Non-Audible Speech Classification Using Deep Learning Approaches

Rommel Fernandes, Lei Huang, Gustavo Vejarano
Department of Electrical Engineering and Computer Science
Loyola Marymount University
Los Angeles, CA, USA
rferna16@lion.lmu.edu, lei.huang@lmu.edu, gustavo.vejarano@lmu.edu

***Abstract*—Research advancement of human-computer interaction (HCI) has recently been made to help post-stroke victims dealing with physiological problems such as speech impediments due to aphasia. This paper investigates different deep learning approaches used for non-audible speech recognition using electromyography (EMG) signals with a novel approach employing continuous wavelet transforms (CWT) and convolutional neural networks (CNNs). To compare its performance with other popular deep learning approaches, we collected facial surface EMG bio-signals from subjects with binary and multi-class labels, trained and tested four models, including a long-short term memory(LSTM) model, a bi-directional LSTM model, a 1-D CNN model, and our proposed CWT-CNN model. Experimental results show that our proposed approach performs better than the LSTM models, but is less efficient than the 1-D CNN model on our collected data set. In comparison with previous research, we gained insights on how to improve the performance of the model for binary and multi-class silent speech recognition.**

***Index Terms*—deep learning, electromyography, silent speech interfaces, human-computer interaction, wavelet transform**

I. INTRODUCTION

Over the past few years, HCI has been an increasing field of study. HCI can be described as a feedback loop between human and computer. With the increased usage of wearable devices, such as watches, heart rate monitors, and other smart sensors, researchers extract bio-signal information to recognize typical human activities. For example, electrocardiogram (ECG) signals have been used to detect irregular heartbeats [1], while electroencephalography (EEG) [2] and electromyography (EMG) [3] signals have been used to predict body movements.

Silent speech interface(SSI) is one type of HCI. It employs a signal-extracting system such as EMG and EEG to collect signals of silent or non-audible speech, then recognize different attributes from the signals using machine learning algorithms. SSI systems can help post-stroke victims dealing with physiological problems such as speech impediments due to aphasia. In the past, machine learning algorithms typically used for speech recognition, such as decision tree, support vector machine, Naïve Bayes, and Hidden Markov Models, were also employed as classifiers for silent speech signals. These traditional machine learning algorithms require extensive feature extraction from signals, and only shallow features can be learned from those approaches, leading to undermined performance. With recent advancement in deep learning and it's breakthrough performance applied to speech recognition, state-of-the-art SSI systems have employed deep learning technologies to classify silent speech signals [4].

This paper focuses on recognizing EMG captured non-audible speech, which is caused by the inability to verbalize words or sentences through the use of sound in an effective way. Compared with other methods that capture non-audible speech such as electroencephalography (EEG), near infrared sensors (fNIRS), implants for speech and motor cortex (ECOG), and video camera lip-reading, EMG is non-invasive and most cost-effective. Therefore, we employ EMG to capture non-audible speech in our project. We then investigate different deep learning approaches, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), as models for recognizing the captured EMG signals. In general, RNN models such as LSTM are prevalent in modeling time-sequence signals including speech, while CNN models are used for multidimensional signals such as images and videos. Similar to speech signals, EMG signals are time sequences, which are suitable for RNN modeling. On the other hand, EMG signals are always collected through multiple sensor arrays placed at different locations, so they can also be arranged as multidimensional tensors, which are suitable for CNN modeling. In this work, we compare RNN and CNN models for classifying EMG signals. In addition, we propose a novel approach to applying CNN on the scalograms of EMG signals through a continuous wavelet transform.

The rest of this paper is organized as follows. Section II discusses several recent and related works. Section III describes our experimental setup used to capture and preprocess EMG data samples as well as four different deep learning approaches we trained and tested. Section IV presents and discusses the experimental results. Finally, Section V concludes this paper and provides recommendations for future work.

II. RELATED WORK

Research conducted using EMG to predict speech for SSI systems has taken place for over two decades. Before the deep learning era, using EMG to recognize speech patterns involved heavy feature extraction of the data along with discrete

mathematical modeling. Recently, there have been well-documented results of using a combination of mathematical modeling and deep learning to predict speech using EMG. Some research addresses syllable and single-word prediction [5], [6]. Other research has used EMG signal 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 words from 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.

In [9], new approaches with machine learning models such as Restricted Boltzmann Machine algorithms and deep neural networks (DNN) were introduced. Their corpus consisted of 25 sessions from 20 speakers comprising of 200 read English-language utterances such as phonemes, consonants, and vowels. Their results showed that DNN models performed better for phoneme related classifications using EMG inputs.

The work performed by [7] continued some of the primary research done in [9]. The authors investigated LSTM models and compared them to other models such as GMMs. Their results showed that LSTM models performed better than GMM.

The work done in [10] used the same data set and corpus of [7]. Their primary work focused on evaluating the performance of CNN based models for EMG-to-Speech conversion. The researchers converted EMG signals to mel-frequency cepstral coefficients (MFCCs) and extracted feature vectors from multiple channels, and then used two CNN based models, a LeNet-inspired model and a ResNet-32 based encoder-decoder model. Their results showed that the CNN based architectures can outperform a plain DNN based conversion system.

In [8], the authors built a proof of concept SSI system that used a one-dimensional CNN as a classifier. Seven electrodes were placed around the throat and face. Subjects in the research did not open their mouth, make any sound, or provide any muscle articulation to train the models. Their quantitative results for the 1-D CNN network reported an average accuracy of 92.01% for all subjects. Their corpus included individual words and short phrases.

From the above-mentioned research work, it has been shown that deep learning models have improved the performance of EMG signal modeling over traditional models, and CNN models outperform plain DNN models. However, there has not been any comparison between RNN and CNN in recognizing EMG silent speech signals. In addition, some previous research applied models on extracted feature vectors of EMG signals, while others on the original data collected. Based on these observations, we compare the performance of several RNN and CNN models, and propose a different data representation method using wavelet transformed signals, which provides time-frequency information of the signals.

III. SYSTEM DESCRIPTION

The system we used in our experiments consists of three main subsystems: data acquisition, data preprocessing, and model training and testing. In the data acquisition, the original multi-channel EMG data are collected from 10 different human subjects who volunteered to participate in the experiment. The collected raw data are then cleaned, aligned and labeled with corresponding labels in the data preprocessing subsystem. Finally, the formatted and divided data sets are used to train and test different deep learning approaches that recognize speech from the EMG signals. In the following subsections, we describe each subsystem in detail.

A. Data Acquisition

Since our main focus is on comparing different deep learning approaches, we simplified our data acquisition process by using three channels positioned on the cheek and throat area as suggested in [6]. Compared to a far greater number of channels used in previous works in [8], [9], [7], [6], this minimum number of electrodes reduced discomfort experienced by the subjects.

The data acquisition device consists of two Shimmer3 EMG units, each with a 24 MHz CPU. The EMG units have 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 the work done in [5]. The areas where the EMG electrodes were placed are as follows: *Depressor anguli oris* (EMG1), *Zygomaticus major* (EMG2), and *Anterior belly of the digastric* (EMG3), as shown in Fig. 1a. Each bipolar electrode of the muscle group is placed approximately 2 cm apart based on the unit specifications in Fig. 1a. After proper placement of the electrodes on the subject, the EMG units are placed on the upper torso and shoulder using comfort straps. Data including the EMG signal strength and timestamps is transmitted via Bluetooth to a Linux (Ubuntu) Intel laptop using open-source Python libraries.

We captured two types of labeled annotations for our sample data. Our *first set* of annotations consists of the labels for the words *yes* and *no*. Our *second set* of annotations consists 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 without making sound Fig. 1b. The label persists on the screen for two seconds; it is then followed by the word *relax* (as shown in Fig. 1b), which persists on the screen for two more seconds. The next label in the annotations is randomly generated and displayed on the screen for a total of 50 labels per annotated set. The subject performs this task for the *first set* and *second set*.

B. Data Preprocessing

Once the data is captured from the subjects, it has to be cleaned and aligned before being an input into a deep learning model. To remove noises from the recorded signals, low-pass and high-pass filters are applied.

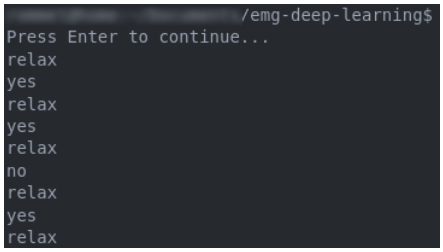
First, we applied a low-pass filter with a cutoff frequency of 4 Hz, assuming muscle movements are mostly below 4 Hz. We then applied a high-pass filter with a cutoff frequency of 0.5 Hz, which removes the DC offset. The filters are designed around a window sinc function. The coinciding timestamps of the EMG data with the annotations are mapped together to create an input-output relationship. Fig. 2 shows samples of the preprocessed EMG signals from three channels with the respective annotated labels. Each row in Fig. 2 represents an annotated event, and the first three columns represent the EMG channels. With channel 1 representing the first column, channel 2 representing the second, and channel 3 representing the third column. Each sample was segmented with a two-second fixed window size as the subject silently reads the annotated label on the screen. The signal that is captured during the word *relax* is then discarded from the data set.

C. Model Training and Testing

Four different deep learning approaches are trained and tested using the same EMG data sets we collected and preprocessed as described in previous subsections.

The first model is a single directional LSTM model that is similar to the bi-directional model in [7]. It consists of an LSTM layer with a hidden dimension of 100 units followed by a dense layer of 80 units, and an output layer. It uses a batch size of 32. Binary cross-entropy is used as the loss function for the binary classification of *yes* and *no*.

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



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

Fig. 1. Subject reading annotated labels on screen while connected to EMG units and electrodes.

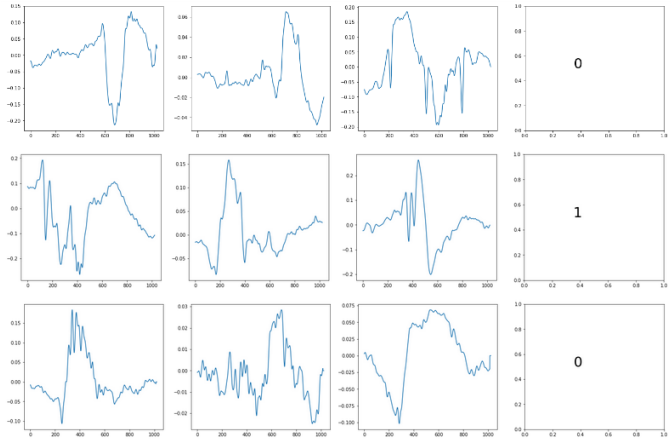


Fig. 2. Signals after cleaning and adding low-pass and high-pass filters. First Column: EMG1, Second Column: EMG2, Third Column: EMG3, Fourth Column: Annotated Labels

The second model is a bi-directional LSTM model as proposed in [7], which uses two hidden layers of 100 and 80 units with a sigmoid activation.

The third model is a 1-D CNN model as proposed in [8]. It consists of two convolutional layers with 400 units with max pooling, followed by a fully connected layer of 200 units with a sigmoid activation. Each convolutional layer used a dropout of 0.25.

The fourth model is a two-dimensional (2-D) CNN model, which requires the inputs to be two-dimensional signals. One way of formatting the EMG signals as a 2-D array is to stack up one-dimensional signals from multiple channels. However, since we only used three channels, there is not enough information on the channel dimension to be explored by CNN units. Inspired by some previous work that extracted frequency domain features for 1-dimensional signal recognition, we propose to add frequency as the second dimension. We first generated a scalogram, which is a 2-D time-frequency representation of the original signal, for each channel using a continuous wavelet transform. For each sample, we stack up the three channels' scalogram in a similar fashion to that of the three color channels in an image. Fig. 3 shows EMG signals with their scalogram. We choose a frequency bin size of 256

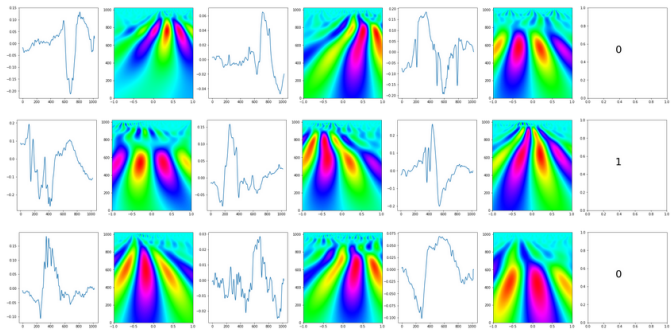


Fig. 3. Sample signals with corresponding scalograms and labels

TABLE I
TESTING PERFORMANCES OF DIFFERENT MODELS

Class Label	LSTM			BI-LSTM			1-D CNN			2-D CNN		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
no	0.60	0.83	0.70	0.49	0.62	0.55	0.85	0.79	0.82	0.75	0.72	0.73
yes	0.77	0.52	0.62	0.56	0.42	0.48	0.83	0.88	0.85	0.76	0.79	0.77
Average	0.69	0.66	0.66	0.53	0.52	0.51	0.84	0.84	0.84	0.76	0.76	0.76

TABLE II
1-D CNN MODEL RESULTS

Class Label	Trained Data		Test Data	
	Prec.	Rec.	Prec.	Rec.
no	0.96	0.87	0.85	0.79
yes	0.87	0.96	0.83	0.88
Average	0.92	0.91	0.84	0.84

TABLE III
2-D CNN MODEL RESULTS

Class Label	Training Data		Testing Data	
	Prec.	Rec.	Prec.	Rec.
no	0.97	0.95	0.75	0.72
yes	0.95	0.97	0.76	0.79
Average	0.96	0.96	0.76	0.76

and a Mexican hat wavelet function to generate the scalogram. The coefficients from the output of the scaleogram are then used together with the corresponding label to train and test the CNN model. The CNN model has a similar architecture to that of LeNet, which can be represented as conv1-pool1-conv2-pool2-flat-FC1-FC2, with a 0.5 dropout before the FC layers. The two convolutional layers have 32 and 64 units respectively, each with a filter size of 3x3, and a max-pooling kernal size of 2x2, with a stride of 2. This is then followed by the two dense layers, with FC1 having 512 units and FC2 having two units as outputs. A smaller batch size of 16 is used due to the large data size for each sample.

The preprocessed data sets are divided into a training set and testing set using an 80/20 split. All models are trained using an Adam optimizer with a learning rate of 1e-4. Training of the models are conducted on a Google Cloud Computing Engine with four virtual CPUs, 26 GB of memory, and one NVIDIA Tesla K80 Graphics Processing Unit (GPU).

IV. EXPERIMENTAL RESULTS

For the binary classification experiment, we evaluate both precision and recall values, as well as the F1 score for each of the four models. The testing results are documented in Table I. Precision (denoted as Prec. in Table I) measures the fraction of correctly identified relevant samples over all identified relevant samples, while recall (denoted as Rec. in Table I) measures the model's capability of identifying relevant samples over all relevant samples in the data set. F1 score is a balanced accuracy measurement calculated as the harmonic mean of precision and recall values. The four models described in Section III-C is denoted as LSTM, Bi-LSTM, 1-D CNN, and 2-D CNN, respectively.

It can be observed from Table I that the 1-D CNN performs the best, while the 2-D CNN model achieves better performance than the LSTM models. The larger number of units in each convolutional layer used in the 1-D CNN model than in the 2-D CNN model helps to identify more critical features in the local data sequence. Although the time-frequency representation by the scalograms provides an additional dimension of information in the frequency domain, the added amount of data limits the number of units in each convolutional layer. In addition, the continuous-time wavelet transformation of data requires more processing power. Therefore, the simpler 1-D CNN model is more computationally efficient than the 2-D model.

Comparing the CNN models to the RNN models, our results showed that both CNN models outperform the RNN models in all measurements. One possible reason is due to the simple words *yes* and *no* we are predicting may not have significant long-term dependencies within the time sequences. Therefore, CNN models with small filter size can focus more on looking for local patterns. LSTM models should be more suitable for other EMG-SSI applications that aim at predicting longer phrases or sentences.

Comparing the two RNN models, the more complicated bi-directional LSTM model performs worse than the simpler single-directional LSTM model. Similarly, comparing the two CNN models, the more complicated 2-D CNN model performs worse than the simpler 1-D CNN model. One major contributing factor is the small size of the data set used in our experiment. As a result, a larger model suffers more severely from overfitting, which can be shown by better training than testing performances. From Tables II and III, a higher discrepancy between the training and testing performances can be observed in the results of the 2-D CNN model (shown in Table III) than those of the 1-D CNN model (shown in Table II).

It should be noted that in [7], [8], [10], the authors used similar deep learning approaches to the ones we used, and reported higher accuracy than our experimental results presented in this paper. One of the main contributing factors is the much larger amount of data they had available for training and testing purposes. In [8], approximately 31 hours of training data was captured in order to train their one-dimensional CNN model. In [7], [10], the authors trained a variety of different models, utilized the on-going corpus of data from previous research projects that focused on EMG-SSI systems. We have shown for a small data corpus, CNN based models perform better than LSTM models. With additional data, we believe that the performance measurements in our

experiments can be improved significantly.

V. CONCLUSION AND FUTURE WORK

We have experimented with an SSI system that recognizes non-audible speech from surface EMG signals using different deep learning approaches. This type of system can help people who suffer from speech-related problems. We compared four different deep learning approaches, including two RNN models (LSTM and bi-directional LSTM) and two CNN models (1-D and 2-D CNN models). In order to apply the 2-D CNN model, we proposed to transform the original time sequence signals into their scalograms using continuous wavelet transform. Based on limited data we collected from a small number of subjects using simplified data acquisition device and process, the experimental results showed that CNN models perform better than LSTM models, and simpler models such as single directional LSTM and 1-D CNN models outperform their more complicated counterparts such as bi-directional LSTM and 2-D CNN models, respectively.

There are many ways to improve the performance of the SSI system. In data acquisition, it is imperative to acquire more annotated EMG data with proper labels in order to overcome overfitting of powerful deep learning models. However, recruiting more subjects to volunteer can be challenging. One alternative approach is to train a customized SSI system, that is tailored to specific subjects. We have also collected data for more words and phrases such as the ten digits. Appropriate deep learning models are to be created and trained in order to recognize more meanings of silent speech. In data preprocessing, we eliminated the captured signal when the subject was told to *rest*. Future models could incorporate the *rest* instances as an additional class so that the model will learn to identify false positives. In model training and testing, hyper-parameters should be optimized for each model.

REFERENCES

[1] "Classify Time Series Using Wavelet Analysis and Deep Learning - MATLAB & Simulink Example," .

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

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

[4] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu, "Deep Learning for Sensor-based Activity Recognition: A Survey," July 2017.

[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, 2005, pp. 331–336, IEEE.

[7] Matthias Janke and Lorenz 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.

[8] Arnav Kapur, Shreyas Kapur, and Pattie Maes, "AlterEgo: A Personalized Wearable Silent Speech Interface," in *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI '18*, Tokyo, Japan, 2018, pp. 43–53, ACM Press.

[9] Michael Wand and Tanja 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, Aug. 2014, pp. 4200–4203, IEEE.

[10] Lorenz Diener, Gerrit Felsch, Miguel Angrick, and Tanja Schultz, "Session-Independent Array-Based EMG-to-Speech Conversion using Convolu- tional Neural Networks," *13th ITG Conference on Speech Communication*, p. 5, 2018.