Proceedings of 12th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2022 Month: October Volume: 12

Silent Speech Recognition in Nepali

Rhimesh Lwagun[a], Sanjay Rijal[b], Rabin Nepal[c], Upendra Subedi[d], Dinesh Baniya Kshatri[e]

Thapathali Campus, IOE, Tribhuvan University, Nepal
[a] rimeshlwagun10@gmail.com, [b] sanjay.732419@thc.tu.edu.np, [c] rabin47nepal@gmail.com,
[d] upendrasubedi2@gmail.com, [e] dinesh@ioe.edu.np

Abstract

Since the development of the very first computer, human-computer interaction has always required to have some form of physical activity as an input to the computer. Although these methods are accurate and hassle-free, they fail to be intelligible to the differently-abled. Speech interaction tackles this issue to some extent but it is still subjected to privacy issues. The proposed system in this research project confronts these problems and provides a secure and seamless interaction between a human and a computer using silent speech recognition. The surface electromyography (sEMG) signals, during the silent speech, are recorded from the facial muscles of a speaker using 8-channel gold cup electrodes and filtered to remove noise and other unwanted signals. The spectrogram of the processed signal is then extracted to train a Convolution Neural Network (CNN). The trained model is finally deployed to predict the utterances.

Keywords

Silent speech, Surface Electromyography, Spectrogram

1. Introduction

Speech being a convenient and natural way of conveying messages has been an integral part of human information exchange. The advent of modern electronics has made it easier to interact with these devices through natural speech. However, human speech is a complex process involving muscle movements with myriad nerve impulses fired from neuromuscular junctions in a pattern; generated as a result of muscle fiber action potential in the cell membranes triggering the neurotransmitters (acetylcholine) [1]. These biosignals produced during the speech process can be utilized to predict the utterances which is the underlying theory of human-computer interaction through silent speech.

Human speech can be exemplified using a three-stage model; Conceptualization, Formulation, and Articulation. The first two stages occur in the brain, whereas the last stage occurs in the motor system under the control of the brain. Articulation again can be further subdivided into three stages; Respiration, Phonation, and Articulation. Respiration, also known as breathing, refers to the inhalation and exhalation of air by the contraction and expansion of the diaphragm. During the inhalation, the lungs expand, causing the air to flow from the mouth to the lungs with the glottis relatively open. During exhalation, the lungs contract, pushing the air from the lungs toward the mouth, which provides energy for human speech. For most languages, the production of sound occurs during exhalation which is why humans cannot generally speak while inhaling. Phonation is the process that modifies the pulmonic air in such a way that it produces acoustic signals. During phonation, the vocal folds vibrate causing a change in air pressure generating acoustic waves which then get amplified internally through resonance. The rate of vibration of the vocal folds is perceived by the listener as modifications in pitch and/or loudness. On the other hand, articulation implies all the actions of the organs of the vocal tract that affect modifications of the signal generated by the voice source resulting in speech events that can be identified as vowels, consonants, or other phonological units [2]. This mode of speaking naturally occurs while reading which helps the human brain to comprehend what is being read and potentially reduces the cognitive load on the brain. It should be clearly established that this internal articulation process is voluntary and occurs in the articulatory system upon receiving the stimulus from the brain. Internal articulation occurs due to the cumulative action of the brain and the peripheral nervous system but is very different from the speech conception happening in the brain [2].

Silent speech occurs without respiration and phonation and is the main activity upon which this research is based [3]. The silent electromyography (sEMG) signals from articulatory muscles can be extracted and interpolated into an abstraction of speech. This research project emphasizes this speech method and employs it in achieving seamless human-computer communication measures.

2. Literature Review

There have been many attempts at Speech recognition using EMGs. [4] presented a wearable silent speech interface, AlterEgo which basically is a natural extension of the user's cognition by enabling a seamless conversation with machines and people allowing users to send arbitrary text input to a computing device using natural language processing, without discernible muscle movements or any voice commands. [5] describes the method of speech recognition by surface EMG signals by recording the electric active potentials of articulatory muscles followed by decoding into a speech that the speaker vocalized. Speech recognition using EMG signals dates back to the 1980s. 93% accuracy was observed on 10-word vocabulary. It suggested a satisfactory result can be obtained even for the silently articulated signals.

Similarly, [6] attempted to perform session-independent subvocal speech recognition leveraging character-level recurrent neural networks and the connectionist temporal classification loss. Recently [7] carried out human-computer interaction using sEMG signals (vocalized with ten phonetically distinct words) in two different articulation modes; with subtle facial muscle movement and internally articulated speech, and achieved a sound CNN performance leveraging temporal as well as spectral feature vectors. They further deployed the model predicting the utterances a significant number of times.

3. Dataset Preparation

3.1 Muscle selection

During internal articulation (without acoustic vocalization), no significant movement of facial muscles or articulators is observed, however, more than 15 muscles directly involved in the speech system are neurologically activated. These particular muscles receive feeble electrical signals from the peripheral nervous system (PNS) [8]. Depending on the extent of involvement, various facial muscles were selected for signal extraction. Many factors such as noise susceptibility, signal strength, cross-talk, signal frequency, nature of electrode, electrode-muscle distance, electrode configuration, and convenience for placement are considered during muscle selection. Some of the basic criteria to be followed while selecting the muscles are as follows:

- Select the muscle depending on the convenience of electrode placement.
- Distance from the electrode to the muscle should be minimum.
- Size of the muscle should be large enough to avoid cross-talks.
- Size and type of electrodes should be considered before selecting the muscles.
- Configuration of electrodes (bipolar or monopolar) determines which muscle to select.

Table 1: Selected Facial Muscles for sEMG Signal Extraction Involved in Activation of Active Articulators

EMG Channel Number	Name of the Muscle
1	Levator Angulis Oris
2	Zygomaticus Minor
3	Zygomaticus Major
4	Orbicularis Oris
5	Omohyoid
6	Anterior Belly of Digastric
7	Mylohyoid
8	Depressor Anguli Oris

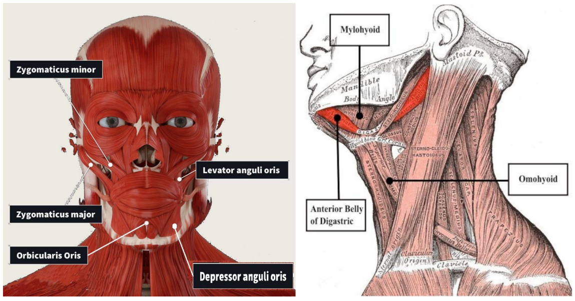

Figure 1: Selected Active Facial Muscles

3.2 Dataset Preparation Method

Datasets were collected using gold-cup electrodes (1.45 mm diameter of conductive area) [9] and the Cyton board of OpenBCI [10] using all the available 8 channels in monopolar configuration at a sampling rate of 250 Hz. Eight signal electrodes were placed at different selected facial muscles. The ground and the reference electrodes were attached to the ear lobes. Before placing the electrodes on the target muscles, the area was cleaned with isopropyl alcohol, and a generous amount of electrolyte (Ten20 conductive paste) was applied to the electrodes facilitating both conduction and adhesion. The electrodes were further secured to the target muscles using medical tape.

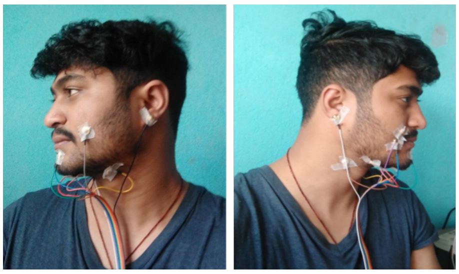

Figure 2: OpenBCI Electrodes Placement

Table 2: Dataset Description

Speaker	Session Count	File Count	Session Length (sec)
RL	8	86	9480
US	9	79	10320
SR	1	12	480

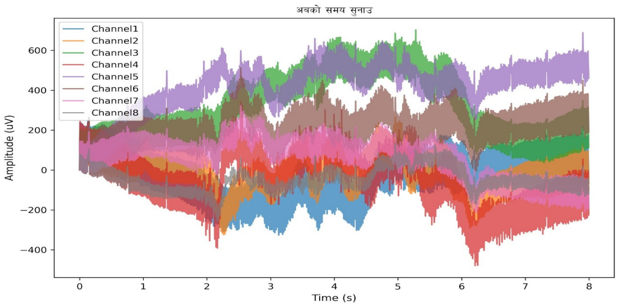

Figure 3: Raw sEMG Signal Obtained Directly from Facial Muscles Through Electrodes

A total of 5 phonetically different Nepali sentences were selected for the preparation of the dataset: “बत्तिको अवस्था बदल”, “आजको मौसम बताउ”, “एउटा सङ्गीत बजाउ”, “पङ्खाको स्थिती बदल” and “अबको समय सुनाउ” which translates to “Turn on/off the light”, “What is today’s weather?”, “Play a song”, “Turn on/off the fan” and “What time is it now?” respectively. The dataset as shown in **Table 2** contains raw EMG data from the selected facial muscles of 3 different male subjects of age range 22 to 24 years old. The sentences were chosen such that they are phonetically different and of approximately equal duration. The extracted signals from all the channels were saved in a *.txt* file using OpenBCI GUI [11]. The dataset from two of the users (US and RL) was used to train the model while that of the user SR was used as a test dataset.

The raw sEMG signal for the word “अबको समय सुनाउ” directly extracted from active facial muscles using gold-cup electrodes and OpenBCI hardware is shown in **Figure 3**.

4. Silent Speech Recognition Model

4.1 Preprocessing

After, the digitization of the signal; still consisting of various noises and artifacts, requires preprocessing before feature extraction. Since the prominent internally articulated signals fall within the range of 1.5-50 Hz, inevitable noises such as line noise and ECG artifacts need to be removed. The line noise from the surrounding electrical power line, encountered at 50 Hz along with its harmonics is suppressed by cascading three notch filters of order 1. Afterward, the ECG artifacts, due to the electrical activity from the heart, are minimized using Ricker (Mexican Hat) Wavelet, a pulse-shape signal generally used in determining the pulse period and mimicking the impulsive portion of the pulse [12].

$$R(f) = (1 - 2\pi^2 f^2 t^2) e^{-\pi^2 f^2 t^2} \tag{1}$$

The frequency spectrum of this wavelet is real and non-negative.

$$|R(f)| = R(f) \tag{2}$$

Also, the differentiation of the function in the frequency domain is zero i.e.:

$$\frac{dR(f)}{df} = 0 \tag{3}$$

Thus, any delay in the time domain only affects the magnitude of the function while the phase remains constant which makes the detection of the impulses convenient [13]. Since the wavelet mimics the

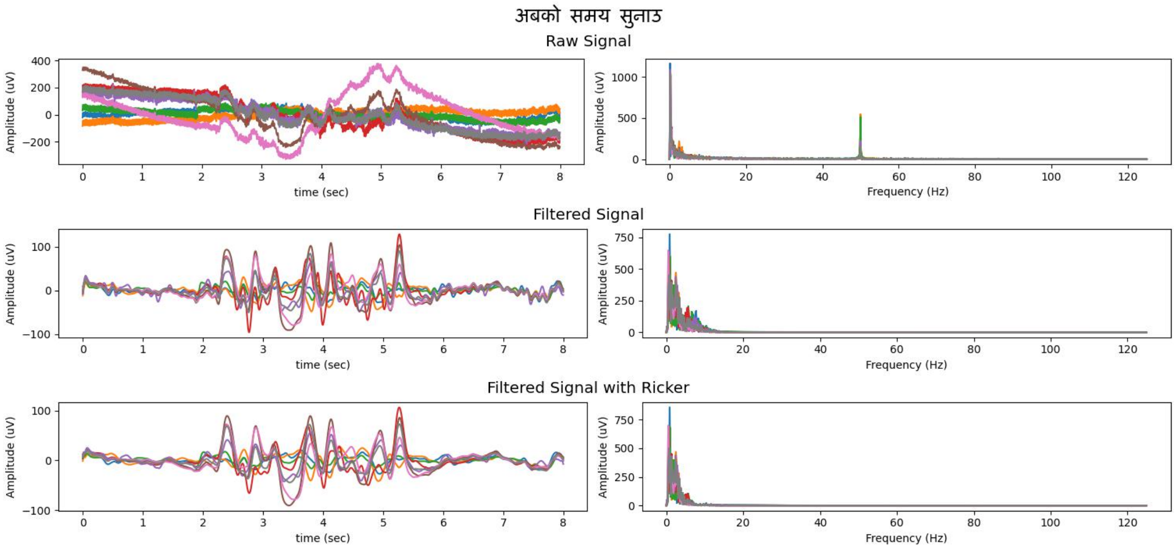

Figure 4: Preprocessing of Raw sEMG Signals with Digital Filters and Ricker Wavelet

heartbeat pulse, it is convoluted with the EMG signal to extract the ECG signal which is then removed from the original EMG signal without losing much of its features. And to focus only on the signal of interest, it is passed through a 1st-order Butterworth band-pass filter with a lower cutoff frequency of 1.5 Hz and a higher cutoff frequency of 50 Hz. This is done to ensure the removal of other noises in the signal. The gradual improvement in the signal is depicted in the **Figure 4** as the aforementioned preprocessing steps are applied to the signal. The subplot shows the EMG signal of interest after attenuating most of the ECG artifacts and line noises. The EMG signals produced during internal articulation are faint as they are in microvolts (within $\pm 200\mu$V) with low frequency (<20 Hz).

Any machine learning performs better when there's an ample amount of data. The amount of data collected is sufficient but augmenting the data gives it a fair advantage to learn more from the features. The raw data is augmented by the time shifting of the signal. Mainly, the training dataset is augmented with a random shifting factor (t_0); the right shift causes time delay while the left shift results in time advancement. Thus, the randomization in the shifting factor and shifting direction produces additional unique data to train the model.

4.2 Feature Selection

The obtained filtered signal shows the presence of an EMG signal, but still, it does not provide information to both the human and the machine. Since the EMG signal is very different from the speech signal, it is necessary to explore feature extraction methods suitable for EMG. After an exhaustive study of various time domain and frequency domain features, it was found that time domain features produce more desirable results than any other features. However, one has to combine multiple time domain features, which is computationally heavy as a single utterance contains 8 channels. Thus, Spectrogram, a feature that gives information on time and frequency was chosen. Spectrogram combined the information of spectral and time-domain (with trade-offs) with some computation.

One of the most prevailing spectral features for signal analysis is Short Time Fourier Transform (STFT), a frequency domain feature representation. STFT is obtained by introducing a sliding window to the time-variant signal, so it is also known as the time-dependent Fourier transform. This window adds a new temporal dimension to the frequency response by suppressing the input signal outside a certain region (window). Moreover, it reduces spectral leakage, a phenomenon of smearing of the spectrum due to leakage of energy from the frequency components nearby [14]. If an input signal $x(t)$ is introduced to the sliding window with window function $\gamma(t)$ then the discrete-time STFT of the signal is given by,

$$X(\tau,\omega) = \int_{-\infty}^{+\infty} x(t)\gamma(t-\tau)e^{-j\omega n}dt \tag{4}$$

Here $\gamma(\tau)$ is the window interval centered at zero.

In general, STFT is complex-valued so spectrograms are often used for further processing of the signal as they represent the pattern of energy change of the signal which may not be visible with STFT or PSD. The spectrogram is simply the rectified square of STFT given by,

$$S(\tau,\omega) = \left| \int_{-\infty}^{+\infty} x(t)\gamma(t-\tau)e^{-j\omega n}dt \right|^2 \tag{5}$$

The spectrogram parameters were set as per **Table 3**. The overlap percentage was specified such that it complies with the Constant Overlap Add (COLA) constraints [15]. The spectrogram features were normalized channel-wise to emphasize the regions with a higher probability of occurrence of the signal. According to the Heisenberg uncertainty principle, the time-frequency windows cannot be made arbitrarily small, and a perfect time-frequency resolution cannot be achieved [16]. Therefore, the selection of window size should be done considering whether the time resolution or the frequency resolution is required.

$$\triangle t.\triangle\omega \geq \frac{1}{2} \tag{6}$$

where $\triangle t$ is the radius of the time window $\gamma(t)$ centered at $\tau(0)$ while $\triangle\omega$ is the frequency window $\Gamma(\omega)$ centered at ω_o.

Table 3: Spectrogram Parameters for Feature Extraction

Spectrogram Parameters	Values
Sampling Rate	250 Hz
Window	Periodic Hanning Window (M=60)
Per Segment Length	60
Overlap	75%
FFT Length	60

5. Neural Network

5.1 Model Architecture

Various machine learning models were employed, trained, and tested on the spectrogram dataset of the extracted sEMG signals. Based on the performance of the model; train accuracy and loss, validation accuracy and loss, the trend of train and validation outputs, confusion matrices, and precision-recall scores, the optimal model was selected which in this specific case is Convolution Neural Network (CNN).

The architecture of the employed neural network is as shown in the **Figure 6**. CNN model detects the patterns, and power range of signals detected in a specific frequency range for a specific utterance, in the feature vectors and convolute using defined kernels for different layers.

The input to the CNN was 4-dimensional spectrogram feature tensors. Two different 2D convolution layers; Conv2D1 (output filters = 64, kernel size = (3,3), activation function = ReLU) and Conv2D2 (output filters = 512, kernel size = (3,3), activation function = ReLU) were the initial feature mapping layers of the model, both followed by batch normalization (momentum = 0.99, epsilon = 0.001) as per necessity, minimizing the training time and test error for the minority class in highly imbalanced dataset [17] and max pooling layers of stride size (2,2).

After flattening the feature space by flattening the layer, four different dense layers each with ReLU activation function, and 20-25% dropout to each dense layer. Moreover, kernel and activity regularization were added to the layers depending on the regularization requirements. After a series of experiments, kernel, and activity regularization were set to 0.001 while the bias regularization was found to have no significant impact on the model response. For the output, another dense layer with a softmax activation function was applied.

5.2 Neural Network Response

The dataset was randomly split into a train (90%) and validation (10%) while an extra recorded session (of speaker SR) was used as the test dataset. The model response experimented on various other optimizers like stochastic gradient descent (SGD), Adaptive Gradient (Adagrad), Adaptive Delta (Adadelta), and Adaptive Moment Estimation (Adam) among which SGD was found to be the most efficient for the model. The model was thus optimized using an SGD optimizer with an initial learning rate = 0.0025 and momentum = 0.9. Since the dataset was multi-labeled sparse categorical cross-entropy loss function was used and the model was trained for 100 epochs each with a batch size of 32.

From **Figure 7** it can be observed that the accuracy of the model reaches near 80% in the training dataset while near 70% in validation data. The accuracy in the test dataset was 74.75% which is satisfactory as per the model performance. In both graphs, the dataset

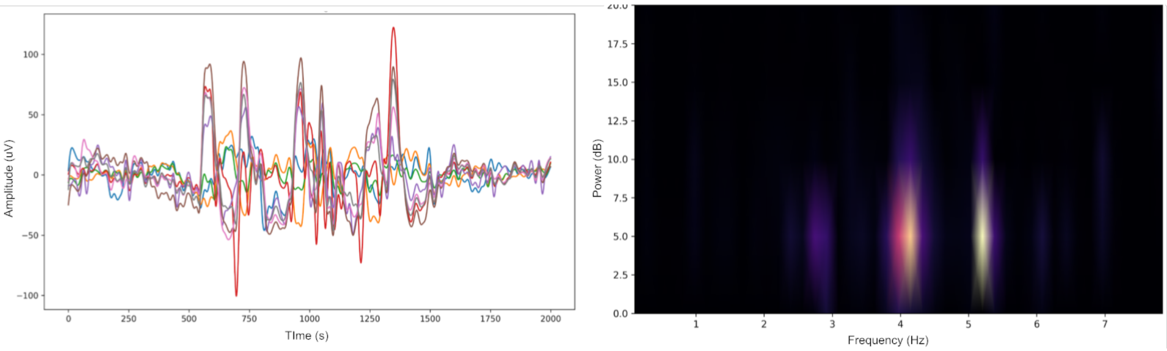

Figure 5: Time Domain (Left) and Frequency Domain/Spectrogram (Right) Analysis of Filtered sEMG Signal for Utterance “अबको समय सुनाउ”

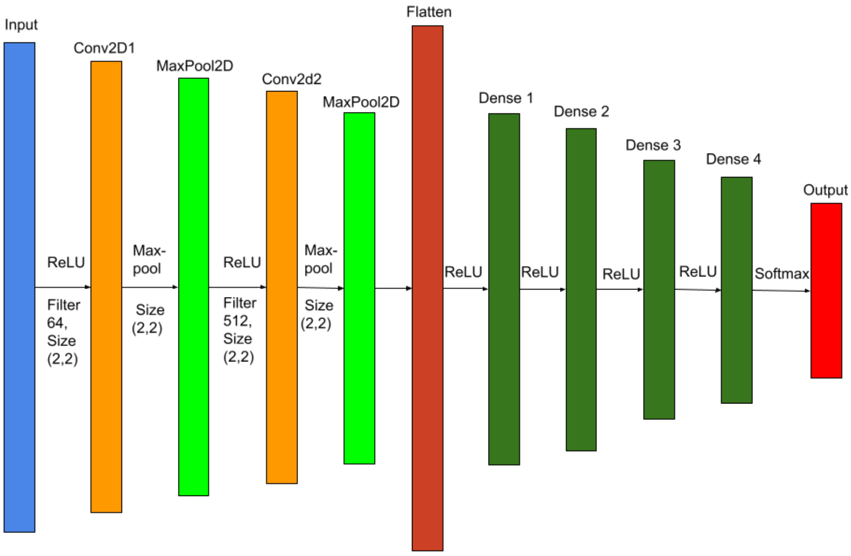

Figure 6: Architecture of the Designed CNN Model

plots have a high positive correlation and consistency. Loss plots, supporting this conclusion, show a gradual and consistent decrease in the losses of the model during both the training and validation phases. The overlap between both losses presents that the model is far from an over or under-fitted response. The result is due to the tuning of the model parameters coupled with the augmented training dataset, which provides the model with an ample amount of data to train. Furthermore, the accuracy plot shows that the model beings to saturate after 80 epochs, since the model loss is decreasing so, it is trained for 100 epochs.

The model response was also visualized with other procedures like confusion matrix, precision score, recall score, and F1 score. The confusion matrix, **Figure 8** shows significantly correct predictions for all the sentences; with very few false predictions. This prediction is also supported by the precision, recall, and F1 scores obtained as 0.75, 0.74, and 0.74 respectively as a weighted average.

The proposed system provides insight into the recognition of silently uttered discrete sentences. The prediction of the sentences is not up to the state-of-the-art, but in most cases, the predictions are correct. The consistency of EMG electrodes during each session, muscle fatigue, limited dataset, time alignment of utterance, and user dependency on the system poses some problems to the recognition model.

Initially, for the recognition of silent utterances, it was hypothesized that for every phonetically different sentence, the EMG patterns during articulation must be different, and for the same sentences must be similar. To test this assumption, amplitude (μV) vs. time (ms) (i.e. raw signal) plot of phonetically different words, as shown in **Figure 4** were analyzed which led to the conclusion that the plot for different articulations is different. However, similar articulations were also found to be different to some minor extent, mainly in terms of amplitude, time, and rate of utterance.

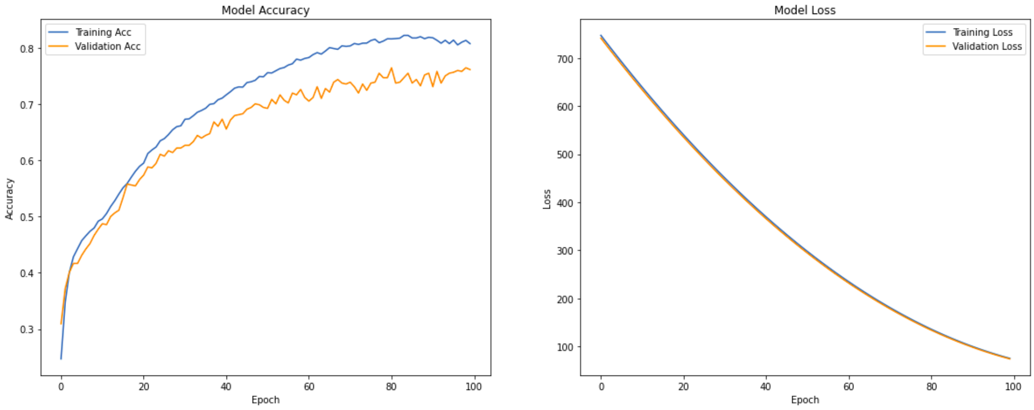

Figure 7: Performance of CNN model

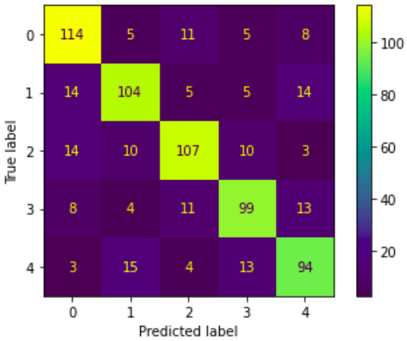

Figure 8: Confusion Matrix of the Model. The Labels 0, 1, 2, 3, and 4 respectively Denote the Sentences “बत्तिको अवस्था बदल” “आजको मौसम बताउ” “एउटा सङ्गीत बजाउ” “पङ्खाको स्थिती बदल” and “अबको समय सुनाउ”

A major limitation of the research method is during data collection from the user which includes some impeding factors such as muscle fatigue, the imprecise placing of electrodes, and insufficient area of contact of the electrode. Furthermore, the speed of the articulation differs in every instance, making it difficult to maintain some level of consistency within a session and across sessions. Another limitation of the work is dataset is limited to specific age ranges and gender. Extending the dataset by extracting signals from speakers of more diverse age and gender groups can bring diversity to the dataset.

6. Conclusion

The research shows the possibility that silently uttered discrete Nepali sentences can be predicted using silent speech recognition. Through the use of STFT as a feature and the CNN model, the recognition system can predict internally articulated words from the weak and unpredictable EMG signal. sEMG signals extracted from the facial muscles using electrodes are transmitted to the Cyton board where all the signal processing activities are accomplished with fewer improvements to the initial setup. The signals after processing are prepared as input for the machine learning model and the model is trained to predict the uttered sentences. The predicted output of the model is then displayed with accuracy and loss. Thus, EMG signals which are prominent within the frequency range of 1.5-50 Hz can be used to interact with a remote computer system. The research highlights that human-computer interaction using neuromuscular signals is feasible with a permissible error rate. The dependency on the physicality of the speaker, speaking rate, and the operating environment adds further complications to the project.

7. Future Enhancements

Some of the possible future enhancements to the system can be:

- Conduct the signal extraction procedure in various sessions which introduces variation in data making the data session independent. Moreover, it aids in reducing the effects of muscle fatigue.
- Implement the features of seamless communication with which the project can be made applicable in the medical sector involving speech abnormalities as well as human-computer interaction.
- Combine silent speech communication with virtual assistants such as Alexa, Google Assistant, and so on instead of voiced

communication. And integrate IOT devices for home automation.

- Develop a wearable device with a minimal number of electrodes while increasing their reusability for longer sessions.

Acknowledgments

The authors would like to express their deepest gratitude to the OpenBCI Community for sponsoring this research project and to the Department of Computer and Electronics, Thapathali Campus for providing an opportunity to conduct the research.

References

[1] Kirma Sembulingam and Prema Sembulingam. *Essentials of medical physiology*. JP Medical Ltd, 2012.

[2] H. Timothy Bunnell. William j. hardcastle and john laver (eds.), the handbook of phonetic sciences. oxford: Blackwell, 1997. pp. vii 904. *Journal of Linguistics*, 35(1):167–222, 1999.

[3] J Anthony Seikel, David G Drumright, and Douglas W King. *Anatomy & physiology for speech, language, and hearing*. Cengage Learning, 2015.

[4] Arnav Kapur, Shreyas Kapur, and Pattie Maes. Alterego: A personalized wearable silent speech interface. pages 43–53, 03 2018.

[5] Michael Wand and Tanja Schultz. Session-independent emg-based speech recognition. *BIOSIGNALS 2011 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, pages 295–300, 01 2011.

[6] Pol Rosello, Pamela Toman, and N. Agarwala. End-to-end neural networks for subvocal speech recognition. 2017.

[7] R. Nepal, R. Lwagun, S. Rijal, U. Subedi, and D.B. Kshatri. Human-computer interaction using neuromuscular signals. Bachelor's thesis, 2021.

[8] Gary Kamen and David A Gabriel. *Essentials of electromyography*. Human Kinetics Publishers, 2009.

[9] OpenBCI. Gold Cup Electrodes, 2019. https://shop.openbci.com/products/openbci-gold-cup-electrodes.

[10] OpenBCI. Cyton Biosensing Board, 2019. https://shop.openbci.com/products/cyton-biosensing-board-8-channel?variant=38958638542.

[11] OpenBCI. OpenBCI GUI, 2019. https://docs.openbci.com/Software/OpenBCISoftware/GUIDocs/.

[12] Hyeon Kyu Lee and Young-Seok Choi. Application of continuous wavelet transform and convolutional neural network in decoding motor imagery brain-computer interface. *Entropy*, 21(12):1199, 2019.

[13] Yanghua Wang. The ricker wavelet and the lambert w function. *Geophysical Journal International*, 200(1):111–115, 2015.

[14] National Instruments. Understanding ffts and windowing. https://www.ni.com/en-us/innovations/white-papers/06/understanding-ffts-and-windowing.html.

[15] Julius O. Smith III. *Spectral Audio Signal Processing*. W3K Publishing, 2011.

[16] Alfred Mertins and Dr Alfred Mertins. *Signal analysis: wavelets, filter banks, time-frequency transforms and applications*. John Wiley & Sons, Inc., 1999.

[17] Veysel Kocaman, Ofer M Shir, and Thomas Bäck. The unreasonable effectiveness of the final batch normalization layer. In *International Symposium on Visual Computing*, pages 81–93. Springer, 2021.