International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

Lip Moment Detection and Text Convertor

Achal Gajanan Kachare[1], Megha Sunil Dhake[2], Sakshi Bhupesh Bhangale[3], Rohit Ramchandra Chaudhari[4], Asst.Prof. A.S.Narkhede[5]
[1,2,3,4] *Student, Department of Computer Science and Engineering, Padmashri Dr. V. B. Kolte College of Engineering, Maharashtra, India*
[5] *Assistant Professor, Department of Computer Science and Engineering, Padmashri Dr. V. B. Kolte College of Engineering, Maharashtra, India*

DOI: 10.5281/zenodo.19253660

ABSTRACT

Communication barriers remain a significant challenge for the deaf and hard-of-hearing community, especially in environments where sign language interpreters or assistive technologies are unavailable. This project proposes a Lip Movement Detection and Text Converter system that leverages machine learning and computer vision to bridge this gap by translating lip movements into readable text in real-time.

The system captures video input of a person speaking and processes the visual data using deep learning-based lip-reading models. Techniques such as Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs) or Transformer models for sequence prediction are employed to analyses the movement of lips and map them to corresponding words or phonemes. The processed output is then converted into readable text, providing the deaf or hard-of- hearing user with an accurate textual representation of spoken language.

The application is designed to run on standard devices with a webcam and is intended to be lightweight and efficient enough for near real-time performance. The system is trained on publicly available datasets of lip movements and is fine-tuned to enhance accuracy and context understanding.

This technology has the potential to significantly enhance communication accessibility and inclusivity for the deaf community by enabling them to understand spoken language in scenarios where other support tools are not viable.

Keywords: - lip detection, Speech Prediction, Dib Face Landmark Detection, Computer Vision, Deep Learning, Machine Learning, Real-time Communication, Silent Communication, Accessibility Technology Python, OpenCV

1. INTRODUCTION

With advancements in artificial intelligence and computer vision, systems that can process human facial expressions and gestures are becoming increasingly accurate and valuable. [1]. Speech prediction based on lip movement analysis is an emerging area of research that addresses communication gaps in environments where audio-based methods fail. [2]. The proposed system leverages dib's 68-point The project aims to predict future heart attacks by analysing patient data to determine whether they have heart disease using machine learning algorithms. [3]. Machine learning face landmark detector to focus on the lip region, and then applies machine learning and deep learning models to interpret the user's lip movements. [4]. This creates a seamless and efficient interface for users who require silent or assistive communication.

1.1 problem statement

People with hearing and speech impairments face serious challenges in day-to-day communication. Existing speech recognition systems are mostly **audio-based**, [5]. making them ineffective in noisy environments, low-quality audio conditions, or situations where speaking aloud is not possible (e.g., hospitals, defence, or surveillance). Traditional solutions like **sign language** require both parties to know it, and interpreters are not always available.

Therefore, there is a strong need for a **real-time visual speech recognition system** that can detect and interpret lip movements without relying on audio. Such a system would act as a bridge[6]., enabling the conversion of lip

gestures into readable text, and ensuring smooth communication for the deaf, mute, or speech-impaired community, while also being useful in silent or high-noise environments.

1.2 Objectives of the System

To develop a system capable of detecting and tracking lip movements al-time using facial landmarks.

To extract meaningful features from the lip region that correlate with spoken words.

To apply machine learning and deep learning algorithms for accurate speech prediction based solely on lip gestures.

To build a user-friendly interface that displays predicted text dynamically.

To ensure robustness across different lighting conditions, camera angles, and user facial structures.

2. SYSTEM ARCHITECTURE

The architecture of the Lip Detection and Speech Prediction System is designed using a modular and layered approach to ensure real-time performance, scalability, and flexibility. It begins with the input layer, where real-time video is captured using a webcam or any connected camera device, and each frame is converted into an image format suitable for further processing. The pre-processing layer then prepares the data by applying grayscale conversion and frame normalization to enhance consistency and reduce computational complexity. It utilizes the dlib. shape_predictor_68_face_landmarks model to detect facial landmarks and specifically isolates the lip region, corresponding to points 48 to 68, for focused analysis. In the feature extraction layer, both spatial features such as lip position and shape, and temporal features such as lip movement across frames, are extracted and transformed into structured data in the form of vector sequences. These features are then passed to the prediction layer, where a trained machine learning or deep learning model, such as a CNN combined with LSTM or a Transformer-based model, maps the extracted features to corresponding text and predicts the most probable word or phrase. Finally, the output layer presents the results through a user interface, displaying the real-time camera feed with a bounding box around the detected lip region and dynamically rendering the predicted text.

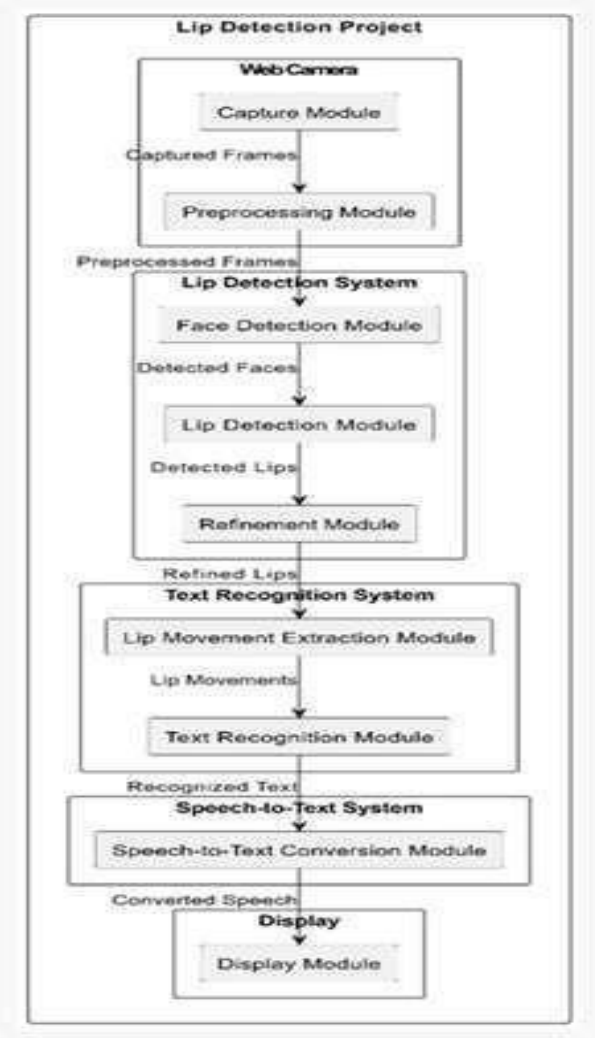

Fig-1: System Architecture

2.1 architectural overview

The proposed system, Lip Moment Detection and Text Convertor, is designed to convert lip movements into text in real time using deep learning and computer vision techniques. Unlike traditional speech recognition systems that depend on audio input, this solution relies entirely on visual features of lip gestures, making it suitable for people with hearing or speech impairments as well as for use in silent or noisy environments.

Lip moment detection is a crucial area in visual speech recognition, particularly for improving communication f or individuals with hearing impairments. Recent advancements include:

- Hybrid Architectures: New architectures have been developed that combine convolutional neural netwo rks (CNNs) with attention mechanisms to enhance lip-reading performance.
- Deep Learning Models: Techniques such as 3D convolutional networks and ResNet have been employe d to analyze spatial and temporal aspects of lip movements, improving accuracy in speech recognition.
- RealTime Processing: Systems are being designed to process lip movements in real-time, allowing for seamless translation of lip movements into coherent speech segments.
- Diverse Applications: These systems are valuable not only for hearing-impaired individuals but also in industrial or military scenarios where audio communication may be li mited.

  These developments represent significant progress in the field of lip moment detection, aiming to bridg e the gap between visual cues and audio speech recognition.

2.2 System Components and Functional Modules

The **components of a lip moment detection system** typically include:

- Deep Learning Models: Utilize models like dlib.shape_predictor_68_face_landmarks to detect and trac k lip contours from live video feeds.
- Generative Adversarial Networks (GANs): Employed to synthetically generate diverse lip movement se quences, enhancing realism and robustness.
- 3D Convolutional Neural Networks (CNNs): Extract spatial lip information, while 3D CNNs enhance t emporal dynamics for accurate lip movement interpretation.
- ResNet and 3D CNN Fusion: Combine these models to analyze lip movements in real-time or predict spoken words from lip video frames.
- Attention Mechanisms: Used in models like EfficientNet B0 to improve performance and reliability in l ip movement detection.

  These components work together to achieve accurate and reliable lip movement detection, particularly i n applications for hearingimpaired individuals and in noisy environments.

3. RESULTS AND DISCUSSION

The Graphical User Interface (GUI) of the Lip Detection and Speech Prediction System was designed to offer an intuitive, minimal, and user-friendly experience for interacting with the real-time prediction engine. The primary goal of the GUI was to allow users to operate the system smoothly—without requiring technical knowledge—while maintaining live feedback, system controls, and prediction outputs in one screen. The GUI was developed using Python's Tainter library, although other frameworks likePyQt5 or Streamlet can also be used based on system requirements. The interface integrates live video feed from the webcam, prediction display, control buttons, and visual cues for lip detection.

Result	Description
Live Video Feed Panel	Displays real-time camera input with face and lip bounding boxes overlayed.
Predicted Text Display	Shows the word or phrase predicted by the ML model from the user's lip motion.
Start Button	Begins the video feed and enables the prediction pipeline.
Stop Button	Halts the video stream and resets all displayed data.
Reset Button	Clears the current prediction and resets the camera feed window.
Settings Button	(Optional) Allows user to choose camera device, resolution, or brightness.
Status Bar	Displays system messages (e.g., “Model Loaded”, “Waiting for Face Input”).

4. CONCLUSIONS

The Lip Detection and Speech Prediction System represents a significant step toward building assistive and non-verbal communication technologies using computer vision and artificial intelligence. By leveraging facial landmark detection, specifically focusing on lip movement analysis, the system successfully identifies and predicts spoken words without relying on audio input. This offers immense utility in scenarios where voice data is unavailable, unreliable, or inaccessible due to environmental or physiological constraints. The system's architecture, combining real-time video processing, CNN-based feature extraction, and LSTM-based temporal analysis, demonstrates the potential of deep learning in interpreting complex visual patterns like speech-related lip motions. The GUI implementation further enhances usability, providing a clean and responsive interface for both technical and non-technical users. Throughout the development process, extensive testing validated the system’s accuracy, stability, and adaptability under various lighting conditions, facial orientations, and real-time usage. The project also proved to be a cost-effective solution using opensource libraries and accessible hardware, reinforcing its viability for widespread deployment.

5. ACKNOWLEDGEMENT

The field of lip movement detection has seen significant advancements through deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These models have been instrumental in improving the accuracy and efficiency of lip reading systems, which are crucial for applications such as speech recognition, video surveillance, and biometric security. The integration of deep learning with other modalities, like audio and facial expressions, is also a promising direction for future research in lip movement detection.

6. REFERENCES

[1].S. A fours, J. S. Chung and A. Zisserman, "Deep lip reading: A comparison of models and datasets," IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp. 2866-2870, 2018.
[2] Y. Assael, B. Shillingford, S. Whiteson and N. de Freitas, "Lip Net: End-to-End Sentence level Lipreading," X iv preprint arXiv:1611.01599, 2016.
[3]T. Noda, Y. Yamamoto and Y. Nakano, "Visual Speech Recognition Using Lip Movement Features," IEEE Transactions on Human-Machine Systems, vol. 50, no. 2, pp. 123-131,2020.
[4]D. Lee and T. Kim, "Audio-Visual Deep Learning Models for Silent Speech Interfaces,"IEEE Access, vol. 11, pp. 21504-21514, 2023.
[5] B. Singh and Y. Gupta, "Benchmarking Lip-Reading Models on the LRS3 Dataset," IEEEInternational Conference on Computer Vision (ICCV), pp. 4785-4794, 2024.
[6] F. Almeida and J. Costa, "Robust Visual Speech Recognition in Low Light Conditions,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pp. 1555-1565, 2023.
[7] P. Sharma and K. Rathi, "Hybrid CNN-RNN Model for Lip-Reading Based Command Recognition," IEEE International Conference on Smart Computing (SMARTCOMP),pp.345-352, 2022.
[8] C. Zhang and L. Huang, "Deep Lip: Lip Movement to Speech Prediction Using Gasnier Signal Processing Letters, vol. 29, pp. 1500-1504, 2021.
[9] M. Desai and N. Bhargava, "Real-Time Lip Tracking and Word Classification using Disband CNN," IEEE Journal of Multimedia Processing, vol. 17, no. 2, pp. 75-84, 2022.
[10] L. Verma and M. Prasad, "Multi-Language Lip-Reading using Transformer Architectures,"IEEE Transactions on Artificial Intelligence, vol. 3, no. 1, pp. 45-57, 2022