IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⵘ Peer-reviewed & Refereed journal ⵘ Vol. 11, Issue 1, January 2024

DOI: 10.17148/IARJSET.2024.11115

Decoding the Unspoken: Deep Learning-Based Recognition of Silently Spoken English Vowels Using sEMG Signals

Rajendra Kachhwaha[1], Rajesh Bhadada[2]

Research Scholar, Department of Computer Science and Engineering, MBM University, Jodhpur, India[1]

Professor, Department of Electronics and Communication Engineering, MBM University, Jodhpur, India[2]

Abstract: In the advancing field of rehabilitation technology and human-machine interfaces, surface electromyography (sEMG) has emerged as a critical non-invasive method for interpreting human intentions, particularly in developing advanced prosthetics and silent speech recognition systems. However, despite its potential, challenges such as noise interference and the necessity for precise electrode placement have constrained its accuracy. This paper explores the application of advanced deep learning (DL) models, including Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Deep Neural Networks (DNN), and Convolutional Neural Networks (CNNs) in both 1-dimensional (1D) and 2-dimensional (2D) formats to improve the interpretability of sEMG signals for silent speech recognition. The proposed setup utilized a multithreading queuing (MTQ) based novel three-channel low-cost sEMG data acquisition system for English vowel recognition. The two channels are responsible for collecting and extracting data, while the third channel helps visualize data in real-time. It involves data acquisition using disposable electrodes across key facial muscles, followed by employing a range of DL models to process and classify the sEMG signals. Our findings suggest that advanced DL models, particularly the CNN-2D model, outperformed other state-of-the-art methods by achieving 90% accuracy in vowel recognition, showcasing the potential of deploying low-cost hardware with new predictive paradigms in sEMG analysis.

Keywords: Surface Electromyography (sEMG), Silent Speech Recognition, Deep Learning Models, Rehabilitation Technology

I. INTRODUCTION

In the rapidly evolving domain of rehabilitation technology and human-machine interfaces, the use of surface electromyography (sEMG) signals emerges as a pivotal approach for interpreting human intentions [1][2]. sEMG, a non-invasive method for measuring muscle activity, offers insights into muscle fibers' electrical signals during contractions [3], playing a critical role in developing advanced prosthetics and enabling silent speech recognition systems [4][5].

Despite its potential, challenges such as noise interference and the need for precise electrode placement hamper signal accuracy [6][7][8]**.** The progression of sEMG data analysis has shifted from understanding muscle activity to applying advanced computational models, aiming to improve sEMG signal interpretability. Machine learning classifiers, including K-nearest neighbors (KNN), Support vector machines (SVM), and Artificial neural networks (ANN), have facilitated low-cost sEMG data acquisition systems' effectiveness in silent speech recognition, comparable to commercial setups [7][9][10][11].

Our prior research [7] has been instrumental in developing algorithms for efficient sEMG data collection and processing, demonstrating significant accuracies in recognizing English vowels from facial muscle activity. Furthermore, we have focused on a novel three-channel multithreading queuing (MTQ) based low-cost sEMG data acquisition system that supports real-time signal visualization, highlighting the potential of merging cost-effective technology with a new predictive paradigm [10].

The emergence of deep learning (DL) has significantly enhanced computational modeling for sEMG analysis. DL models, such as Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Deep Neural Networks (DNN), and Convolutional Neural Networks (CNNs) in both 1-dimensional and 2-dimensional formats have proven effective in various fields, including natural language processing and image recognition. Their ability to learn complex features from data makes them ideal for analyzing sEMG signals, capturing essential temporal and spatial patterns.

DOI: 10.17148/IARJSET.2024.11115

This research aims to further extend our previous work [7][10] by incorporating advanced DL models into the sEMG analysis framework, specifically for silent speech recognition. By using the same low-cost hardware and data collection methodologies, we seek to assess the effectiveness of LSTM, Bi-LSTM, DNN, CNN-1D, and CNN-2D models in identifying sEMG signals' complex patterns.

This approach is anticipated to improve classification accuracy, robustness, and adaptability, marking a significant advancement in sEMG analysis. Moreover, the contributions of our study include:

1. Integrating advanced deep learning models to enhance feature learning from raw sEMG data, eliminating manual feature engineering.

2. Enabling real-time sEMG signal processing and visualization through a multithreading queuing-based algorithm, for applications requiring immediate feedback.

3. Providing a low-cost, high-performance solution that makes sophisticated rehabilitation technologies more accessible.

4. We assess the quantifiable effectiveness of the proposed low-cost, high-performance setup by comparing it with various state-of-the-art methods.

II. RELATED WORK

This section synthesizes the literature on vowel-based sEMG systems, including insights from our previous studies, to highlight the progression toward more accurate and accessible silent speech recognition technologies. Kumar et al. [12] were pioneers in applying ANN to sEMG data for speech recognition, achieving an impressive 88% success rate in vowel recognition from three facial muscles.

This early work set the foundation for subsequent studies focusing on eliminating auditory clues from speech recognition. Arjunan et al. [13][14] extended this line of research by successfully classifying five English vowels with up to 86% accuracy using ANN and data collected from four facial muscles, marking a significant step forward in the field. Larraz et al. [11] further explored the potential of sEMG in silent speech vowel recognition, achieving over 70% accuracy using a raw dataset from eight muscles, demonstrating the diversity of muscle involvement. Mostafa et al. [15]

Introduced a novel non-invasive tool for recognizing eleven Bangla vowels, showcasing the versatility of sEMG in speech recognition across languages with an overall accuracy of 82.3%. Agnihotri et al. [16] demonstrated the robustness of neural networks in classifying three English vowels with an 85% recognition rate. Japanese vowels were also explored by Takabatake et al. [17][18], who reported classification accuracies of 33% and 62.33% using KNN and SVM classifiers, respectively highlighting the challenges and opportunities in vowel classification. Fraiwan et al. [19] and Manabe et al. [20] both developed customized hardware for sEMG data collection, achieving accuracies up to 77% in recognizing Arabic vowels and significant identification rates for vowels uttered in Japanese, respectively. Chandrashekhar [9] developed a Silent Speech Interface (SSI) using an MWM sensor for English vowel recognition, achieving an 80% success rate with the SVM approach, illustrating the effectiveness of specialized hardware. Building upon these foundational studies, our previous research [7] and [10] introduced an innovative sEMG data acquisition system, showcasing significant improvements in classification accuracy.

The authors in [7] utilized custom machine learning classifiers, achieving nearly 83% accuracy in English vowel recognition, emphasizing our contribution to enhancing classification techniques. These advancements highlight the potential of combining low-cost hardware with sophisticated data processing techniques to enhance silent speech recognition systems further.

Table 1 offers a comprehensive synopsis of the research papers under discussion, capturing essential elements, including the number of classes, number of subjects, number of channels, muscles under observations, employed classifiers, and the reported accuracies. This summary provides a quick reference to understand the scope and results of each research by making it easier to grasp the evolution of sEMG-based silent speech recognition technologies.

DOI: 10.17148/IARJSET.2024.11115

TABLE 1 SUMMARY OF RELATED WORK

Ref.	Vowel & Hardware Used	Classes	Subjects	Channels	Muscles	Classifier	Accuracy (%)
[12]	English (Comm.)	5	3	3	Mentalis, Messetter, Depressor anguli oris	ANN	88
[13]	English (Comm.)	5	3	4	Zygomaticus major, Mentalis, Masseter, Depressor anguli oris	ANN	80
[14]	English, German (Comm.)	3-3	3	4	Zygomaticus major, Mentalis, Masseter, Depressor anguli oris	ANN	86
[21]	English (Comm.)	5	-	4	Zygomaticus major, Mentalis, Masseter, Depressor anguli oris	ANN	60
[11]	Spanish (Low Co.)	30	3	8	Levator labii superioris, Risorius, Platysma, Zygomaticus major, Orbicularis oris, Depressor anguli oris, Depressor labii inferioris, Digastric	DT, DT with Ada Boost	42, 62
[19]	Arabic (Low Co.)	3	20	3	Orbicularis Oris, Triangularis, Risorius	RF	82
[15]	Bangla (Low Co.)	11	8	3	Massester, Buccinators, Depressor	ANN	82
[16]	English (Comm.)	5	-	-	Zygomaticus major, Mentalis, Masseter, Depressor anguli oris	ANN	85
[17]	Japanese	5	1	3	Orbicularis Oris, Zygomatic, Depressor angle oris	KNN	33
[18]	Japanese	5	1	3	Orbicularis Oris, Zygomatic, Depressor angle oris	SVM	62
[9]	English (Low Co.)	5	1	-	Submental triangle (area under neck)	CNN, SVM, KNN	55, 80, 67
[7]	English (Low Co.)	5	1	3	Orbiclaris Oris, Masseter, Digastric	ANN, SVM, KNN	82, 83, 84

Comm.: Commercial Hardware, Low Co.: Low-cost/Self-Developed Hardware, DT: Decision Tree, RF: Random Forest

III. METHODOLOGY

This section includes the details about data acquisition, muscles under observation, experimental procedure, preparation of dataset, and architectural details of each deep learning model, including layers, activation functions, optimization techniques, and loss functions used.

A. Data Acquisition

The choice of the MTQ technique and hardware for data collection is driven by its proven efficiency in previous studies [10], aiming to achieve high-quality data while maintaining cost-effectiveness and user convenience. This is employed to collect silently spoken data of English vowels A, E, I, O, U, and 'Silence' when the person remains quiet. To uniquely identify a category of each recorded instance inside the collection, each class is automatically represented by an integer number during recordings and presented in Table 2. Using integer numbers to represent each class simplifies the data analysis and allows for easier comparison between different studies.

TABLE 2 VOCABULARY CODES

Syllabus	Silence	A	E	I	O	U
Syllabus	Silence	A	E	I	O	U
Unique Number	0	1	2	3	4	5

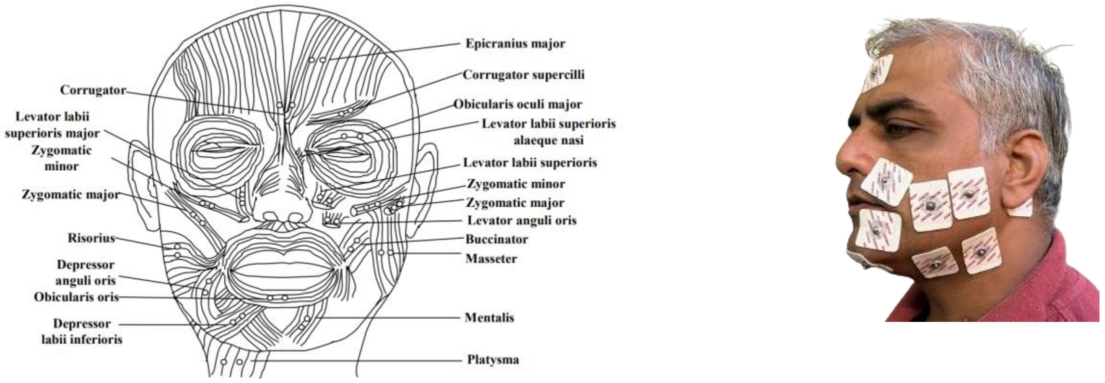
DOI: 10.17148/IARJSET.2024.11115

In this study, sEMG signals from facial muscles were recorded using disposable Ag/AgCl electrodes that had a gel surface area of 2 cm^2 and a sensor area of 0.8 cm^2. This electrode configuration is selected to optimize signal quality and minimize skin irritation, ensuring participant comfort throughout the data collection process.

B. Muscles Under Observation

Researchers have employed distinct facial muscles to obtain sEMG data for the production of silently articulated syllables. Identifying the most informative facial muscles for sEMG data collection is crucial for silent speech recognition accuracy. This research involved an examination of three specific muscles located in the facial and neck region, namely the Orbicularis Oris (M1), Masseter (M2), and Digastric (M3) muscles, as depicted in Figure 1 (a) [4][5] and Figure 1 (b) shows actual placement of electrodes on subject's face and neck region. These muscles are chosen based on their significant involvement in speech articulation and the potential for clear sEMG signal differentiation.

The Orbicularis Oris muscle, also known as M1, is a circular muscle located in the region of the lips. The Masseter muscle extends laterally along the mandibular ramus angle and surface from the zygomatic arch. The Digastric muscle located in the cervical region can elevate the hyoid bone, thereby facilitating an increased orifice of the oral cavity. The utilization of the digastric muscle has been demonstrated to indicate the involvement of the neck in the generation of silent speech. The selection of these particular muscles was predicated upon their capacity to furnish precise and dependable sEMG information to recognize silent speech.



(a) (b)

Fig. 1 Facial muscles location: (a) Human face muscles (b) Actual electrode placement

C. Algorithm Integration with MTQ

The integration of the three-channel MTQ technique with a multi-threading approach underscores our commitment to real-time, efficient data processing. This technique ensures efficient data processing, minimizing the risk of erroneous data collection.

The MTQ's design allows for a seamless flow of data between collection, processing, and visualization, highlighting its utility in handling high-throughput sEMG data. Two successive data structures (queues) are utilized to handle temporary data, while data transmission is managed by three independent processes (T1, T2, and T3). T1 receives data, T2 transmits data across queues, and T3 changes the visual data structure and creates a real-time graph with recording whenever necessary.

The data structure's size can be restricted to a range from 300 to 1000 sample values. Here, for this study, it is set to 300 for each channel. The system constantly checks the state of the connection, and data processing happens only when the hardware is correctly connected and the user interacts with the interface controls.

Data processing includes capturing the signal, transforming, and creating a dataset that includes temporal, demographic, and measurement data. When users direct the system to reset or terminate, it performs the appropriate cleanup tasks. These methods end the algorithm's execution and show an acknowledgment message. The MTQ system is responsible for the initial handling and preprocessing of sEMG signals, making them suitable for input into the DL models. The methodological flowchart is presented in the following Figure 2.

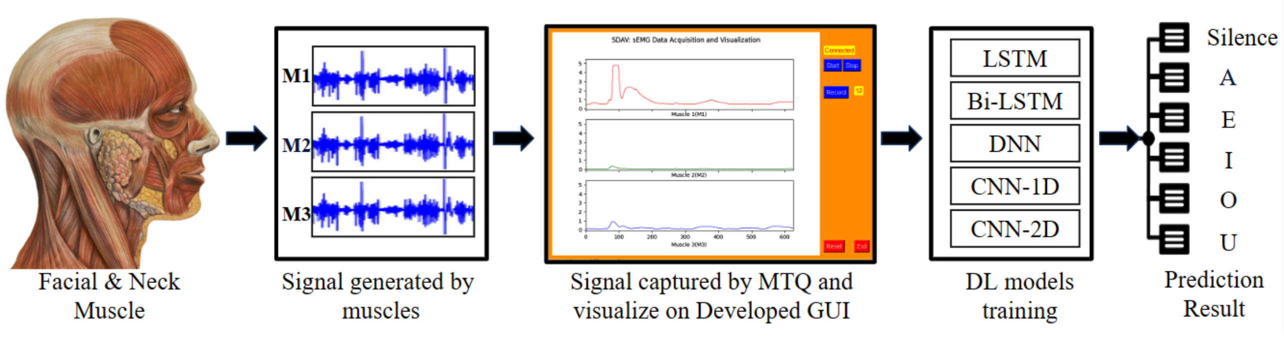

M1
M2
M3
SDAV: sEMG Data Acquisition and Visualization
LSTM
Bi-LSTM
DNN
CNN-1D
CNN-2D
Silence
A
E
I
O
U
Facial & Neck Muscle
Signal generated by muscles
Signal captured by MTQ and visualize on Developed GUI
DL models training
Prediction Result

Fig. 2 Flowchart of methodology

D. Experimental Procedure

To ensure verifiability and eliminate the potential influence of individual human characteristics such as age, gender, accent, fluency, and the like, the data was consistently gathered from a male subject who was 40 years of age, and in good health, with no known speech impairments. The selection of a single subject also aimed to minimize variability in the dataset, focusing on the system's ability to recognize silent speech patterns. The individual in question exhibited proficient communication skills in the English language. The participant was informed about the methodology and procedures involved in the recording process. The data was collected by positioning the participant in a solitary chair in front of a computer monitor. The temperature of the room was regulated within the range of 24-26 ºC to prevent the occurrence of perspiration [10].

E. Dataset Preparation

Getting the dataset ready for a classification task is a key step in making a model that is accurate and reliable. During experimental sessions, 50 recordings for each vocabulary content (A, E, I, O, U, Silence) were obtained using [10], from three facial muscles. The length of one sEMG recording data value is 900 and all recording is made in a CSV file. Here, we are using a ratio of 80% - 20% as our training and testing datasets.

F. Model Architecture

To detect potential patterns in collected sEMG data, it is necessary to employ diverse classification algorithms. The implementation of these techniques can facilitate comprehension of the fundamental muscle activation patterns associated with diverse tasks or movements. This study employed five distinct deep-learning methodologies to investigate and capture different aspects of the sEMG data, leveraging the strengths of deep learning in handling time-series data and extracting complex features. The neural network models under consideration are LSTM, Bi-LSTM, DNN, and two variants of CNN. The LSTM networks were developed to solve the vanishing gradients that commonly affect conventional recurrent neural networks (RNN). The issue of vanishing gradients arises in the context of backpropagation training, wherein the gradients utilized by the algorithm diminish to an extremely small magnitude, thereby impeding the comprehension of long-term connections [22]. LSTM networks consist of interconnected cell clusters. Every individual cell is equipped with three gates that are responsible for controlling the flow of information into and out of the cell, as well as into the memory storage. The input gate regulates the inflow of data into the cell, the output gate governs the outflow of data, and the forget gate determines the retention or elimination of data. The data flow in question is regulated by gates that employ sigmoid activation functions [22]. The core LSTM unit is defined by the following equations:

$$f_t = \sigma\,(W_f \cdot [h_{t-1}, x_t] + b_f\,)$$
$$i_t = \sigma\,(W_i \cdot [h_{t-1}, x_t] + b_i\,)$$
$$\check{C}_t = tanh\,(W_C \cdot [h_{t-1}, x_t] + b_C\,)$$
$$C_t = f_t * C_{t-1} + i_t * \check{C}_t$$
$$o_t = \sigma\,(W_o\,[h_{t-1}, x_t] + b_o\,)$$
$$h_t = o_t * tanh\,(C_t)$$

Where σ is the sigmoid function, W and b are weights and biases, and h_t, C_t are the hidden state and cell state at time t.

Standard LSTM models are limited in their ability to learn in a unidirectional manner, which hinders their capacity to predict novel words. The Bi-LSTM is a type of neural network that addresses the limitations of conventional LSTM models. It achieves this by processing input sequences in both forward and backward directions. This approach enables the Bi-LSTM to capture complex phrase patterns and context more accurately [23]. The Bi-LSTM architecture employs a pair of LSTMs to handle input sequences bidirectionally, so it can predict the present phrase or label by utilizing both preceding and succeeding data. Bi-LSTM combines the forward LSTM $\overrightarrow{h}_t$ and backward LSTM $\overleftarrow{h}_t$ as:

$$h_t = [\overrightarrow{h}_{t,} \overleftarrow{h}_t]$$

The increased utilization of DNNs can be attributed to their exceptional proficiency in handling intricate tasks [24]. A fundamental characteristic of DNN is the utilization of feedback connections, whereby the outputs of a given layer of neurons are subsequently fed as inputs into the subsequent layer of neurons. Due to their multi-layered architecture, DNNs can acquire and communicate complex and abstract connections between input and output. The backpropagation algorithm is employed in the training of DNN to adjust the synaptic weights connecting neurons based on the discrepancy between the anticipated and observed output. Through multiple iterations, the precision of the network progressively enhances until it attains the desired threshold [24]. The output of each layer in DNN is given by:

$$a^{[l]} = g^{[l]} \left(W^{[l]} \cdot a^{[l-1]} + b^{[l]}\right)$$

Where $g^{[l]}$ is the activation function for layer l

CNNs are a type of artificial neural network that has been extensively utilized in various domains, including but not limited to image recognition and natural language processing. The fundamental building block of CNN architecture is the convolutional layer. In a convolutional layer, the input data undergoes convolution with a fraction of the input at each filter. The convolutional layer produces feature maps that accentuate distinct characteristics of the input. Subsequently, the feature maps are transmitted to the succeeding layer of the network. CNNs possess a notable edge over conventional machine learning algorithms owing to their innate capability to autonomously acquire features from unprocessed data, as stated by [25]. The present study employs two distinct forms of CNN, namely CNN-1D and CNN-2D. The CNN-1D model specializes in processing one-dimensional sequence data, also referred to as Conv-1D. CNN-1D designed for time-series data, it applies 1D convolution operations, capturing temporal dependencies. The convolution operation in CNN-1D is:

$$C_t = f\,(W \cdot X_{t:t+k-1} + b)$$

Where $X_{t:t+k-1}$ represents the input segment, W is the filter, b is the bias, and f is the activation function

On the other hand, CNN-2D is capable of accommodating a diverse array of two-dimensional inputs and is also known as Conv-2D. CNN-2D is useful for capturing spatial features from multichannel sEMG data. The 2D convolution operation is given by:

$$C_{i,j} = f\left(\sum_m \sum_n W_{m,n} \,.\, X_{i+m,j+n} + b\right)$$

Where $W_{m,n}$ is the filter applied to the input X at position (i, j)

Both CNN-1D and CNN-2D employ a series of convolutional layers on the input sEMG data to extract features. The deployed classifiers' customization information is summarised in Table 3.

TABLE 3 CUSTOMIZATION DETAILS OF EACH TECHNIQUE

Classifier	Hyper-parameters
LSTM	units ← 50, input_shape ← (30,30), activation←'softmax', epoch←50, batch_size←8
Bi-LSTM	units ← 50, input_shape ← (30,30), activation←'softmax', epoch←50, batch_size←8
DNN	filters←32, kernel_size←5, layer activation←'relu', epoch←50, input_shape←(900, 1), learning_rate←0.0001, beta_1←0.9, beta_2←0.999, optimizer←Adam, loss←'categorical_crossentropy', last layer activation←'softmax', batch_size←8
CNN-1D	filters←32, kernel_size←5, layer activation←'relu', epoch←50, input_shape←(900, 1), learning_rate←0.0001, beta_1←0.9, beta_2←0.999, optimizer←Adam, loss←'categorical_crossentropy', last layer activation←'softmax', batch_size←8

DOI: 10.17148/IARJSET.2024.11115

CNN-2D	filters←32, kernel_size←(5, 5), layer activation←'relu', epoch←50, input_shape←(30, 30, 1), learning_rate←0.0001, beta_1←0.9, beta_2←0.999, optimizer←Adam, loss←'categorical_crossentropy', last layer activation←'softmax', batch_size←8

G. Evaluation Metrics

Evaluation of model performance is conducted by assessing its accuracy. It indicates what proportion of the events in the collection has been correctly classified. The percentage of accurate predictions produced by the classifier (including TP and TN) over the total number of predictions and denoted by the following formula:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Where:

- TP (True Positive): Number of successfully detected positive instances by the classifier
- TN (True Negative): Number of negative instances properly detected by the classifier
- FP (False Positive): Number of negative examples that the classifier wrongly categorized as positive
- FN (False Negative): Number of positive examples that the classifier mistakenly labeled as negative

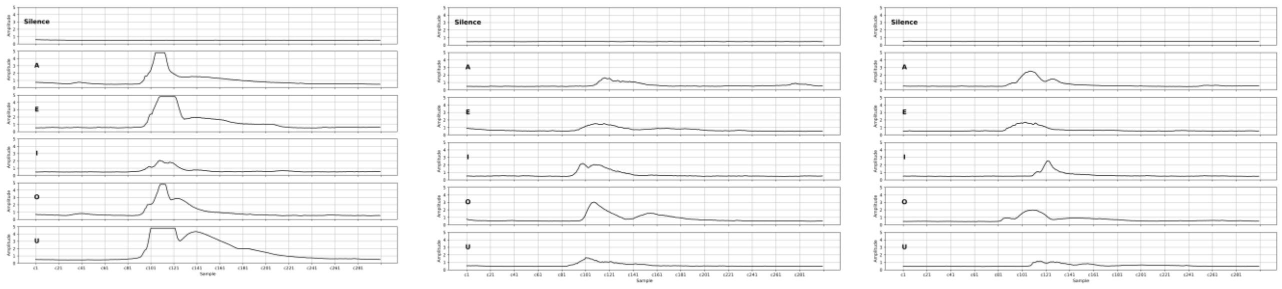
It is possible to assess a classifier's efficacy by looking at its accuracy rate. When there is a disparity between classes or when misclassifying certain classes is more costly than misclassifying others, this approach may not be the best solution. A further assessment measure used is the confusion matrix. A confusion matrix is a tabular representation that illustrates the accuracy of a classifier's classification by presenting the accurate and inaccurate classifications for each category. It is an appropriate approach for evaluating the performance of a classifier and identifying erroneous classifications.

IV. OBSERVATIONS AND RESULTS

This study delves into the comparative analysis of advanced deep learning architectures, including LSTM, Bi-LSTM, DNN, and two variants of CNN (CNN-1D, CNN-2D), in the context of silent speech recognition using sEMG data. Employing the MTQ method for data acquisition [10], this research aims to elucidate the efficacy of these models in discerning silently spoken English vowels, thereby offering insights into their potential advantages and limitations for future applications.

A. Observations

The initial phase of our investigation involved the visualization of real-time raw sEMG patterns corresponding to each vowel across three distinct facial muscles, as facilitated by the MTQ technique. This visualization, illustrated in Figure 3, revealed that the sEMG patterns for each vowel are uniquely distinguishable, thereby providing a robust foundation for the subsequent classification analysis.



(a) For muscle M1 (b) For muscle M2 (c) For muscle M3

Fig. 3 Recorded sEMG data for vowel dataset from each muscle

B. Results

The classification outcomes (refer Figure 4), post 50 epochs of model training, shed light on the performance of each deep learning architecture concerning accuracy, loss, and confusion matrix metrics for both training and testing datasets.

The LSTM model achieved an accuracy of 85% and 78.33% on the training and testing datasets, respectively. The corresponding losses were 0.4313 and 0.3668. The findings indicate that the performance of the model on the training and testing datasets was nearly identical. From Figure 4 (i) (c), the confusion matrix was employed to evaluate the performance of the model. According to the confusion matrix, the model exhibited a relatively lower accuracy level in identifying vowels, particularly concerning the vowels 'A' and 'O', in comparison to the remaining vowels.
It is recommended that additional refinement of the model may enhance its efficacy in discriminating among vowel phonemes.

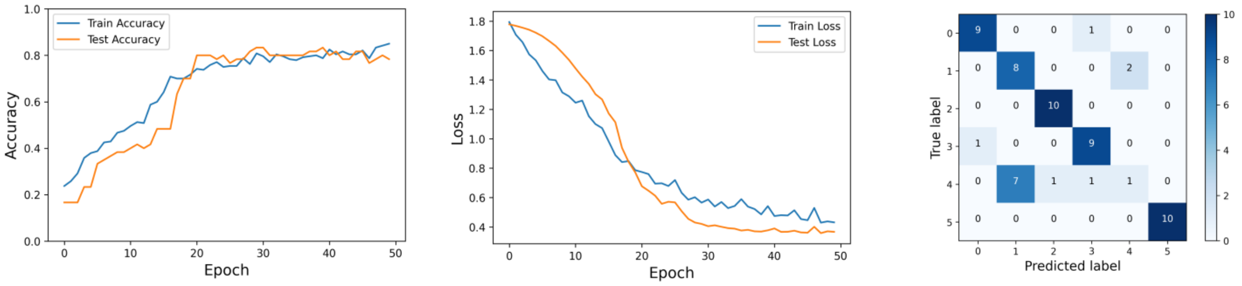
The Bi-LSTM model achieved an accuracy of 85.83% and 80% on the training and testing datasets, respectively, with corresponding losses of 0.3837 and 0.6358. The findings suggest that the performance of the model is satisfactory on both the training and testing datasets. From Figure 4 (ii) (c), to see how well the model worked. The confusion matrix revealed that the model was slightly less accurate in identifying the vowel 'O' than other vowels. It is suggested that further fine-tuning of the model could improve its performance in distinguishing between different vowel data.

The DNN model achieved an accuracy of 98.75% and 80% on the training and testing datasets, respectively. The corresponding losses were 0.0585 and 0.8665. The results show that the model performs well on the training datasets but slightly less on the testing ones. From Figure 4 (iii) (c), the confusion matrix revealed that the model correctly predicts the 'E', 'O', and 'U' vowels from the testing dataset while being slightly less accurate in identifying the vowel 'A' and 'I' than other vowels. This shows that the model's ability to discriminate between distinct vowel data may be fine-tuned even more.

The CNN-1D model achieved accuracies of 100% and 83.33% on the training and testing datasets, respectively. The corresponding losses were 0.0006 and 0.6363. The findings indicate that the model exhibits satisfactory performance on the training datasets, albeit marginally lower on the testing datasets. From Figure 4 (iv) (c), to see how well the model worked. The confusion matrix revealed that the model almost correctly predicts all vowels from the testing dataset while being slightly less accurate in identifying the vowel 'I' than other vowels.

Remarkably, the CNN-2D model achieved an accuracy of 99.58% and 90% on the training and testing datasets, respectively. The corresponding losses were 0.0165 and 0.3233. The findings indicate that the model performs satisfactorily on the training and testing datasets. Figure 4 (v) (c), was employed to evaluate the efficacy of the model. The confusion matrix analysis indicates that the model demonstrates a high accuracy level in predicting most vowels from the testing dataset, with slightly lower accuracy in identifying the vowel 'I' compared to the other vowels.

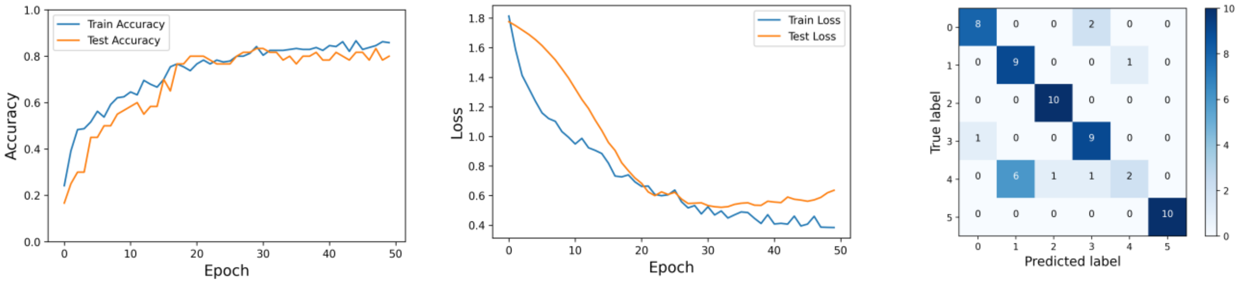
Further, Table 4 presents the classification outcome achieved after 50 epochs, with respect to the accuracy and loss values for both the training and testing datasets.



(a) (b) (c)

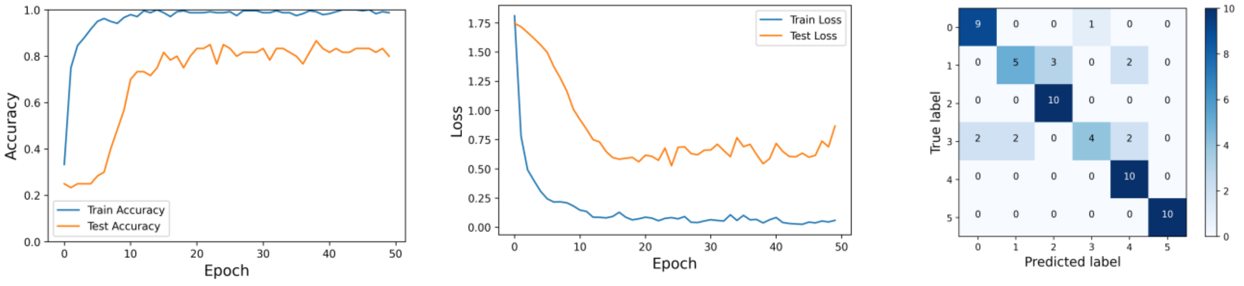
(i) Using LSTM

DOI: 10.17148/IARJSET.2024.11115



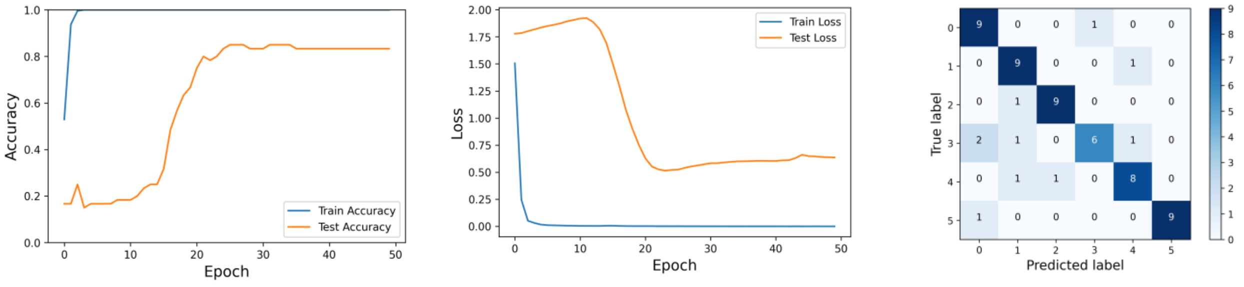
(a) (b) (c)

(ii) Using Bi-LSTM



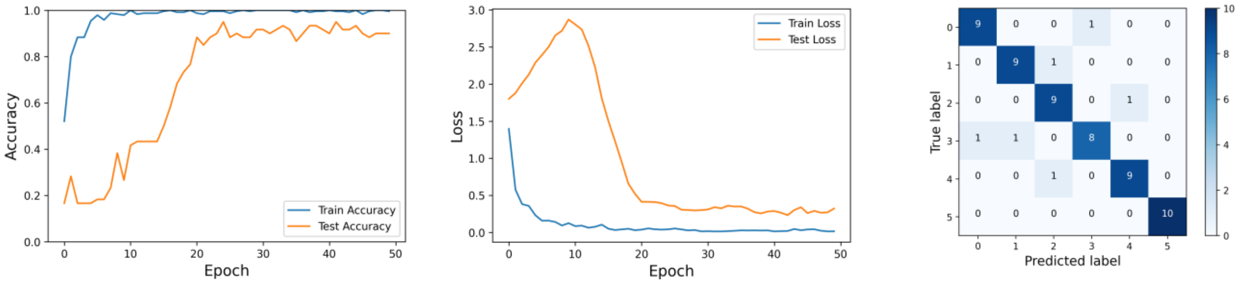
(a) (b) (c)

(iii) Using DNN



(a) (b) (c)

(iv) Using CNN-1D



(a) (b) (c)

(v) Using CNN-2D

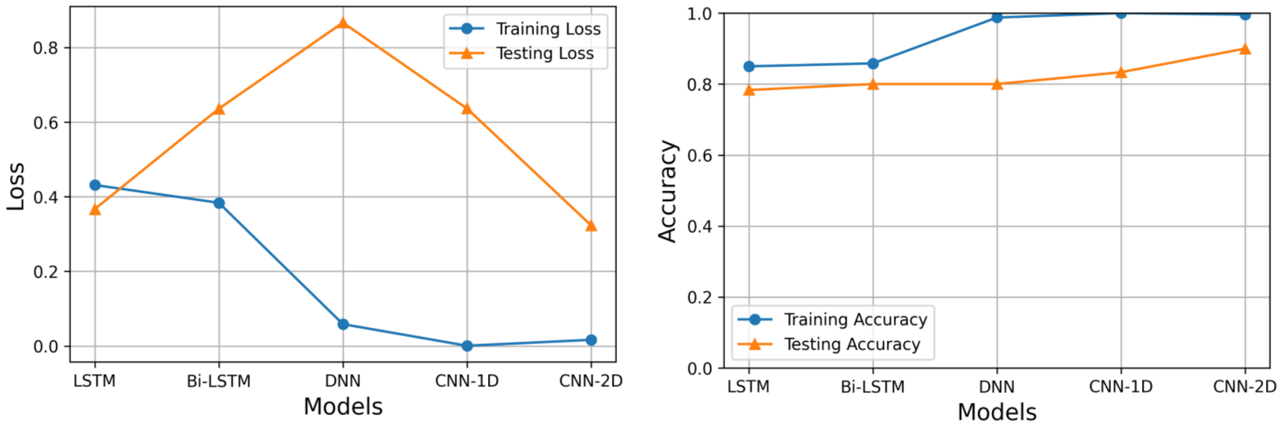
Fig. 4 Classification result in terms of (a) Accuracy, (b) Loss, (c) Confusion matrix for each deployed classifiers

TABLE 4 Comparison of the result obtained from various DL models

Model	Training Loss	Training Accuracy	Testing Loss	Testing Accuracy
LSTM	0.4313	85%	0.3668	78.33%
Bi-LSTM	0.3837	85.83%	0.6358	80%
DNN	0.0585	98.75%	0.8665	80%
CNN-1D	0.0006	100%	0.6363	83.33%
CNN-2D	0.0165	99.58%	0.3233	90%

The comparison results, from Table 4, indicate that the Bi-LSTM model exhibits superior performance compared to the LSTM models, achieving an accuracy of 80% and a loss of 0.6358 for the testing dataset. The DNN model exhibits a commensurate level of precision, at 80%, when contrasted with the Bi-LSTM model.

The utilization of both Bi-LSTM and DNN models results in a 2.13% increase in accuracy when compared to the LSTM model. The CNN-1D model exhibits favourable outcomes, achieving an accuracy rate of 83.33% and a loss value of 0.6363 when applied to the testing dataset. This represents an enhancement of 6.38%, 4.16%, and 4.16% over the LSTM, Bi-LSTM, and DNN models, respectively. The CNN-2D model exhibits superior performance compared to alternative models, achieving an accuracy of 90% and a loss value of 0.3233. The aforementioned model demonstrates an enhancement in precision by 14.89%, 12.5%, 12.5%, and 8% for the LSTM, Bi-LSTM, DNN, and CNN-1D models, correspondingly. The graphic representation of the aforementioned Table 4 is presented in Figure 5.



(a) (b)

Fig. 5 Comparison of (a) Loss, (b) Accuracy, obtained employed classifiers

The findings indicate that the CNN-2D model exhibits the highest level of suitability for the given dataset and may be effectively employed in analogous classification endeavors. Additional investigation may be conducted to examine alternative methods for maximizing the model and enhancing its efficacy.

C. Comparison with State-of-the-art Methods

In this section, we compare the results of this low-cost high-accuracy setup with existing state-of-the-art methods. This comparison with previous research elucidates significant advancements in the field of silent speech recognition using sEMG signals, particularly emphasizing the efficacy of low-cost hardware solutions. Our proposed method, leveraging a self-developed, low-cost hardware setup combined with a CNN in 2-dimensional format (CNN-2D), achieved an impressive accuracy of 90%. The comparison results against different methods are tabulated in Table 5. Starting with the findings from Kumar et al. [12], which used commercial hardware paired with Artificial Neural Networks (ANN) to achieve an accuracy of 88%, our study demonstrates a 2.27% improvement. This comparison highlights the narrowing gap between commercial and low-cost hardware in terms of performance, with our CNN-2D model surpassing the established benchmark by a significant margin. Arjunan et al., utilize commercial hardware and ANN, reported accuracies of 80% [13] and 86% [14], respectively. The improvements in accuracy achieved in our proposed low-cost high-accuracy setup are 12.5% and 4.65%, emphasizing the advancements in algorithmic efficiency and hardware capability over time.

TABLE 5 COMPARE WITH PREVIOUS STUDIES

Ref.	Hardware Type	Classifier	Accuracy (%)	Percentage Improvement
[12]	Commercial	ANN	88	↑ 2.27%
[13]	Commercial	ANN	80	↑ 12.5%
[14]	Commercial	ANN	86	↑ 4.65%
[21]	Commercial	ANN	60	↑ 50%
[16]	Commercial	ANN	85	↑ 5.88%
[9]	Low-cost/ Self Developed	CNN SVM KNN	55 80 67	↑ 63.6% ↑ 12.5% ↑ 34.3%
[7]	Low-cost/ Self Developed	ANN SVM KNN	82 83 84	↑ 9.75% ↑ 8.43% ↑ 7.14%

The study by Naik et al. [12], which also employed commercial hardware and ANN, showed a relatively lower accuracy of 60%. Our approach marks a substantial 50% improvement, underlining the significant strides made in both the understanding of sEMG data and the development of more sophisticated DL models. Umesh et al. [16], with an 85% accuracy using commercial hardware, observed a 5.88% improvement in our study. This comparison further solidifies the argument that low-cost hardware, when coupled with advanced DL techniques, can achieve or even surpass the performance of more expensive commercial setups. Chandrashekhar's research [9] is particularly noteworthy as it directly compares commercial and self-developed, low-cost hardware. Using a variety of classifiers (CNN, SVM, KNN), he achieved accuracies ranging from 55% to 80%. Our study's CNN-2D model outperforms these results significantly, with percentage improvements ranging from 63.6% for CNN to 12.5% for SVM, and 34.3% for KNN, showcasing the effectiveness of our approach in leveraging low-cost hardware for high-accuracy applications. The research by Kachhwaha et al. [7], utilizing self-developed, low-cost hardware with ANN, SVM, and KNN classifiers, achieved accuracies of 82%, 83%, and 84%, respectively. Our study presents improvements of 9.75%, 8.43%, and 7.14% over these results, indicating the superior capability of our CNN-2D model in extracting and learning from the complex features of sEMG data. These outcomes of our proposed low-cost high-accuracy setup not only demonstrate a significant improvement over existing methodologies but also contribute to the ongoing discourse on making technologies more accessible and affordable. The results underscore the viability of CNN-2D models as a promising approach for enhancing the accuracy and efficiency of sEMG-based communication systems, paving the way for further research and development in this exciting domain.

V. CONCLUSION

In this research focused on the evaluation of a sEMG-based system for speech recognition, specifically targeting individuals with speech disabilities. This study utilizes various deep learning models such as LSTM, Bi-LSTM, DNN, CNN-1D, CNN-2D, to recognize silently spoken English vowels from three facial muscles. The quantitative results revealed a high classification accuracy of 90% for sEMG signals, indicating the promising potential of this technology for real-world applications. Notably, the CNN-2D model showcased its efficacy by 14.89%, 12.5%, 12.5%, and 8% for the LSTM, Bi-LSTM, DNN, and CNN-1D models, respectively. By leveraging deep learning techniques, this research provides a stepping stone toward the practical implementation of sEMG-based speech recognition systems. Further research efforts can focus on refining the system's accuracy and expanding its scope to encompass a wider range of words and phrases. Ultimately, the integration of sEMG-based speech recognition systems into mainstream assistive technologies holds great potential in improving the communication and overall quality of life for those in need.

REFERENCES

[1]. J. V. Basmajian, “Muscles alive. their functions revealed by electromyography,” *Academic Medicine*, vol. 37, p. 802, August 1962.

[2]. C. J. De Luca, “Physiology and mathematics of myoelectric signals,” *IEEE Transactions on Biomedical Engineering*, vol. BME-26, pp. 313–325, June 1979.

[3]. C. J. De Luca, “Surface electromyography: Detection and recording,” *Technical Report*, vol. 10, pp. 1–10, June 2002.

[4]. B. Lapatki, D. Stegeman, and I. Jonas, “A surface emg electrode for the simultaneous observation of multiple facial muscles,” *Journal of neuroscience methods*, vol. 123, no. 2, pp. 117–128, 2003.

[5]. A. Merlo, D. Farina, and R. Merletti, “A fast and reliable technique for muscle activity detection from surface emg signals,” *IEEE transactions on biomedical engineering*, vol. 50, pp. 316–323, March 2003.

[6]. R. Kachhwaha, A. P. Vyas, and R. Bhadada, “Adaptive threshold-based approach for facial muscle activity detection in silent speech emg recording,” in *Proceedings of 6th International Conference on Recent Trends in Computing, (Singapore)*, pp. 83– 98, Springer Singapore, 2021.

[7]. R. Kachhwaha, A. P. Vyas, R. Bhadada, and R. Kachhwaha, “Sdav 1.0: A low-cost sEMG data acquisition and processing system for rehabilitation,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, pp. 48–56, March 2023.

[8]. A. P. Vyas and R. Bhadada, “Feature extraction cum frequency analysis system for facial surface electromyography signals based human speech recognition,” *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 5, pp. 1998–2006, December 2017.

[9]. V. Chandrashekhar, “The classification of EMG signals using machine learning for the construction of a silent speech interface,” *The Young Researcher*, vol. 5, pp. 265–283, July 2021.

[10]. R. Kachhwaha, A. P. Vyas, R. Bhadada, and R. Kachhwaha, “A multithreading queuing based sEMG data acquisition system for rehabilitation with real-time visualization,” *Journal of Rehabilitation Sciences and Research*, January 2024. Accepted for publication.

[11]. E. Lopez-Larraz, O. M. Mozos, J. M. Antelis, and J. Minguez, “Syllable-based speech recognition using emg,” in *Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 4699–4702, IEEE, Buenos Aires, Argentina, September 2010.

[12]. S. Kumar, D. K. Kumar, M. Alemu, and M. Burry, “Emg based voice recognition,” in *Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference*, pp. 593–597, IEEE, Melbourne, VIC, Australia, December 2004.

[13]. S. P. Arjunan, D. K. Kumar, W. C. Yau, and H. Weghorn, “Unspoken vowel recognition using facial electromyogram,” in *International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2191–2194, IEEE, New York City, USA, September 2006.

[14]. S. P. Arjunan, H. Weghorn, D. K. Kumar, and W. C. Yau, “Vowel recognition of english and german language using facial movement (semg) for speech control based hci,” in *Proceedings of the HCSNet workshop on Use of vision in human-computer interaction*, Volume 56, pp. 13–18, Canberra, Australia, November 2006.

[15]. S. Mostafa, M. Awal, M. Ahmad, and M. Rashid, “Voiceless bangla vowel recognition using semg signal,” *SpringerPlus*, vol. 5, pp. 1–15, September 2016.

[16]. U. Agnihotri, A. S. Arora, and A. Garg, “Vowel recognition using facial movement (semg) for speech control based hci,” *International Journal of Engineering Research & Technology ACMEE*, vol. 4, pp. 1–5, April 2016.

[17]. R. Takabatake, S.-i. Ito, M. Ito, and M. Fukumi, “Vowels recognition using emg measured with dry type sensors,” in *The International Conference on Electrical Engineering*, no. 90049, pp. 1–5, Okinawa, Japan, July 2016.

[18]. R. Takabatake, S.-i. Ito, M. Ito, and M. Fukumi, “Japanese vowels recognition using linear discriminant analysis and surface electromyogram measured with bipolar dry type sensors,” in *Proceedings of 5th IIAE International Conference on Intelligent Systems and Image Processing*, no. 1468, pp. 5–11, Waikiki, Hawaii, USA, September 2017.

[19]. L. Fraiwan, K. Lweesy, A. Al-Nemrawi, S. Addabass, and R. Saifan, “Voiceless arabic vowels recognition using facial emg,” *Medical & Biological Engineering & Computing*, vol. 49, pp. 811–818, March 2011.

[20]. H. Manabe, A. Hiraiwa, and T. Sugimura, “Unvoiced speech recognition using emg-mime speech recognition,” in *CHI’03 extended abstracts on Human Factors in Computing Systems*, pp. 794–795, Ft. Lauderdale, Florida, USA, April 2003.

[21]. G. R. Naik, D. K. Kumar, and S. P. Arjunan, “Reliability of facial muscle activity to identify vowel utterance,” in *TENCON 2008-2008 IEEE Region 10 Conference*, pp. 1–6, IEEE, Hyderabad, India, November 2008.

[22]. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, pp. 1735–1780, November 1997.

[23]. A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” *Neural networks*, vol. 18, pp. 602–610, July 2005.

[24]. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, May 2015.

[25]. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, November 1998.