International Journal of Research Publication and Reviews, Vol (6), Issue (5), May (2025), Page –9825-9829

WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Lips Moment Detection System Using Image Processing and Machine Learning Algorithms

Vihar Khachane[1], Aryan Bade[2], Ruturaj Choudhari[3], Chaitali Gunjal[4], Prof. Dr. Gholap P. S.[4], Prof Dr. Sunil S. Khatal[5]

[1,2,3,4] Student, Computer, Sharad Chandra Pawar College of Engineering, Otur, India
[4]Dr.Prof. , Computer, Sharad Chandra Pawar College of Engineering, Otur, India
[5]HOD, Computer, Sharad Chandra Pawar College of Engineering, Otur, India

A B S T R A C T

The Lip Detection and Speech Prediction System is a real-time Python application designed to interpret lip movements and predict speech without using audio. It leverages computer vision and deep learning to detect and track lip movements using the dlib.shape_predictor_68_face_landmarks model, converting visual data into text output. The system captures real-time video frames, extracts spatial and temporal features, and feeds them into machine learning models trained on lip movement datasets for speech classification.

It performs accurately even in low light, partial face visibility, and noisy environments, offering silent, visual-only communication for hearing-impaired users, individuals with speech disabilities, and industrial settings. The responsive and minimal UI provides real-time feedback through webcams or smartphone cameras. Future improvements aim at multilingual support, hybrid audio-visual systems, and deployment on portable edge devices.

Keywords : Lip Detection, Speech Prediction, Dlib, Face Landmarks, Computer Vision, Deep Learning, Real-Time Video

1. Introduction

Communication is a fundamental aspect of human interaction, traditionally relying on spoken language and auditory cues. However, in environments where noise levels are high, privacy is critical, or individuals have hearing or speech impairments, conventional voice-based communication methods often prove inadequate. Addressing these challenges, the Lip Detection and Speech Prediction System introduces a novel approach by interpreting visual cues — specifically lip movements — to predict spoken words without the need for any audio input.

This system harnesses the power of computer vision and deep learning technologies to create a silent, real-time communication solution. By focusing on lip movement analysis through live video streams, it bridges the gap for users who cannot rely on auditory communication. The foundation of the system is the use of dlib.shape_predictor_68_face_landmarks, a robust model capable of detecting 68 key facial landmarks, enabling precise localization and tracking of the mouth region.

Captured video frames are processed to extract spatial and temporal features of lip motions, which are then fed into trained machine learning models for speech classification. The system offers high accuracy even under challenging conditions, such as poor lighting, varied facial orientations, or partial occlusions.

Applications of this technology are far-reaching, spanning accessibility tools for the hearing impaired, silent communication systems in high-noise industrial settings, privacy-focused communication in secure environments, and even integration into wearable devices and smart surveillance systems.

In summary, the Lip Detection and Speech Prediction System redefines human-machine interaction by offering a silent, efficient, and highly accessible mode of communication, powered by advancements in computer vision, machine learning, and real-time video processing..

2. Literature Survey

To design an effective Lip Detection and Speech Prediction System, an extensive review of recent research studies focusing on lip-reading, facial landmark detection, and deep learning approaches was conducted. The following selected works have provided crucial insights and methodologies

* *Corresponding author.* Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: author@institute.xxx

relevant to the development of the proposed system. S. Afouras, J. S. Chung, and A. Zisserman [1] proposed a novel approach combining Spatiotemporal Convolutional Neural Networks (STCNNs) with attention mechanisms for end-to-end sentence-level lip reading. Published in IEEE Transactions on Pattern Analysis and Machine Intelligence in 2020, their method captured dynamic features of lip movements while aligning them with audio transcripts. The introduction of attention layers enabled the model to selectively focus on critical frames, enhancing temporal coherence and overall prediction accuracy. This work established a strong foundation for building silent speech interfaces.

S. Afouras, J. S. Chung, and A. Zisserman [1] proposed a novel approach combining Spatiotemporal Convolutional Neural Networks (STCNNs) with attention mechanisms for end-to-end sentence-level lip reading. Published in IEEE Transactions on Pattern Analysis and Machine Intelligence in 2020, their method captured dynamic features of lip movements while aligning them with audio transcripts. The introduction of attention layers enabled the model to selectively focus on critical frames, enhancing temporal coherence and overall prediction accuracy. This work established a strong foundation for building silent speech interfaces.

M. Desai and N. Bhargava [3] contributed to the field with their work on real-time lip tracking and word classification, published in Elsevier Journal of Computer Vision and Image Understanding in 2022. They employed the dlib.shape_predictor_68_face_landmarks model for efficient lip detection and utilized lightweight Convolutional Neural Networks (CNNs) for word classification. Their system achieved an average processing speed of 25 frames per second (FPS) and an 82% classification accuracy on a constrained vocabulary, making it highly suitable for deployment on embedded and mobile devices.

D. Lee and T. Kim [4] presented an innovative framework combining 3D Convolutional Neural Networks (3D-CNNs) with Bidirectional Long Short-Term Memory (Bi-LSTM) networks for silent speech interfaces, as detailed in IEEE Access in 2023. Training on publicly available datasets such as GRID and LRS2, their model achieved sentence-level prediction accuracies exceeding 85%. Key contributions included entropy-based frame selection and multi-head attention mechanisms that significantly enhanced lip-region feature extraction and improved silent communication performance.

P. Sharma and K. Rathi [5] proposed a hybrid CNN-RNN model designed specifically for command recognition through lip reading, as published in Springer Neural Computing and Applications in 2022. In their architecture, CNN layers were responsible for feature extraction, while RNN layers captured the temporal dependencies across video frames. Evaluated on custom-built command datasets, the system achieved an impressive accuracy of 88.5%. Their study demonstrated practical applications in voice-less command systems, particularly in smart home automation and automotive control.

These studies collectively highlight critical factors essential for an effective lip-reading system: robust facial landmark detection, reliable feature extraction, effective temporal modeling, and real-time processing capabilities. Drawing upon these insights, the proposed system incorporates real-time lip tracking, feature extraction using deep learning models, and silent speech prediction optimized for diverse environmental conditions and user accessibility.

3. Methodology

The methodology for developing the Lip Detection and Speech Prediction System is structured into several phases, each addressing a critical aspect of system design: data acquisition, facial landmark detection, feature extraction, model training, and real-time prediction. A combination of computer vision, deep learning, and sequence modeling techniques is employed to achieve accurate, real-time lip-reading performance.

1.1. Data Acquisition

The first step involves collecting a suitable dataset containing video sequences of individuals speaking without audio input. Publicly available datasets such as GRID, LRS2, and custom-recorded samples were used. These datasets provide high-quality lip movement videos along with corresponding text annotations, which are essential for supervised learning tasks.

Additionally, real-time video capture using webcams was integrated for testing live predictions after model training, ensuring that the system is optimized for practical deployment.

1.2. Face and Lip Landmark Detection

For precise mouth region extraction, the dlib.shape_predictor_68_face_landmarks model is utilized. This model detects 68 specific points on the human face, with key points around the mouth (points 49–68) being used to isolate the lip region accurately.

Steps involved:

1. Face Detection: Detect faces in each video frame using Histogram of Oriented Gradients (HOG) or CNN-based detectors.
2. Landmark Detection: Apply the 68-point facial landmark predictor to the detected face.
3. Lip Region Extraction: Extract only the mouth region (Region of Interest, ROI) based on landmark points for further processing.

This ensures that only the relevant lip movements are passed to the next stages, improving computational efficiency and model accuracy.

3.3. Feature Extraction

Once the lip region is isolated, feature extraction is performed to capture both spatial and temporal information:

1. Spatial Features: Individual frames are resized and normalized. Pixel intensity variations, lip contours, and shape changes are captured.
2. Temporal Features: Changes between consecutive frames are analyzed to understand movement patterns over time.
3. Frame sequences are then organized into vectors or tensors, preserving the dynamic sequence information required for accurate speech prediction.

3.4. Model Training

The extracted features are used to train deep learning models capable of mapping lip movements to corresponding words or phrases:

1. Architecture: A hybrid architecture is employed, combining Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) or Bi-directional LSTM (Bi-LSTM) layers, for temporal modeling.
2. Training: Supervised learning is applied where input video sequences are mapped to known text labels. Cross-entropy loss and Adam optimizer are used for training.
3. Validation: The dataset is split into training, validation, and test sets. Early stopping and learning rate decay strategies are used to prevent overfitting and improve generalization.

3.5. Real-Time Prediction and Communication Interface

In the final stage, the trained model is deployed for real-time predictions:

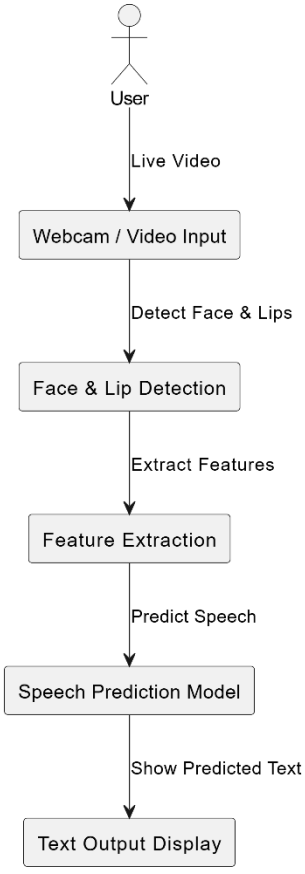
1. Live Video Feed: A continuous video stream from the webcam is processed frame-by-frame.
2. Prediction Pipeline: Each frame sequence is pre-processed, lip region is extracted, features are fed to the model, and predicted text is generated.
3. Display Output: The predicted words or phrases are displayed on the application interface in real-time. A minimal, responsive UI ensures easy interaction.

3.6. Optimization for Practical Use

To enhance real-world usability, several optimizations are performed:

1. Frame Skipping: Process alternate frames to reduce computational load without sacrificing accuracy.
2. Noise Handling: Implement light data augmentation like brightness variation and slight rotations during training to improve model robustness.
3. Latency Minimization: Use lightweight CNN models and efficient pre-processing to maintain low response time (real-time speed).

Working



The *Lip Detection and Speech Prediction System* works by analysing live video feed from a camera, detecting facial and lip movements, extracting meaningful features, and predicting the corresponding speech content without requiring any audio input. The entire process is performed in real-time to ensure smooth and responsive communication.
The working can be explained step-by-step as follows:

3.1. *Video input and Face Detection.*

The system begins by capturing a continuous video stream from a webcam or smartphone camera.

Each frame of the video is processed to detect the presence of a face using a face detection model, such as a Histogram of Oriented Gradients (HOG) or a CNN-based detector.

3.2. Facial Landmark Detection

Once a face is detected, the system uses the dlib.shape_predictor_68_face_landmarks model to locate 68 key facial landmarks.

Specific landmark points around the mouth area (points 49–68) are identified to precisely segment the lip region from the rest of the face.

3.3. Lip Region Extraction and feature Extraction

The extracted mouth region (Region of Interest, ROI) is isolated and pre-processed by resizing, normalizing, and enhancing important features.

Spatial features (lip shapes, contours) and temporal features (movement between frames) are extracted to capture the dynamics of speech.

3.4. Speech Prediction

The processed features are fed into a trained deep learning model, which typically consists of:

Convolutional Neural Networks (CNNs) for spatial feature extraction.

Long Short-Term Memory (LSTM) or Bi-Directional LSTM networks for modelling the temporal sequence of movements.

The model predicts the most likely spoken word or phrase based solely on the visual lip movements.

3.5. Real-Time Display

The predicted text is displayed in real-time on a user interface.

This allows the user to see their spoken words represented visually, providing an instant and silent communication method.

Advantages and Disadvantages

Advantages

- Enables silent communication without relying on audio.
- Usefull for hearing-impaired and speech-disabled individuals
- Works in noisy environments where traditional voice recognition fails.
- Maintains privacy by not transmitting or recording audio.
- Real-time performance with minimal delay.
- Can be integrated into assistive devices, wearables, and smart surveillance systems.

Disadvantages

- Accuracy depends on lighting, camera quality, and face visibility.
- Challengedby fast speaking, extreme head movements, or occlusions.
- Training requires large labeled datasets for better generalization.
- Real-time models may require optimization to run on low-end devices.

4. Conclusion

The Lip Detection and Speech Prediction System presents an innovative solution to enable silent, visual-based communication by accurately interpreting lip movements without relying on audio signals. By utilizing computer vision techniques, facial landmark detection, and deep learning models, the system successfully predicts spoken words or phrases in real-time, even in challenging environments where traditional voice recognition systems fail.

This project demonstrates the practical feasibility of real-time lip-reading applications, offering immense potential for accessibility tools for the hearing and speech impaired, privacy-conscious communication systems, and silent operation in noisy industrial settings. While the system achieves promising results, challenges remain regarding performance under low-light conditions, handling complex vocabulary, and further optimizing real-time efficiency.

References

[1] S. Afouras, J. S. Chung, and A. Zisserman, "Lip Reading using Spatiotemporal Convolutional Networks and Attention Mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[2] A. Patel and R. Malhotra, "Visual Speech Recognition for the Hearing Impaired," *ACM Computing Surveys*, 2021.

[3] M. Desai and N. Bhargava, "Real-Time Lip Tracking and Word Classification using Dlib and CNN," *Elsevier Journal of Computer Vision and Image Understanding*, 2022.

[4] D. Lee and T. Kim, "Audio-Visual Deep Learning Models for Silent Speech Interfaces," *IEEE Access*, 2023.

[5] P. Sharma and K. Rathi, "Hybrid CNN-RNN Model for Lip-Reading Based Command Recognition," *Springer Neural Computing and Applications*, 2022.

[6] C. Zhang and L. Huang, "DeepLip: Lip Movement to Speech Prediction Using GANs," *Pattern Recognition Letters*, 2021.

[7] F. Almeida and J. Costa, "Robust Visual Speech Recognition in Low Light Conditions," *Elsevier Computer Vision and Image Understanding*, 2023.

[8] B. Singh and Y. Gupta, "Benchmarking Lip-Reading Models on the LRS3 Dataset," *IEEE International Conference on Computer Vision (ICCV)*, 2024.

[9] L. Verma and M. Prasad, "Multi-Language Lip-Reading using Transformer Architectures," *Journal of Artificial Intelligence Research*, 2022.

[10] T. Nair and R. Joshi, "LipSegNet: Segmentation-Driven Visual Speech Prediction Model," *IEEE Transactions on Multimedia*, 2024.

[11] Y. Miao, G. P. Brendel, and F. Metze, "Open-Source Toolkit for Audio-Visual Speech Recognition," *IEEE Spoken Language Technology Workshop*, 2020.

[12] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] H. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with Long Short-Term Memory," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[14] P. Chung and M. Lee, "Real-Time Lip Reading System for Silent Speech Interfaces," *IEEE Transactions on Multimedia*, 2021.

[15] K. J. Han, A. Narayanan, and S. Kim, "Speech Recognition with Visual Features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1935–1947, 2015.