International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Deep Learning Framework for Lip Motion-Tospeech Translation in Robots

Shruthi M V
Department of Electronics and Communication
Sri Siddhartha School of Engineering
Tumkur, India

Soumya S
Department of Electronics and Communication
Sri Siddhartha School of Engineering
Tumkur, India

Sushma B N
Department of Electronics and Communication
Sri Siddhartha School of Engineering
Tumkur, India

Abstract **- Speech is the most natural and fastest way of communication for humans. Oral communication plays an important role in social life. Human speech starts in the brain and ends in the mouth. The lungs, vocal cords, tongue are also very important, as are lips, throat and jaw. The person with speech impairment caused by partial or complete removal of larynx face difficulty in communication. Putting it simply, the AI-Driven Lip Motion Translator for Robotic Communication is a new system that allows robots to understand and react to human lip movement without speech. The system records real-time video of a person's lip movements using a camera and processes it with artificial intelligence (AI) to recognize the spoken words. Recognized speech is then translated to another language or converted to audio output enabling a smooth communication between humans and robots.**

This paper also has communication modules such as GSM, RF transmitters for wireless data transfer and robotic output units like servo motors, relays and drivers to perform synchronized actions. The system synchronizes speech with realistic lip and facial movements, making robotic communication more natural, expressive and human-like. The proposed model can be applied in assistive technology, multilingual communication, healthcare robots, and service robotics.

Keywords – Deep learning, Lip-to-Speech Translation, Human–Robot Interaction, Multimodal Learning

I. INTRODUCTION

The important part of communication is speech. Speech is natural way for communication for a person to express his feelings or to tell anything. Human speech perception is, as is well known, a multimodal process. Effective and fluent conversation requires a joint effort from both interlocutors. In spoken human dialogue, this effort often takes place in real time as speech is occurring. Watching the movements of lips visually gives important information about the place of pronunciation of articulation. This is often referred to as lip-reading which is to make sense of what someone is saying by watching the movement of his/her lips.

For listeners who are suffering from hearing impairment or in noisy environments it is important to provide information visually. Most of the studies are only concerned with visual information to improve speech recognition. Audio features remain the main contribution and more important role than visual features .

In some of the cases, the environment is noisy and it is very difficult to extract the useful features from the audio. Some applications need speech recognition in adverse acoustic environments. For instance, it is very hard to understand the speech of person where there is noisy crowd of people like detecting a speech of person from a distance or through glass of window.

A Silent Speech Interface has been proposed for speech processing in the absence of intelligible acoustic signal, and it has been used as an aid for the dumb person or speech handicapped. A Silent Speech Interface (SSI) is a device which enables the communication to be present when voice is not. In noisy situations, the SSI is a promising way to process speech. Lip movements provide valuable visual cues that help in understanding spoken words, especially when audio signals are unclear. In noisy environments, relying only on speech signals can reduce recognition accuracy, making visual information important. Silent Speech Interfaces use lip movements and facial features to recognize speech even when no audible voice is present.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

II. LITERATURE SURVEY

A. Hunt, W., Soorati, M., et al (2024) A Survey Of Language-Based Communication.

This survey focuses on the role of language-based communication within human-robot teams. While it may not specifically detail lip motion translation, its abstract discusses the integration of Natural Language Processing (NLP) and Large Language Models (LLMs) to improve human-robot interaction and communication efficiency. It addresses how robots can interpret and act upon human instructions, providing context on the communication aspects of the user's topic.

B. Shwetha K S, Rohith M K, Sakshi Prashant Yandagoudar, Sinchana (2024) Silent Interpreter: Analysis Of Lip Movement And Extracting Speech Using Deep Learning

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

C. N. Rathipriya , N. Maheswari (2024) A Comprehensive Review Of Recent Advances In Deep Learning For Automated Lip Reading

This article provides an up-to-date review of the rapid development in automated lip-reading (ALR) technology using deep learning. It covers recent approaches and models from the early 2020s to 2023. The review highlights the key advancements, challenges, and future directions in using AI to accurately interpret speech from lip movements, which is directly relevant to developing robust "lip motion translators".

D. Wang, H., et al (2020) A Survey Of Research of Lipreading Technology

This paper provides a comprehensive survey on the development of lip-reading technology. It discusses various methods used for visual feature extraction from lip movements, such as Discrete Cosine Transform (DCT), and different classification models including Hidden Markov Models (HMM), Support Vector Machines (SVM), and Deep Neural Networks (DNNs). The abstract emphasizes the application of these techniques to recognize speech from visual information, which is a foundational element for any lip motion translator.

III. SYSTEM DETAILS

This system consists of three stages, as shown in Figure 1. The first stage is capturing of lip movements using webcam. The second stage is extraction of the visual features from the lip movement sequence. The role of the final stage is to recognize the input utterance using a k-NN classifier rate until after the text has been formatted and styled.

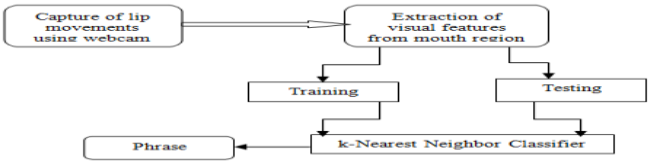


Figure 1: System Proposed

A. System Description

The webcam is used to capture the lip movements by the system. For one specific lip movement(word) twenty frames are generated and features are extracted and histograms are obtained for each frame. Histograms are averaged and trained for classifier . Here for speech recognition k-Nearest Neighbor classifier are used.

Average of histograms that are obtained in real time are fed to the classifier and it is tested with the stored histograms that has been trained earlier. If the histograms that are obtained in real time matches with the stored histogram than the word that has been uttered by the person will be displayed in the form text. Text to speech system is used to convert the text into speech.

B. Visual Feature Extraction

Visual feature extraction is facilitated using the resulting landmark locations from the AAM tracking. Center of the mouth region is found by determining the centroid of the corresponding mouth points. A region around the mouth is extracted using the scale and rotation information from the AAM global transform. The mouth region is rotated around the center of the mouth to align with the mean face shape, which has a horizontally aligned mouth. The extracted ROI is a square whose side length is chosen such that the ratio of the global transform scale to the ROI side length is equivalent to extracting a ROI with side length of 40 pixels in the mean face shape (scale = 1.0).

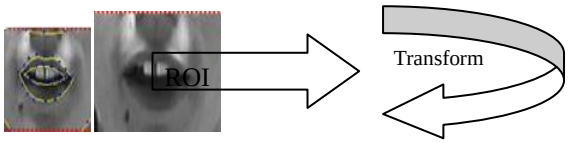


Figure 2: Illustration of the visual feature extraction process

Given the AAM tracking data and the scale and rotation normalized ROI, the next step is to extract visual features. Visual features can be either geometric based or appearance based. Geometric based features are represented by physical distances or shape parameters. Some common examples of geometric features are height and width of the mouth, landmark locations, or fitted curve parameters. Appearance based visual features attempt to capture the ROI as a whole. In an extreme case, all pixel intensities in the ROI could be used as visual features, however, the 'curse of dimensionality' and visual noise make this an infeasible and poor feature choice.

Appearance based features greatly reduce the dimensionality of the data by compacting the transform energy within the top few transformation coefficients. Collecting these feature vectors over a number of frames (t=1,2,...,n) leads to the final

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

visual observation, Ov = [Ov 1 , Ov 2 , Ov 3 , ...Ov n]. Isolated recognition is manually triggered in this system.

IV. METHODOLOGY

A. *Data Collection*

Capture video datasets of human speakers showing clear lip and facial movements. Record corresponding audio and translated speech in multiple languages.

B. *Pre-Processing*

Extract facial landmarks and lip-region frames using computer vision techniques. Normalize and align video and audio data for consistent training.

C. *Feature Extraction*

Use Convolution Neural Networks (CNNs) or Vision Transformers (ViT) to analyze lip movements. • Apply Audio and Speech Recognition models (ASR/TTS) for voice and language processing.

D. *AI Model Training*

Train a Lip-Reading Model to decode speech content from lip movements. Integrate a Neural Machine Translation (NMT) model to convert decoded speech into target languages. Synchronize translated speech with lip and facial motion using Generative Adversarial Networks (GANs) or Lip-Sync Models (e.g., Wav2Lip).

E. *Integration With Robot Interface*

Implement the trained model into a robot's communication system.Use real-time camera and speaker modules for live lip-motion decoding and speech playback.

F. *Output Generation*

Robot reproduces synchronized lip movements with translated speech. Enables natural, multilingual, and expressive robotic communication.

V. BLOCK DIAGRAM

The block diagram of the paper "Conversion of lip movement to speech: An aid to physically impaired and dumb people" is as shown in the Figure 3

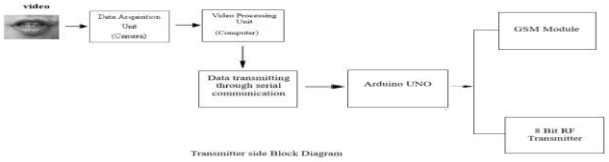


Figure 3: Block diagram of Transmitter.

The AI-Driven Lip Motion Translator system transmitter side is designed to capture a person's lip motions and convert them into meaningful data and then wirelessly transmit the data to a robot or receiver. The whole procedure is done in a series of blocks, each of which is vital target.

The transmitter side block diagram illustrates how lip movement is recorded as video, converted into useful digital data, and then sent wirelessly over RF or GSM. Every block is crucial to the proper conversion and transmission of quiet lip movements to the recipient.

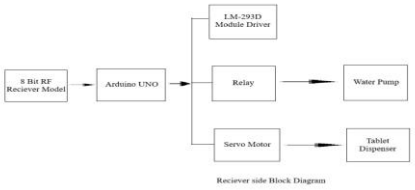


Figure 4: Block diagram of Receiver.

The 8-bit RF Receiver Module on the receiving end gathers the broadcast signal and sends the decoded data to an additional Arduino UNO. With the aid of auxiliary gear, this Arduino manages the necessary output devices. It employs an L293D Motor Driver Module to safely drive motors with enough power for applications that require motor control.

The Arduino triggers a relay, which regulates the water pump, to switch high-power devices. In a similar vein, the Arduino powers a servo motor that runs the tablet dispenser when it comes to mechanical movements. By successfully translating lip gestures into wireless orders that automate chores like pumping water or distributing medication.

VI. FLOW CHART

The overall operation for the Conversion of Lip Movement to Speech is explained in the flow chart as shown in Figure 5

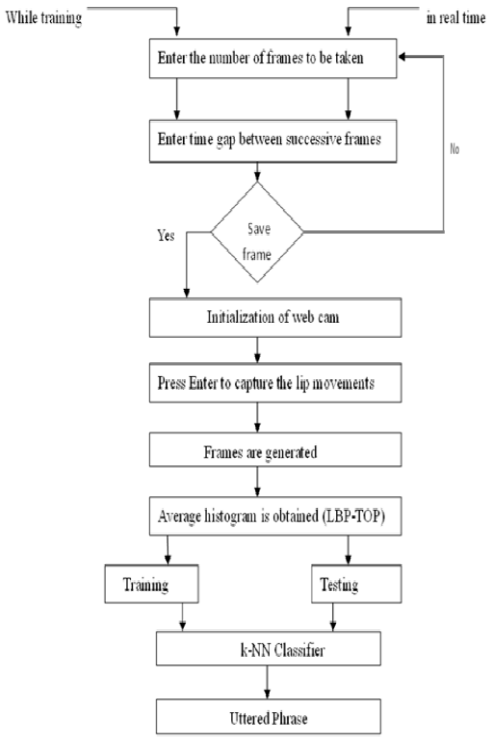


Figure 5: Flow Chart.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

The flowchart explains the process of recognizing spoken sentences using lip movements and machine learning.The system captures lip-movement frames through a webcam based on user-defined settings. The captured frames are processed using the LBP-TOP method to extract numerical features in the form of histograms. These features are used to train or test the machine-learning model. Finally, the k-Nearest Neighbor (k-NN) classifier compares the extracted features with stored data and displays the recognized spoken sentence as the output.

VI.RESULTS

Initially the system is trained for different words and in real time the data is obtained and compared with pre stored data (while training). For an example, the system is trained for the word 'WATER ON'. By capturing the video of lip movement the frames are obtained and histogram is stored. The obtained frames are as shown in the Figure 6.

In real time, video of the lip movement is captured. The frames and histogram are obtained, compared with prestored data (while training). If the data obtained is approximate to any of the stored data then the respective word is uttered and displayed.

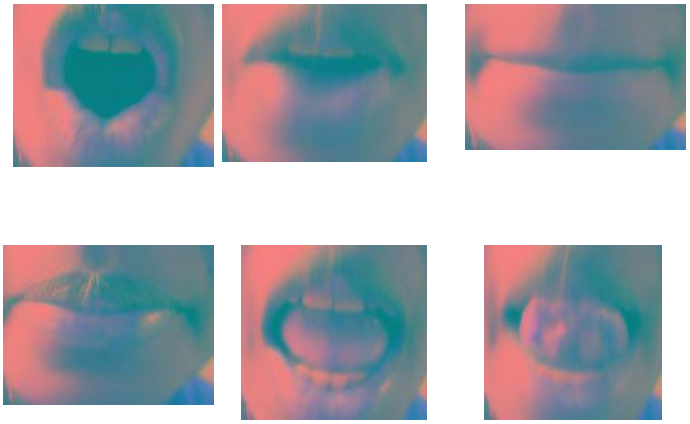
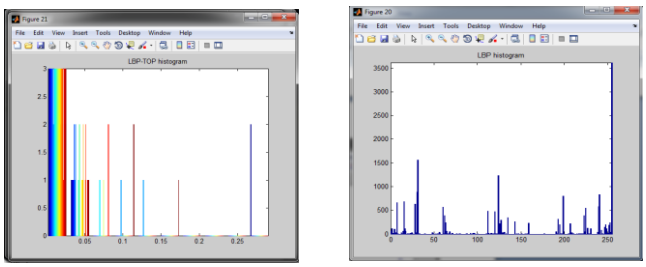


Figure 6: Frames obtained from video

The frames and histogram obtained for the word 'WATER ON' in real time is as shown in Figure 7.



.

Figure 7: Histogram of a word

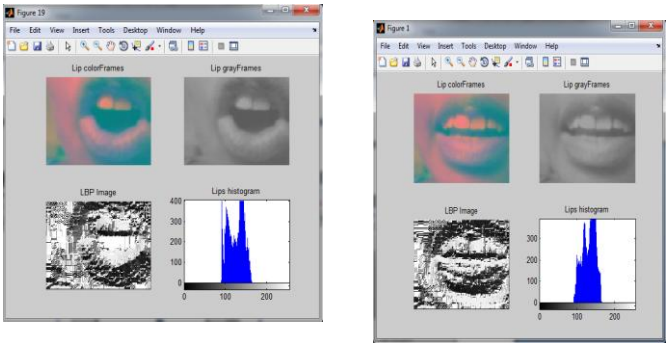


Figure 8: Frames and their histogram

For evaluating the performance of the proposed method, different design experiments were conducted that includes speaker-independent and speaker-dependent

A. Robot working

In this work, the robot works according to commands given by the patient through their lip movements. A camera keeps track of the lip patterns of the patient and the AI system translates them into commands such as asking for water or a tablet.

These commands are subsequently provided wirelessly to the robot where a microcontroller acts as the main brain and decides which action to take. When patients requests water; the controller closes a relay to activate a small water pump safely dispensing a measured amount of water to the patient. When a tablet is requested, the controller activates a servo motor inside the tablet dispenser which dispenses exactly one tablet.

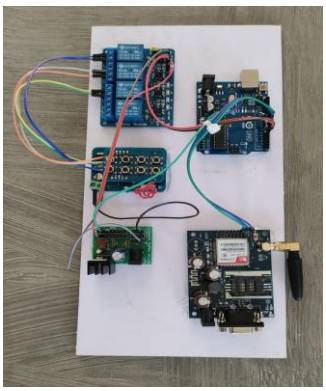


Figure 9: Transmitter side

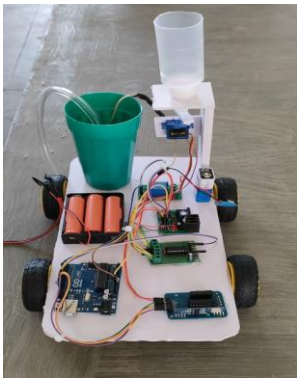


Figure 10: Receiver side

The robot can also approach the patient if needed using motor drivers and wheels. Sensors and programmed limits throughout the process prevent over dispensing and accidental motion.

This way, the robot helps speech-impaired and bedridden patients by understanding their lip movements and automatically giving them water and tablets without the need for a nurse or caretaker to be there every time. . Parts of the Robots-(Transmitter side) and (Receiver side) shown in the figure of 9 and 10.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Table 1 Demonstrates percentage of recognition of Robot

Phrases	Number of times tested	Number of times recognized	Percentage of Recognition
FORWARD	05	4	80
WATER	05	5	90
TABLET DISPENSE	05	4	80

VII. CONCLUSION AND FUTURE WORK

The focus of this study is to strengthen the functional ability of elderly individuals and to assist dumb people to convey information. The entire process starts out with the image acquisition. It is implemented by the recording of the movement of the lips of the person speaking or uttering the alphabets.

This WAV movie has been compromised into PNG static image frames. The extract features of the image and save them as a histogram. Image is compared with histogram of database for categorization. Voice turns on and the said phrase is shown.

Also, the physically challenged will be able to enhance their functional capabilities through the use of a robot and devices that can be operated based on spoken words. So this method is helpful for the elderly, deaf, dumb, blind and physically disabled.

Future Scope: The current findings are encouraging, but the present algorithm still has room for improvement. Future plan is to have research not only on isolated phrases but also on the continuous speech and to improve the accuracy of recognition of uttered phrase. First, we require automated procedures to find optimal parameter settings during training. Second, more sophisticated criteria could be used to enhance the recognition rate.Third, to enhance this project without Laptop by using Raspberry pi. Fourth, other classifiers and faster algorithms will also be observed

REFERENCES

[1] Kim, M., Hong, J., Ro, Y. M, "— Lip to Speech Synthesis with Visual Context Attentional GAN," April 2022.

[2] Akbari et al., 2018. H. Akbari, H. Arora, L. Cao, N. Mesgarani, LIP2AUD-SPEC: speech reconstruction from silent lip movements video. Proceedings of ICASSP2018, pp. 2516- 2520.

[3] Afouras, T., Chung, J. S., Senior, A., et al. — LRS3-TED: a large-scale dataset for visual speech recognition — arXiv (2018).

[4] Cheok et al., 2017. M. J. Cheok, Z. Omar, M. H. Jaward, A review of hand gesture and sign language recognition techniques. International journal of machine learning and cybernetics, vol. 10, no. 1, pp. 131-153.

[5] Hegde et al., "Towards Accurate Lip-to-Speech Synthesis in-the-Wild" October 2023.

[6] Li, X., Wang, X., Wang, K., & Lian, S." A Novel Speech-Driven Lip-Sync Model with CNN and LSTM" May 2022.

[7] Automatic Visual Lip Reading: A Comparative Review of Machine-Learning Approaches — Results in Engineering, September 2025.

[8] Abbasi, A. F., Yousefi-Koma, A., Firouzabadi, S. D., et al.," Integrating Persian Lip Reading in Surena-V Humanoid Robot for Human-Robot Interaction" 2025.

[9] "A Novel Speech to Mouth Articulation System for Realistic Humanoid Robots", Vol101,article number 54,(2021).

[10] Assael, Y. M., Shillingford, B., Whiteson, S., de Freitas, N. — LipNet : End-to-End Sentence-level Lipreading — arXiv / 2016. Landmark end-to-end visual speech recognition model (spatio-temporal CNN + RNN + CTC). Useful for feature extraction / liprepresentation.

[11] Hunt, W., Soorati, M., et al "A Survey Of Language-Based Communication",June 2024.

[12] N. Rathipriya , N. Maheswari, " A Comprehensive Review Of Recent Advances In Deep Learning For Automated Lip Reading" 2024.

[13] Shwetha K S, Rohith M K, Sakshi Prashant Yandagoudar, Sinchana "Silent Interpreter: Analysis Of Lip Movement And Extracting Speech Using Deep Learning" 2024.

[14] K. Vayadande, T. Adsare, N. Agrawal, T. Dharmik, A. Patil and S. Zod, "LipReadNet: A Deep Learning Approach to Lip Reading," 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC), Dharwad, India, 2023, pp. 1-6, doi: 10.1109/ICAISC58445.2023.10200426.