Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

StrokeVoice: A Low-Cost Hybrid Multimodal AAC Ecosystem for Post-Stroke Aphasia and Dysarthria

Phi Van Lam*, Nguyen Cong Hai Dang †, Nguyen Thi Hong Hoa*, Tran Thi Lan *
* University of Transport and Communications, Hanoi, Vietnam
† Archimedes Dong Anh School, Hanoi, Vietnam

***Abstract*—We present StrokeVoice, a low-cost, hybrid multimodal Augmentative and Alternative Communication (AAC) ecosystem designed as a Silent Speech Interface (SSI) for patients with post-stroke aphasia and dysarthria. The system integrates a throat vibration sensor with an acoustic amplification module to capture subtle, non-audible articulatory movements during silent speech. The acquired signals are digitized and transmitted via Bluetooth Low Energy (BLE) to a mobile application, where a lightweight deep learning model performs real-time classification. To enhance communicative utility, we propose a conceptual framework utilizing a Large Language Model (LLM) to reconstruct coherent, full sentences from the discrete, recognized keywords. Experimental evaluations across five subjects demonstrate that StrokeVoice restores intelligible speech states with a system latency of less than 1 second, establishing its viability as a portable, affordable communication aid in both quiet and noisy environments.**

***Index Terms*—Silent Speech Interface, throat vibration sensing, acoustic amplification, TensorFlow Lite, Large Language Model, speech reconstruction.**

Abbreviation	ion
AAC	Augmentative and Alternative Communication
SSI	Silent Speech Interface
BLE	Bluetooth Low Energy
LLM	Large Language Model
ADC	Analog-to-Digital Converter
AI	Artificial Intelligence
FNN	Feedforward Neural Network

I. INTRODUCTION

The rapid expansion of deep learning and intelligent sensing technologies has catalyzed major breakthroughs in pattern recognition and signal processing across diverse fields. In industrial inspection, advanced neural networks have been applied to classify defective components in electric vehicle manufacturing [1] and detect printed circuit board (PCB) flaws via specialized architectures like SEConv-YOLO [2]. Concurrently, embedded smart systems have enhanced public transit infrastructure, as exemplified by IoT-enabled token counting machines [3].

Beyond industrial automation, machine learning and intelligent sensing have increasingly supported human-centric tasks. Researchers have developed image-processing controls for robotic canes to aid visually impaired individuals in maintaining balance [4]. In transportation safety, machine learning models have been deployed for real-time traffic signal recognition and driver warning systems [6]. Deep learning has also demonstrated notable resilience in adverse environments, ranging from YOLOv8-based transmission line fault detection under severe weather conditions [5] to robust object detection in complex power grid infrastructures [7].

Moreover, the feasibility of deploying deep learning algorithms directly on resource-constrained, low-power embedded platforms has been validated by real-time multi-object recognition frameworks for automated robots [8].

Despite these extensive advancements in sensing and automation, assistive communication support for individuals who have lost their speaking capability due to stroke or neurological disorders remains a critical challenge. Conventional voice-based communication devices rely entirely on acoustic signals, rendering them ineffective for users with severe speech motor impairments (such as dysarthria or aphasia) or in high-noise environments.

To overcome these constraints, Silent Speech Interfaces (SSI) have emerged as a compelling alternative. SSI systems reconstruct spoken language from non-acoustic articulatory biosignals, such as vocal tract vibrations or facial muscle activity, bypassing the need for audible voice. Traditional SSI designs often rely on electromyography (EMG) or ultrasound imaging; however, these methods typically demand complex hardware, incur prohibitive costs, and present significant challenges for everyday portable use. Furthermore, decoding raw, noisy biosignals into coherent linguistic concepts remains computationally demanding.

Motivated by these challenges, this paper introduces StrokeVoice, a low-cost, hybrid multimodal Augmentative and Alternative Communication (AAC) ecosystem. StrokeVoice leverages a lightweight throat vibration sensor and an acoustic amplification module to capture non-audible speech articulation. The acquired features are classified in real time using an optimized Feedforward Neural Network (FNN) deployed on a mobile platform, which is conceptually designed to interface with a Large Language Model (LLM) for semantic sentence reconstruction. By combining affordable hardware with localized machine learning, the proposed system aims to provide an accessible, low-latency, and highly deployable assistive communication solution for post-stroke rehabilitation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

II. MAIN CONTENT

A. System Architecture Overview

The operation of the proposed intelligent Silent Speech Interface (SSI) system follows a closed-loop processing pipeline that integrates biosignal acquisition, embedded signal processing, and wireless communication with a mobile platform. The overall workflow of the system is illustrated in Fig. 1.

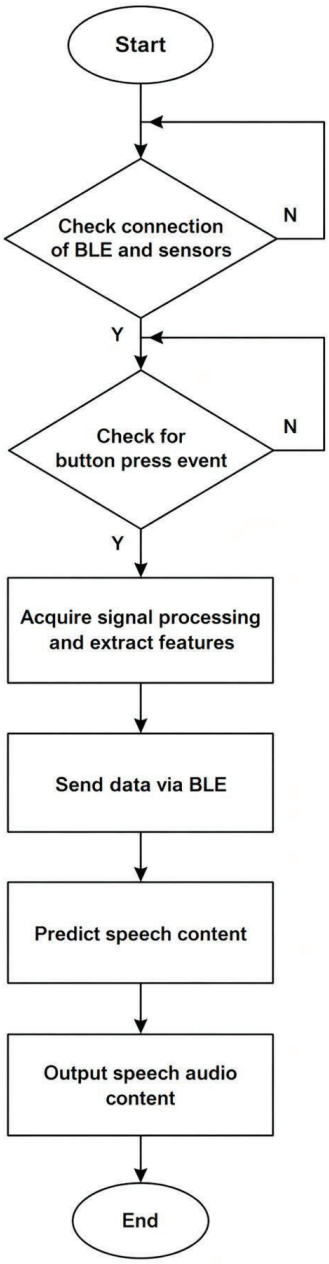


Fig. 1. Overall workflow of the proposed system.

The process begins when the system is powered on and establishes a connection with a mobile device via a wireless communication protocol. Once the connection is successfully initialized, the system enters a standby mode and continuously monitors for a user-triggered event. When the user activates the system by pressing a button, it switches to the signal acquisition mode. At this stage, the sensors simultaneously capture throat vibration signals and auxiliary acoustic signals generated during silent speech.

After acquisition, the analog signals are converted into digital form using an integrated Analog-to-Digital Converter (ADC). The digitized signals are then processed in time windows with an appropriate sampling frequency to preserve essential characteristics of the silent articulation process.

During the signal processing stage, the system extracts representative features from the acquired signals, including amplitude, energy, temporal variation, and other relevant parameters. These features are accumulated and averaged over the duration of the silent speech activity to form a feature vector that represents the spoken content.

When the user releases the button, indicating the end of the recording phase, the system transitions to the data transmission stage. The extracted feature vector is packaged and transmitted to the mobile device via Bluetooth Low Energy (BLE), ensuring real-time communication with low power consumption.

On the mobile platform, the received data is processed by a Feedforward Neural Network (FNN) model to recognize the speech content. To further enhance the semantic coherence and completeness of the output, a language model can be conceptually integrated to refine the predicted results.

Finally, the recognized content is displayed and converted into synthesized speech for audio output. The system then returns to the standby state, ready for the next operation cycle. This iterative processing mechanism ensures continuous, stable, and low-latency performance, making the proposed SSI system suitable for real-time assistive communication applications, particularly for users with speech impairments.

B. Hardware Architecture and Signal Acquisition

The system (Fig. 2) is designed to acquire and process the user's silent speech signals using biosensors attached to the throat area. Unlike traditional image processing systems, the proposed system utilizes mechanical vibration signals and low-intensity acoustic signals for non-audible speech recognition.

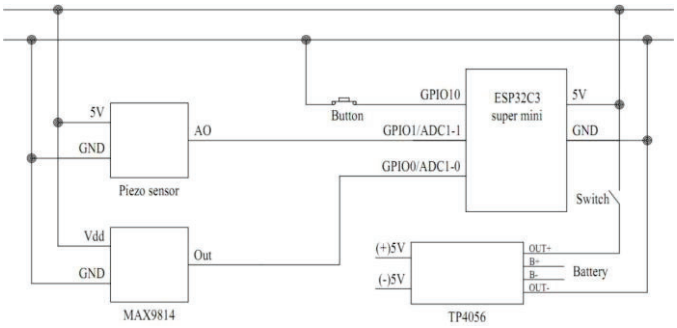


Fig. 2. Overall hardware architecture design of the system.

The signal acquisition block consists of two main components: a throat vibration sensor and an acoustic sensor.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

The vibration sensor is responsible for capturing mechanical vibrations generated when the user performs silent articulation movements, while the acoustic sensor acts as an auxiliary component to provide additional information on the energy and spectral characteristics of the signal.

Signals from the sensors are fed into a microcontroller integrated with an Analog-to-Digital Converter (ADC) for digitization. To ensure the quality of the input signals, the system employs signal amplification circuitry to increase the amplitude and improve the signal-to-noise ratio (SNR).

The signal acquisition process is performed over time windows with a suitable sampling rate, preserving the crucial features of the silent articulation process. Once collected, the data is passed to the processing block for feature extraction.

Fig. 3 illustrates the physical model of the system after fabrication and assembly. The system is designed as a lightweight wearable device, allowing users to employ it in daily communication scenarios.

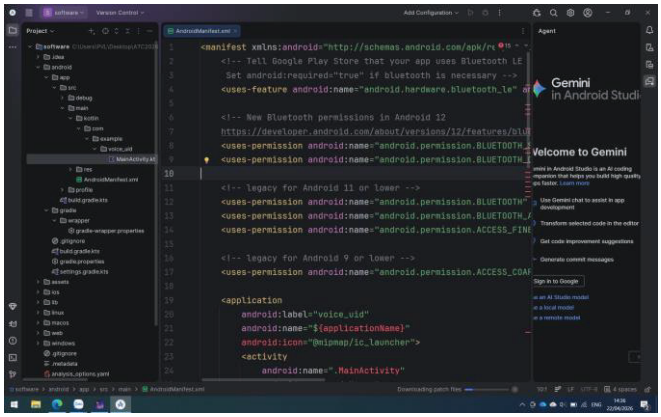


Fig. 3. Physical model of the wearable system.

At the hardware level, the microcontroller serves as the core processing unit for signal acquisition, processing, and data transmission. The sensors are securely positioned on the throat to ensure signal stability during speech. This configuration minimizes the impact of motion artifacts and mechanical noise, thereby enhancing the signal quality.

After preliminary processing, the data is wirelessly transmitted to the mobile device via the Bluetooth Low Energy (BLE) protocol. This approach provides high flexibility, low power consumption, and excellent suitability for wearable devices.

The physical model demonstrates the integration between the embedded hardware and the signal processing system, establishing the foundation for real-time speech recognition.

C. Dataset Construction

To build the dataset for training the silent speech recognition model, the system uses a throat vibration sensor (piezo sensor) combined with an acoustic sensor. Data is collected via an ESP32 microcontroller with a sampling rate of approximately 8 kHz. Each data sample is recorded over a window of 64 consecutive measurement points. From the acquired signals, the system extracts the following time-domain features:

- Root Mean Square (RMS) of the throat vibration signal (Piezo RMS);
- Peak amplitude of the vibration signal (Piezo Peak);
- RMS of the acoustic signal (Mic RMS);
- Zero-Crossing Rate of the acoustic signal (Mic ZCR);
- Log energy of the acoustic signal (Mic Energy);
- Ratio between the vibration and acoustic signals (Ratio).

These features are computed directly on the ESP32 and transmitted to the computer via UART communication. Each data record consists of 6 features and its corresponding class label.

During the collection process, participants silently articulate each target word multiple times. When the user presses the button on the device, the system starts recording data and stops when the button is released. The data samples are saved as CSV files for the model training process.

After collection, the data from all classes is cleaned by removing invalid values, converting to numerical formats, and normalizing the data structure. To prevent class imbalance, the number of samples across all classes is balanced using the undersampling method, where all classes are reduced to have the same number of samples as the minority class. The final dataset is randomly shuffled before being fed into the model training process.

D. Classification Model Architecture

Once the dataset is constructed, the feature vectors are used to train the silent speech recognition model. The goal of the model is to classify the articulation signal into classes corresponding to predefined keywords.

Before feeding the data into the deep learning model, the features are normalized using the StandardScaler method to bring them to a common scale. This normalization improves the stability of the training process, accelerates convergence, and limits the dominance of features with larger magnitudes over others.

The classification model is designed as a Feedforward Neural Network (FNN), consisting of two hidden layers with 32 and 16 neurons respectively, using the ReLU activation function. The output layer consists of 3 neurons using the Softmax activation function to output the probability scores for each class.

The training process utilizes the Adam optimizer and the Sparse Categorical Crossentropy loss function. The dataset is split into training and testing sets with an 80:20 ratio to evaluate the model's generalization capabilities. The detailed model parameters are summarized in Table I.

E. Signal Preprocessing and Feature Extraction

After signal acquisition and digitization, the system performs a sequence of processing steps to enhance data quality and extract meaningful features for the recognition task. The overall signal processing and feature generation pipeline is illustrated in Fig. 6.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

piezo_rms	piezo_peak	mic_rms	mic_zcr	mic_energy	ratio	label
1995.6	2024	742.99	0.109	2.87	2.686	0
1974.74	1990	1171.48	0.094	3.07	1.686	2
1975	1986	741.48	0.094	2.87	2.664	1
1982.59	1991	443.16	0.094	2.65	4.474	1
1977.03	1991	776.51	0.062	2.89	2.546	2
1975.21	1987	634.29	0.062	2.8	3.114	1
1978.34	1989	1365.19	0.078	3.14	1.449	0
1988.04	1993	733.43	0.031	2.87	2.711	1
1995.94	2005	415.2	0	2.62	4.807	1
2002.68	2010	1255.43	0.062	3.1	1.595	1
1973.62	1989	1070.11	0.109	3.03	1.844	2
1969.12	1993	641.78	0.078	2.81	3.068	1
1992.46	2008	517.43	0	2.71	3.851	1
1982.18	1990	689.13	0.016	2.84	2.876	1
1978.91	1991	885.16	0.078	2.95	2.236	2
1976.92	1991	394.91	0	2.6	5.006	2
1982.69	1990	535.89	0.047	2.73	3.7	1
1990.89	2004	402.49	0.094	2.6	4.946	1
1969.77	1989	850.37	0.047	2.93	2.316	2
1973.54	1990	984.95	0.141	2.99	2.004	2
1979.05	1993	962.17	0.141	2.98	2.057	2
1986.66	2004	540.31	0.234	2.73	3.677	0
1981.79	1990	442.85	0.094	2.65	4.475	0
1976.56	2009	704.93	0.234	2.85	2.804	0
1979.01	1992	569.78	0.172	2.76	3.473	0
1973.2	1989	928.42	0.172	2.97	2.125	2
1977.53	1990	357.15	0	2.55	5.537	2
1955.11	1968	1065.69	0.031	3.03	1.835	1
1976.6	1991	1026.97	0.172	3.01	1.925	2
1973.14	1986	499.04	0.094	2.7	3.954	1
1977.98	1990	682.53	0.141	2.83	2.898	1
1976.23	2002	388.74	0.125	2.59	5.084	1

Fig. 4. Constructed dataset for the silent speech recognition model.

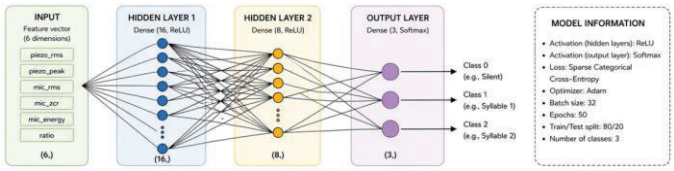


Fig. 5. Deep learning neural network architecture for the classification model.

TABLE I
MODEL PARAMETERS AND SPECIFICATIONS

Parameter	Value
Number of input features	6
Hidden Layer 1	Dense(32), ReLU
Hidden Layer 2	Dense(16), ReLU
Output Layer	Dense(3), Softmax
Optimizer	Adam
Loss Function	Sparse Categorical Crossentropy
Batch Size	16
Epochs	50
Train/Test Ratio	80% / 20%
Data Normalization	StandardScaler
Deployment Format	TensorFlow Lite (.tflite)

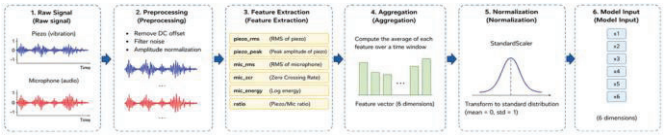


Fig. 6. Signal preprocessing and feature vector generation pipeline.

1) Signal Preprocessing: In the first stage, the raw signals are preprocessed to remove unwanted components and improve signal stability. The DC offset is eliminated by centering the signal around zero mean. In addition, random noise is reduced by applying averaging within each time window. This step enhances the robustness of the input data and minimizes the impact of environmental noise.

2) Feature Extraction: From the preprocessed signals, a set of time-domain features is computed to characterize the silent speech patterns. The selected features include:

- **Root Mean Square (RMS):** represents the overall energy of the signal.
- **Peak Amplitude:** indicates the maximum signal magnitude.
- **Zero-Crossing Rate (ZCR):** reflects the rate of signal variation.
- **Log Energy:** describes the logarithmic energy level of the signal.
- **Signal Ratio:** the ratio between vibration and acoustic signals.

These features are chosen due to their effectiveness in capturing essential characteristics of silent speech while maintaining low computational complexity, making them suitable for embedded systems.

3) Feature Aggregation: The extracted features are computed continuously over sliding windows and then aggregated over the entire duration of the speech activity. Specifically, the features are accumulated and averaged to form a representative feature vector. This approach provides several advantages:

- Reduction of instantaneous noise effects;
- Generation of stable feature representations;
- Improvement in classification accuracy.

4) Data Normalization: Before being fed into the deep learning model, the feature vectors are normalized using the StandardScaler method to ensure that all features share a common scale. This normalization step improves training efficiency and prevents certain features from dominating the learning process.

F. Software Design and Mobile Processing

The control and monitoring software is developed as a mobile application, serving as the central unit for data reception, processing, and visualization of recognition results. The application is built using the Flutter framework with the Dart programming language, enabling cross-platform deployment and ensuring real-time processing performance.

The software architecture follows an event-driven paradigm combined with the Model–View–ViewModel (MVVM) design

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

pattern. In this architecture, events generated from the BLE device, deep learning model, and user interactions trigger corresponding processing tasks. This approach ensures fast responsiveness and facilitates system scalability. The software architecture is illustrated in Fig. 7.

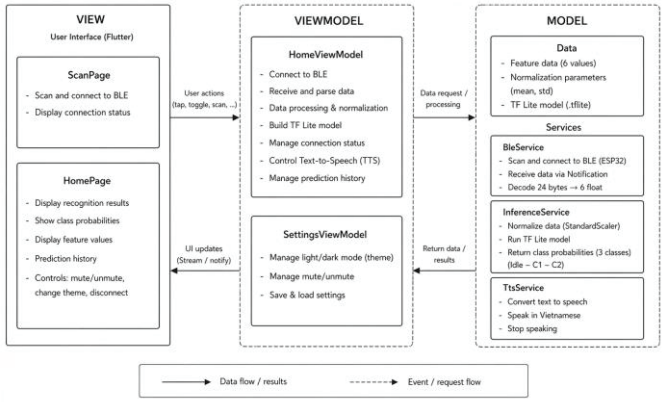


Fig. 7. Software architecture of the mobile application.

The application consists of four main functional modules:

1) BLE Communication and Data Acquisition: The application scans for and connects to the embedded device via Bluetooth Low Energy (BLE). Once the connection is established, feature data is transmitted from the device in the form of a real-valued vector consisting of six elements. The received data is decoded and converted into numerical form for further processing. A notification-based transmission mechanism is employed to ensure low latency and suitability for real-time applications.

2) Deep Learning Inference Module: After data reception, the input features are normalized using pre-trained parameters (StandardScaler). The normalized data is then fed into a TensorFlow Lite model for classification. The model outputs probability scores for each class, and the class with the highest probability is selected as the recognition result. Additionally, the probability distribution is displayed to provide insight into the confidence level of the prediction.

3) Text-to-Speech (TTS) Module: To enable natural communication, the recognition result is converted into speech using a Text-to-Speech (TTS) module. The system utilizes a TTS library to generate Vietnamese speech corresponding to the predicted content. The speech output is triggered only when the prediction confidence exceeds a predefined threshold and the system is not in a silence state. This mechanism helps prevent incorrect audio output under noisy conditions.

4) User Interface and Visualization: The user interface is designed to be intuitive and minimalistic, including the following main components:

- BLE connection status display;
- Recognition result display (e.g., Yes / No / Silence);
- Probability bar for each class;
- Input feature values display;
- History of recent predictions.

In addition, the system provides several user control functions:

- Audio control (mute/unmute);
- Light/dark mode switching;
- Device disconnection.

The interface is updated in real time based on incoming data from the embedded device and inference results from the model, allowing users to effectively monitor and control the system.

III. Experimental Results and Discussion

The performance of the proposed Silent Speech Interface (SSI) system is evaluated in terms of classification accuracy, system stability, and real-time responsiveness. The deep learning model is trained on feature data collected from the throat vibration sensor and auxiliary acoustic signals.

A. Model Training Performance

The training results indicate that the model achieves an accuracy of approximately 82% on the test set. During the initial training phase, the accuracy increases rapidly from around 35% to over 70%, as the model learns fundamental signal characteristics. After approximately 20–30 epochs, the accuracy gradually converges and stabilizes in the range of 80–83%, indicating effective learning without overfitting.

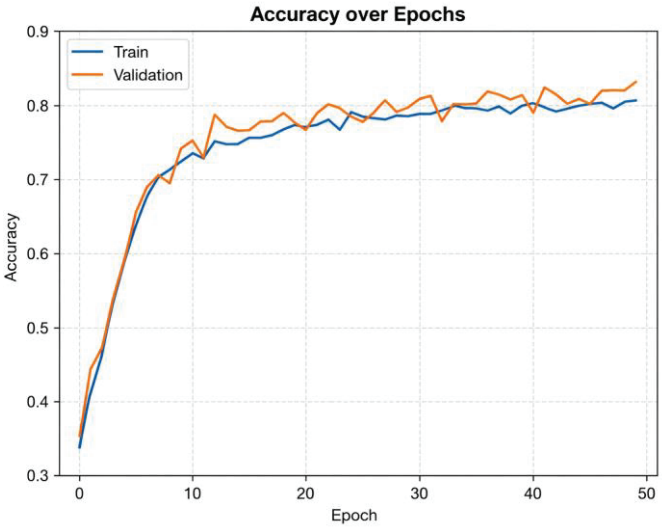


Fig. 8. Model accuracy over training epochs.

Fig. 8 illustrates the evolution of accuracy during training. Both training and validation accuracies increase rapidly in the early stages and later stabilize with minimal divergence. This behavior demonstrates good generalization capability and confirms that overfitting does not occur.

Fig. 9 shows the variation of the loss function during training. The loss decreases sharply in the early epochs and then gradually converges to a stable value of approximately 0.48–0.50. This indicates that the optimization process is effective and the model reaches a convergence state.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

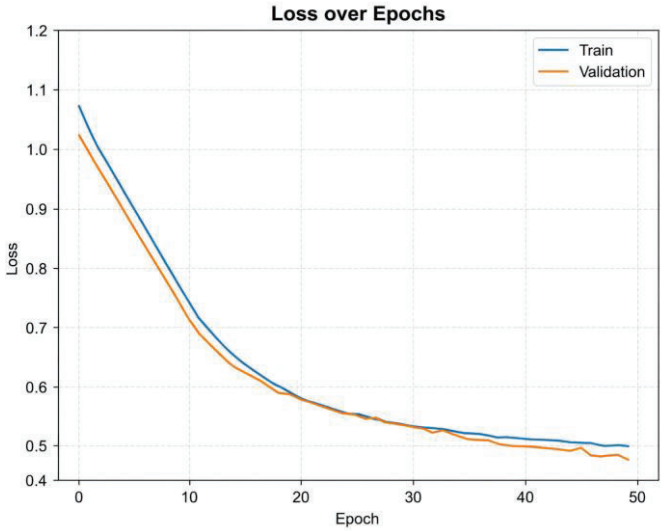


Fig. 9. Loss function over training epochs.

B. System-Level Evaluation

In the real-world deployment, feature data is acquired from the embedded device and transmitted to the mobile application via BLE. The TensorFlow Lite model performs inference directly on the mobile device, ensuring low-latency processing. The system is capable of accurately recognizing three states:

- Yes
- No
- Silent

The prediction results are displayed in real time on the mobile interface, along with the corresponding probability values for each class. Fig. 10 presents the user interface displaying the recognition results. The predicted label is clearly shown along with its confidence level, allowing users to easily interpret the output.

To evaluate the system's robustness, real-world experiments are conducted on five different subjects. The physical characteristics and demographic profiles of the participants are summarized in Table II.

TABLE II
SUBJECT PROFILES FOR SYSTEM EVALUATION

Subject	Age	Gender	Height	Weight
Subject 1	23	Male	167 cm	60 kg
Subject 2	27	Male	170 cm	63 kg
Subject 3	33	Female	157 cm	55 kg
Subject 4	23	Male	170 cm	57 kg
Subject 5	23	Female	158 cm	48 kg

The testing setups and actual experimental trials performed with the subjects are shown in Fig. 11.

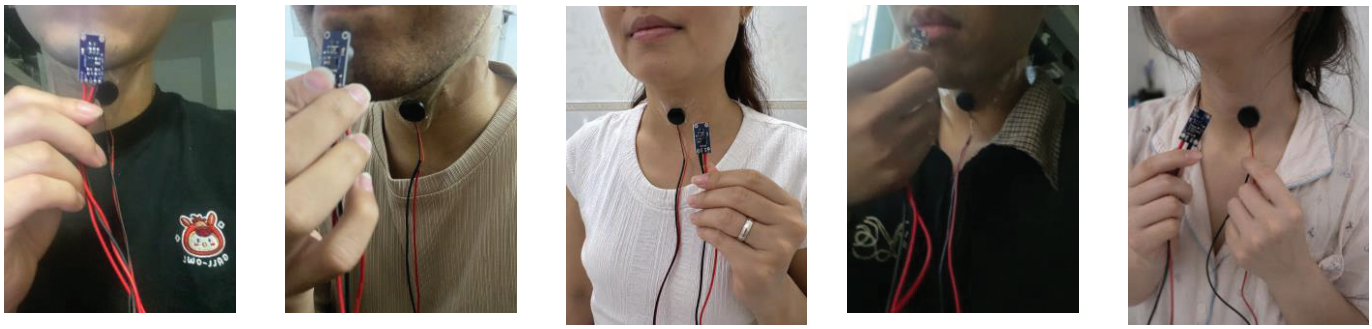
Each subject performs 30 silent articulation trials. The classification performance of the system for each subject across the "Yes", "No", and "Silent" classes is detailed in Table III.

Fig. 10. Recognition results displayed on the mobile application.

TABLE III
SYSTEM RECOGNITION ACCURACY ACROSS DIFFERENT SUBJECTS

Subject	Total Trials	Correct ("Yes")	Correct ("No")	Silence ("Silent")
Subject 1	30	9/10	7/10	9/10
Subject 2	30	9/10	5/10	8/10
Subject 3	30	8/10	6/10	8/10
Subject 4	30	9/10	7/10	8/10
Subject 5	30	7/10	5/10	7/10

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026



(a) Subject 1 (b) Subjec (c) Subject 3 (d) Subject 4 (e) Subject 5

Fig. 11. Experimental testing trials conducted on different subjects.

The testing results across the five subjects indicate that the system exhibits a relatively stable recognition performance, achieving higher accuracy for the "Yes" and "Silent" states. In contrast, the "No" state exhibits slightly lower accuracy, suggesting some degree of acoustic/vibrational similarity and minor classification confusion under weak articulation. Nonetheless, the overall system maintains acceptable and stable performance across all users, confirming its feasibility and potential for practical applications.

C. Confusion Matrix Analysis

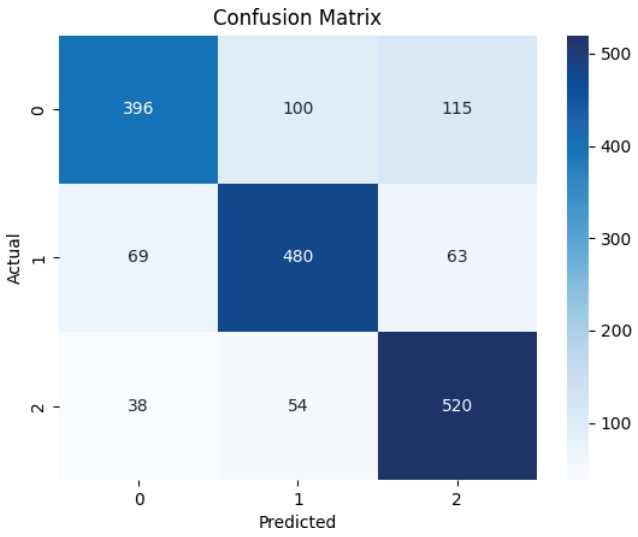


Fig. 12. Confusion matrix of the classification model.

The confusion matrix in Fig. 12 indicates that the classification model performs well across all three classes, with correct predictions concentrated along the main diagonal. Specifically, Class 2 achieves the highest performance with 520 correctly classified samples, followed by Class 1 with 480 samples, and Class 0 with 396 samples. However, Class 0 still exhibits a noticeable misclassification rate, with 100 samples misclassified as Class 1 and 115 samples as Class 2. This suggests that the features of Class 0 share certain similarities with the other two classes, or that the training dataset for this class lacks sufficient diversity.

On the other hand, the misclassification between Class 1 and Class 2 is relatively low, indicating that the model has successfully learned distinctive features to differentiate between these two states. These results confirm the stable recognition capabilities of the model. Nonetheless, expanding the training dataset for Class 0 or refining the feature extraction process could further enhance the overall system accuracy.

D. Latency and Real-Time Performance

The system is designed for real-time operation. The total response time includes:

- Signal acquisition;
- BLE data transmission;
- Model inference;
- Audio output generation.

Experimental results show that the overall latency is less than 1 second, which satisfies the requirement for natural communication. This confirms the feasibility of deploying the proposed SSI system in real-time assistive applications.

IV. CONCLUSION

This paper has presented the design and implementation of an intelligent Silent Speech Interface (SSI) system based on a throat vibration sensor and lightweight deep learning models. The proposed system integrates an embedded signal acquisition module, a wireless data transmission mechanism, and a mobile application for real-time inference and speech synthesis.

The deep learning model, deployed using TensorFlow Lite, enables efficient on-device inference with low latency. Experimental results demonstrate that the system achieves reliable classification accuracy for silent speech signals while maintaining stable real-time performance.

The proposed solution shows strong potential for assistive communication applications, particularly for individuals with speech impairments, as well as in environments where audible communication is not feasible.

In future work, the system can be further improved by expanding the training dataset, optimizing the model architec-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

ture, and incorporating context-aware processing techniques to enhance the quality and naturalness of speech reconstruction.

REFERENCES

[1] Phi Van Lam and Tran Thi Lan, "Deep learning application to classification of products in electric car accessories production", *Transport Magazine*, special number, pp. 18–22, 2023.

[2] S. K. Ong, C. K. Tan, V. M. Baskaran, B. K. Puah and K. H. Liew, "Enhancing Industrial PCB and PCBA Defect Detection: An Efficient and Accurate SEConv-YOLO Approach", *IEEE Access*, vol. 13, pp. 148917–148935, 2025, doi: 10.1109/ACCESS.2025.3601151.

[3] Phi Van Lam, Tran Thi Lan, Trinh Luong Mien, "Design and Implementation of an IoT-Enabled Automatic Token Counting Machine for Metro Systems", Engineering, Technology & Applied Science Research 16(2), 34337-34350, 28 February 2026. DOI: https://doi.org/10.48084/etasr.17737.

[4] Phi Van Lam and Yasutaka Fujimoto, "Image Processing Application in Controlling the Robotic Cane to Support the Blind Maintain Balance When Moving", in *Proc. 5th Int. Conf. on Control and Automation (VCC-A)*, 2019.

[5] H. Liu, X. Jia and J. Dai, "RFEA-YOLO: An Enhanced YOLOv8 for Transmission Line Fault Detection in Rain and Fog", in *Proc. 6th Int. Conf. on Smart Grid and Energy Engineering (SGEE)*, Shanghai, China, 2025, pp. 422–427, doi: 10.1109/SGEE68429.2025.11385661.

[6] Phi Van Lam, Tran Thi Lan and Ngo Minh Quy, "Hệ thống nhận diện tín hiệu đèn giao thông và cảnh báo cho phương tiện phía sau dựa trên mô hình học máy", in *Proc. Conf. on New Technologies and Applications in Electrical, Electronics and Automation*, University of Transport and Communications, Hanoi, Vietnam, May 24, 2025, ISBN: 978-604-76-3106-3.

[7] J. Tian, Y. Tian, B. Liu, Y. Liu and Z. Liang, "A Detection Algorithm for Distribution Network Insulators Based on YOLOv8n-EdgePro", in *Proc. IEEE Int. Conf. on Industrial Technology (ICIT)*, Wuhan, China, 2025, pp. 1–6, doi: 10.1109/ICIT63637.2025.10965212.

[8] Phi Van Lam, Vu Duc Du, Tran Thi Lan, "Real-time multi-object recognition and classification on resource-constrained devices for automated robots", Journal of Measurement, Control and Automation, Vol. 30 No. 2 (2026): May 2026, ISSN 3030-4555. DOI: https://doi.org/10.64032/mca.v30i2