’, ‘Two’. These files will be accompanied by plots produced from the data.

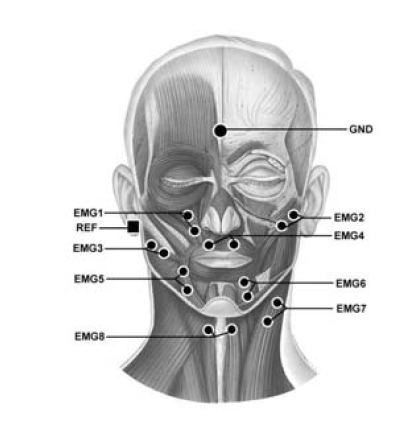


Figure 1. Selected facial muscles for EMG Recording.

***Signal Processing.*** It will be hard to imagine that the signal obtained from the user will not require any post-processing at all. One of the basic steps in signal processing of EMG will be filtering the signals. EMG signals are usually affected by noise, which may be generated by different sources, such as hardware employed for signal amplification and digitization, the movement of cables during data collection and the activity of the motor units distant from the detection point [7]. One of the main advantages of using a filter is that it is easy to implement and fast in real-time applications. The Shimmer3 unit does not include a filtering capably. A simple filter would be a notch filter at around 60Hz which would remove electrical distortion around the data collection from the subjects.

Since we are collecting data with a high sample rate, the data can potentially get distorted. Another signal processing technique might be extracting the linear envelope of an EMG signal. The Linear envelope makes it easy to interpret representation of the raw signal. It is also used to detect when a muscle is active and gives the overall level of activity in a particular muscle at any given time [8], which can be very useful for facial muscle EMG detection.

***Feature Extraction.***One of the advantages of Deep Learning as mentioned previously is the ability to self-extract valuable features so that a user does not have to manually extract features from the signal. Manually extracted signal include Peak-to-Peak values, RMS values of signals, Power Spectral Densities, aggregate values such as min and max, etc. Even though deep learning does not require you to do any feature extraction, adding these values into the model can help boost performance. One example of feature extraction and feature transformation, would be taking the principal components of a signal and using those components as additional features in the model. One of the main purposes for feature extraction is to reduce the dimensionality of the dataset. Principal Component Analysis reduces the dimensionality of the signal by extracting its core components and avoids any redundant information.

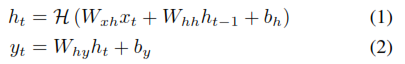
***Training Models.*** Of the research papers investigated, papers [2], [4], [5] have attempted some kind of speech recognition using EMG and Deep Learning. The data collection methods in all three of them are different. What is consistent is the type of algorithms they used to train the data. In [2], they try a variety of models from Gaussian Mixture (GMM) Models, Long-Short Term Memory (LSTM) model, and a Deep Neural Network models. Similarly, paper [4] uses a comparable Deep Neural Network Model to that of [2]. The most recent conference paper [5] uses a type of Deep Neural Network called a Convolutional Neural Network (CNN). In our paper, we plan to investigate the combination of a bidirectional LSTM RNN model with end-to-end training which has given promising results in other speech recognition models [9]. A given RNN can be described as a given input sequence

***x*** *= (x1, …., xT),* which then computes a hidden vector

***h*** *= (h1, …., hT)* and the output of the sequence

***y*** *= (y1, …., yT)* by iterating the following equations from

*t* = 1 to *T,* such that:



where ***W*** denotes the weight matrices and ***b*** terms denote bias vector and ***H*** is the hidden layer of the function. LSTM is then used to with conjunction of the RNN model to address the vanishing gradient problem for the RNN [10]. The vanishing gradient problem limits the RNN from improving after a certain epochs. Therefore, LSTM adds additional non-linearity, so that the RNN has the possibility of improving over time.

***Testing Models.*** In order to test the validly of the model we will be using and 80/20 split of training and testing data. In our case, since we plan to have ten subjects, with two of those subjects, we will use their data to validate the model by feeding the results validating the accuracy. Since this is a classification Deep Learning model, we will be using a classification accuracy score as well as confusion matrix which will help us determine *True Positives* (model correctly predicts the positive class), *True Negatives* (model correctly predicts the negative class), *False Positive* (model incorrectly predicts the positive class) , and *False Negatives* (model incorrectly predicts the negative class). By analyzing the accuracy scores, we can then make a determination of how well our EMG Silent Speech Interface would with live data.

**Results.** If time permits, after reporting our results and completing our analysis, we would like to deploy the model in a production environment. With the model in place, and the electrodes connected to the user, data can be transmitted from the EMG unit via Bluetooth to the computer as it makes real-time predictions as our proof of concept.

# Conclusion

Deep Learning is playing an important role in silent speech interfaces. In this paper we provided a project plan for using Deep Learning and EMG to predict non-audible speech. We reviewed research and found that Deep Learning methods can potentially outperform traditional machine learning methods when classifying words in silent speech interfaces. Finally, we reviewed the necessary tasks that needed to be accomplished in order to meet our objective of creating an EMG silent speech interface system.

References

[1] J .Wang et al. ,Deep learning for sensor-based activity recognition: A Survey, Pattern Recognition Letters (2018)

[2] M. Janke and L. Diener, “EMG-to-speech: Direct generation of speech from facial electromyographic signals,” IEEE/ACM Trans. Audio,,Speech, Lang. Process., vol. 25, no. 12, Dec. 2017.

[3] National Institute on Deafness and Other Communication Disorders (NIDCD), “Statistics on voice, speech, and language,”2016.[Online].Available: <http://www.nidcd.nih.gov/health/statistics/pages/vsl.aspx>

[4] M. Wand T. Schultz "Pattern Learning with Deep Neural Networks in EMG-based Speech Recognition" vol. 35 pp. 4200-4203 2014.

[5] A. Kapur, S. Kapur, and P. Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In 23rd International Conference on Intelligent User Interfaces (IUI ’18). ACM, New York, NY,USA, 43–53. <https://doi.org/10.1145/3172944.3172977>

[6] E. Lopez-Larraz, O. M. Mozos, J. M. Antelis, J. Minguez, "Syllable-based speech recognition using EMG", Proc. Ann. Int. Conf. Eng. Med. Biol. Soc., pp. 4699-4702, 2010.

[7] J. Pauk (2008). Different techniques for EMG signal processing. Journal of Vibroengineering. 10. 571-576.

[8] Shimmer EMG User guide, Rev 1.12. Realtime Technologies., 2017.

[9] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 6645–6649. IEEE, 2013

[10] A. Karpathy, J. Johnson, and F.-F. Li. Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078, 2015

1. [↑](#footnote-ref-2)