iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81385

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Decoding Speech from Labial Movements

Nimmakayala Venkata Lakshmi[1], Cherukuri Sowmya[2], Jujigiri Likitha[3], Amboji Anusha[4]
[1, 2, 3, 4]*Department of Computer Science and Engineering, Bapatla Women's Engineering College, Affiliated to Acharya Nagarjuna University, Approved by AICTE – New Delhi, Andhra Pradesh, India*

Abstract— Decoding speech from labial movements is an emerging area of research that bridges computer vision, linguistics, and artificial intelligence. By analysing the dynamic patterns of lip shapes, positions, and motions, systems can infer spoken words even in the absence of acoustic signals. This technique is particularly valuable in noisy environments, for individuals with hearing impairments, and in applications requiring silent communication. Recent advances in deep learning and image processing have enabled more accurate mapping between visual features and phonetic units, improving recognition rates and robustness across diverse speakers. The integration of multimodal cues, such as facial expressions and contextual language models, further enhances performance, making visual speech decoding a promising complement to traditional audio-based systems. Ultimately, this field contributes to more inclusive human-computer interaction and opens new possibilities for assistive technologies and secure communication.

Keywords: Lip Reading, Computer Vision, Deep Learning, Speech Recognition, CNN, LSTM, Transformer, Visual Speech Decoding, Phoneme Recognition, Assistive Technology.

I. INTRODUCTION

In today's rapidly advancing technological era, communication remains a cornerstone of human interaction and societal participation [1]. For individuals with speech impairments, conveying thoughts, emotions, and needs poses a significant challenge. Traditional solutions such as sign language provide an important means of communication; however, they rely heavily on mutual understanding between parties. This dependence creates barriers, leading to social exclusion and reduced participation in everyday life.

The proliferation of intelligent systems has opened new avenues for assistive communication technologies. Among these, the automated recognition of speech from visual lip movements—commonly referred to as lip reading or visual speech recognition—has emerged as a particularly compelling direction. Unlike audio-dependent systems, lip-reading technologies can function in acoustically hostile environments where conventional microphones fail, as well as in scenarios requiring silent or covert communication [2].

We propose an artificial intelligence-driven system for decoding speech from labial movements using computer vision and deep learning. The system analyses lip movements captured through a standard camera to infer spoken words without the need for audio signals. It is designed to be real-time, portable, and scalable, operating on commodity hardware without specialized equipment. The pipeline comprises video acquisition, face detection, lip localization, feature extraction, sequence modelling using Convolutional Neural Networks (CNNs) combined with Long Short-Term Memory (LSTM) networks, and language model-based output correction.

The proposed system has wide-ranging applications. In noisy industrial or public environments, it enables speech capture where audio is unintelligible. In military and covert operations, it supports silent communication without risk of acoustic interception. In secure videoconferencing, it provides an additional authentication layer. Furthermore, as a multimodal component, it can augment existing audio-based automatic speech recognition systems to improve robustness under degraded acoustic conditions [3].

The remainder of this paper is organised as follows. Section II presents the problem statement. Section III discusses the limitations of existing systems. Section IV describes the proposed system architecture in detail. Section V outlines the algorithms and flowcharts. Section VI discusses results and observations. Section VII concludes the paper, and Section VIII outlines directions for future work.

II. PROBLEM STATEMENT

Individuals with speech impairments face significant barriers in day-to-day communication. Although audio-based automatic speech recognition (ASR) systems have achieved considerable accuracy in controlled environments, they are fundamentally limited by their dependence on acoustic signals [4]. In noisy settings—such as factories, crowded public spaces, or emergency scenarios—the performance of audio-based systems degrades substantially.

Traditional lip-reading approaches, relying on hand-crafted feature descriptors and shallow classifiers, suffer from low accuracy, poor generalisation across speakers, and inability to handle natural speech variation [5]. Moreover, most existing systems operate offline and cannot provide the real-time feedback required for practical assistive applications. There is, therefore, an urgent need for an intelligent, camera-based, and real-time system capable of accurately decoding speech from labial movements without dependence on acoustic input.

III. EXISTING SYSTEM

Several prior approaches to visual speech recognition have been proposed in the literature; however, each exhibits notable limitations. Audio-based automatic speech recognition systems represent the dominant paradigm but fail entirely in the absence of acoustic input or in the presence of significant environmental noise [6]. While noise-cancellation algorithms mitigate some of these effects, their efficacy diminishes under severe acoustic conditions.

Early lip-reading systems relied on hand-crafted features such as Active Appearance Models (AAMs) and Discrete Cosine Transforms (DCTs) combined with Hidden Markov Models (HMMs) for sequence modelling. These approaches demonstrated limited accuracy and poor cross-speaker generalisation due to the high variability in lip shapes, speaking rates, and styles [7]. Furthermore, these systems typically required extensive speaker-dependent training data, making them impractical for real-world deployment.

More recent deep learning-based approaches, such as LipNet [8], demonstrated significant improvements in word-level accuracy on controlled datasets. However, such systems frequently struggle under uncontrolled real-world conditions, including variation in illumination, head pose, and occlusion. Additionally, existing systems lack multimodal integration—the combination of visual features with contextual linguistic knowledge—which is essential for handling the inherent ambiguity in lip movements across phonetically similar phonemes. The absence of a robust, end-to-end real-time lip-to-speech decoding system that is both speaker-independent and language-model-augmented motivates the proposed work.

IV. PROPOSED SYSTEM

The proposed system for decoding speech from labial movements follows a modular pipeline architecture, enabling real-time operation from video acquisition through to transcribed text output. Each module is designed to perform a well-defined function, and together they form an end-to-end visual speech recognition system. The system operates entirely without acoustic input, making it inherently robust to noisy environments and suitable for deployment in acoustically challenging conditions. It is built to be speaker-independent, leveraging deep learning models trained on diverse datasets to generalise across varying lip morphologies, speaking rates, and styles.

The pipeline is designed for scalability and portability, capable of running on commodity hardware such as standard laptops or embedded devices without the need for specialised equipment. A key design principle is modularity, where each component can be independently upgraded or replaced, allowing future enhancements such as the substitution of the Bi-LSTM with Transformer-based architectures or the integration of more advanced language models. The system supports both real-time webcam-based input and offline processing of pre-recorded video, thereby increasing its flexibility across various application scenarios.

Additionally, the system incorporates preprocessing techniques to handle variations in illumination, head pose, and video quality, ensuring robust and consistent feature extraction. Temporal dependencies in lip movements are effectively captured using sequential modeling approaches, enabling the system to learn contextual relationships between consecutive frames. This improves recognition performance, particularly for visually similar phonemes and rapid speech patterns.

The architecture also facilitates the use of data augmentation strategies during training, such as spatial transformations and intensity variations, to enhance generalisation capability. Furthermore, the system can be integrated with assistive technologies, including real-time subtitle generation and silent communication interfaces, making it suitable for applications in accessibility, surveillance, and human-computer interaction. The overall architecture is described below.

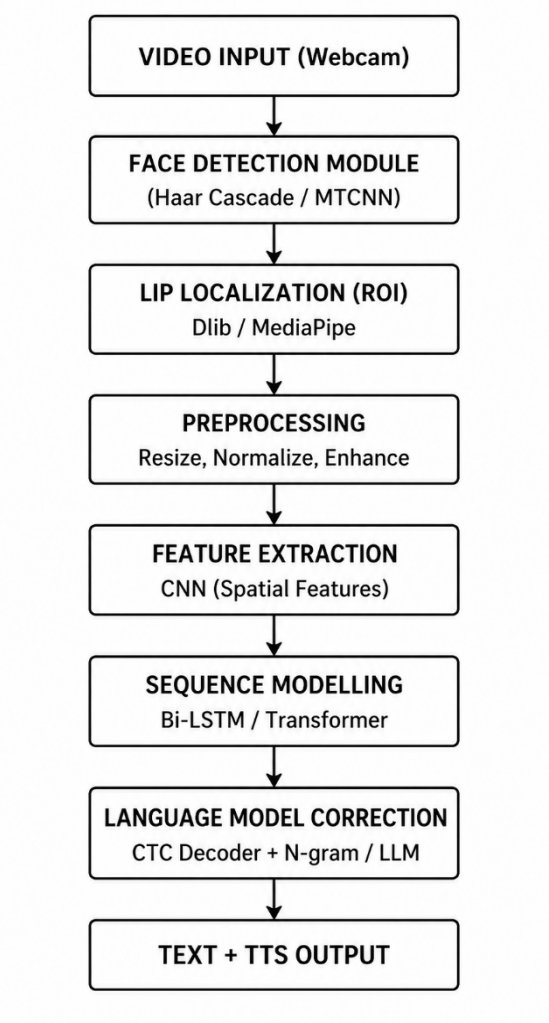


Fig. 1. System Architecture of the Proposed Lip Reading Pipeline

The architecture presented in Fig. 1 illustrates the complete end-to-end pipeline. Video is captured from a standard webcam; each frame is processed through face detection and lip localisation to extract a Region of Interest (ROI). The ROI undergoes preprocessing before being fed into a CNN-based feature extractor, followed by a temporal sequence model and a language model for contextual correction. The final output is presented as text and optionally as synthesised audio via a Text-to-Speech (TTS) engine.

A. Input Acquisition Module

The system captures real-time video from a webcam at a standard frame rate of 25 frames per second (FPS). Each frame is represented as a three-channel RGB image of resolution 640 × 480 pixels. The raw video stream is denoted as:

$$V = \{F_1, F_2, F_3, \ldots, F_n\} \qquad (1)$$

where V denotes the video stream and F_i represents the i-th frame in the sequence. The frame rate controls the temporal resolution of lip movement capture and is a critical parameter for downstream sequence modelling.

B. Face Detection and Lip Localisation

Each captured frame is passed through a face detection algorithm. The system employs the Multi-Task Cascaded Convolutional Network (MTCNN) for robust face detection, which outputs a bounding box B = (x, y, w, h) defining the facial region. Subsequently, the Dlib 68-point facial landmark predictor is applied to identify the lip landmarks, which occupy indices 48 through 67 in the standard Dlib landmark model. The lip region of interest (ROI) is extracted as:

$$R = Crop(F_i, L_{48:67}) \quad (2)$$

where $L_{48:67}$ denotes the set of lip landmarks and Crop denotes the spatial extraction of the corresponding image region. The extracted ROI is normalised to a fixed size of 64 × 128 pixels to ensure consistent input dimensions for the downstream model.

C. Preprocessing

The extracted lip ROI undergoes a series of preprocessing operations to improve feature quality and model robustness. First, the ROI is converted to grayscale to reduce computational complexity while retaining essential shape information. A Gaussian filter is applied to suppress high-frequency noise. The pixel intensities are then normalised to the range [0, 1] using:

$$F' = (F - \mu) / \sigma \quad (3)$$

where μ and σ represent the mean and standard deviation of pixel intensities computed over the training dataset. This normalisation ensures that the input distribution remains stable across frames and speakers, facilitating faster convergence during training.

D. Feature Extraction using CNN

Spatial features are extracted from the preprocessed lip ROI using a Convolutional Neural Network. The CNN architecture comprises three convolutional blocks, each consisting of a convolutional layer, Batch Normalisation, ReLU activation, and Max-Pooling. The feature map produced after the final convolutional block is flattened into a feature vector:

$$f_i = CNN(R_i) \in \mathbb{R}^d \quad (4)$$

where f_i denotes the feature vector for frame i and d is the dimensionality of the extracted feature space. The CNN is pre-trained on a large-scale lip movement dataset and fine-tuned on the target corpus to improve speaker independence.

E. Temporal Sequence Modelling

Speech is an inherently temporal phenomenon; isolated frame features are insufficient for accurate phoneme and word recognition. The sequence of CNN feature vectors, $F = \{f_1, f_2, ..., f_n\}$, is fed into a Bidirectional LSTM (Bi-LSTM) network that models temporal dependencies in both forward and backward directions. The hidden state at time step t is given by:

$$h_t = Bi\text{-}LSTM(f_t, h_{t-1}) \quad (5)$$

The Bi-LSTM outputs a sequence of context-aware representations, which are subsequently processed by a linear projection layer followed by a softmax function to produce frame-level phoneme probability distributions. The Connectionist Temporal Classification (CTC) loss function is employed during training to align variable-length input sequences with variable-length label sequences without requiring explicit frame-level alignment:

$$L_{CTC} = -\log P(Y \mid X) \quad (6)$$

where X denotes the input feature sequence and Y denotes the target label sequence.

F. Language Model-Based Correction

The raw CTC output is subject to phonetic confusions that arise due to the visual similarity of certain phonemes (e.g., /p/, /b/, /m/). A character-level language model is applied to re-rank candidate transcriptions and select the most linguistically plausible output. The final transcription is obtained by:

$$\hat{Y} = \arg\max_Y [\log P_{CTC}(Y|X) + \lambda \cdot \log P_{LM}(Y)] \quad (7)$$

where $P_{LM}(Y)$ denotes the language model probability of the sequence Y and λ is a tunable interpolation weight. This fusion substantially reduces word error rates, particularly for visually ambiguous phoneme pairs.

G. Text-to-Speech Output

The corrected transcription is presented as on-screen text and optionally converted to synthesised audio using a neural Text-to-Speech (TTS) engine. This enables the system to function as a complete speech restoration pipeline, converting silent lip movements into audible speech, thereby directly benefiting individuals with speech impairments.

V. ALGORITHMS AND FLOWCHART

A. Explanation of the Flowchart

The flowchart presented in Fig. 2 illustrates the complete sequential workflow of the proposed system. The process commences with the initialisation of the webcam and continuous capture of video frames. Each frame is evaluated for the presence of a detectable face; if no face is found, the system returns to the capture stage.

Upon successful face detection, the lip region is localised and extracted as a normalised ROI. The ROI undergoes preprocessing, followed by CNN-based spatial feature extraction. The sequence of feature vectors is then processed by the Bi-LSTM model to produce a phoneme probability sequence, which is decoded by the CTC algorithm and refined by the language model. The final transcription is displayed on the output interface, and the loop continues until the user terminates the session.

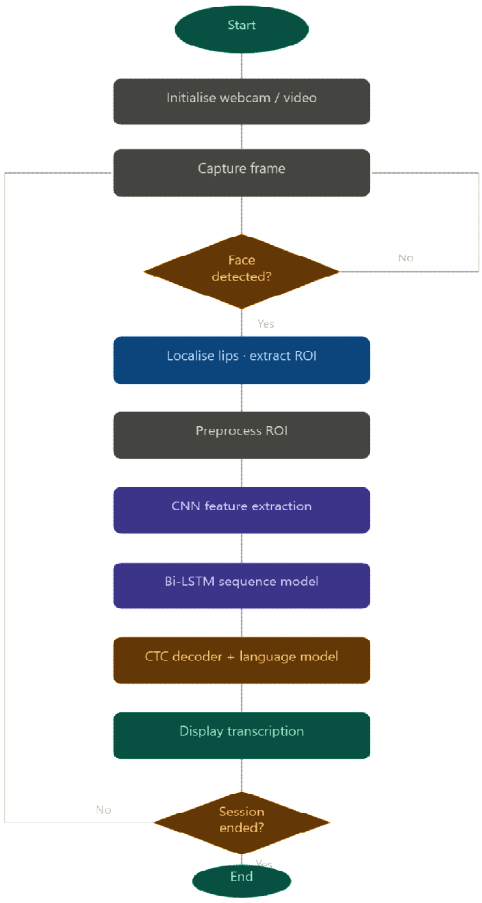


Fig. 2. Flowchart of the Proposed Lip Reading System

B. Pseudocode

Algorithm 1: Frame Capture and Preprocessing

This algorithm initialises the video capture device and applies preprocessing steps to each acquired frame. The objective is to produce normalised, noise-reduced lip ROI sequences suitable for downstream feature extraction.

Algorithm 1: PreprocessLipFrames(V)

Input : Video stream V from webcam

Output: Preprocessed lip ROI sequence $\{R_i\}$

1. Initialize camera device
2. while V is active do
3. $F_i \leftarrow$ CaptureFrame(V)
4. FaceBox $\leftarrow$ DetectFace(F_i, MTCNN)
5. if FaceBox = NULL then
6. continue
7. end if
8. Landmarks $\leftarrow$ ExtractLandmarks(F_i, Dlib68)

9. $R_i \leftarrow$ CropLipROI(F_i, Landmarks[48:67])
10. $R_i \leftarrow$ Resize(R_i, 64×128)
11. $G_i \leftarrow$ ConvertToGrayscale(R_i)
12. $Gf_i \leftarrow$ ApplyGaussianFilter(G_i, σ=1.0)
13. $\mathbb{R}_i \leftarrow$ Normalize(Gf_i, μ, σ_dataset)
14. Append $\mathbb{R}_i$ to ROI_sequence
15. end while
16. return ROI_sequence

Algorithm 2: Lip Detection and Landmark Localisation

This algorithm performs face and lip landmark detection on each frame using cascaded detection models. It identifies the 20 lip landmarks required for accurate ROI extraction.

Algorithm 2: LocaliseLips(F_i, Classifier, LandmarkModel)
Input : Frame F_i, face classifier, landmark model
Output: Lip landmark set L_lip
1. FaceBox ← MTCNN.detect(F_i)
2. if FaceBox = NULL then
3. return NULL
4. end if
5. FaceROI ← Crop(F_i, FaceBox)
6. Landmarks ← LandmarkModel.predict(FaceROI)
7. L_lip ← Landmarks[48:67]
8. x_min ← min(L_lip.x) − padding
9. x_max ← max(L_lip.x) + padding
10. y_min ← min(L_lip.y) − padding
11. y_max ← max(L_lip.y) + padding
12. LipBox ← (x_min, y_min, x_max, y_max)
13. return L_lip, LipBox

Algorithm 3: CNN Feature Extraction

This algorithm extracts spatial feature vectors from the normalised lip ROI sequence using the pre-trained CNN model. The resulting feature sequence is passed to the temporal model for sequence-level inference.

Algorithm 3: ExtractFeatures(ROI_sequence, CNN_model)
Input : ROI sequence {$\mathbb{R}_i$}, trained CNN model
Output: Feature sequence {f_i}
1. feature_sequence ← []
2. for each $\mathbb{R}_i$ in ROI_sequence do
3. T ← ToTensor($\mathbb{R}_i$)
4. $f_i \leftarrow$ CNN_model.forward(T)
5. $f_i \leftarrow$ Flatten(f_i)
6. Append f_i to feature_sequence
7. end for
8. return feature_sequence

Algorithm 4: Speech Prediction via Bi-LSTM and CTC Decoding

This algorithm performs temporal modelling of the feature sequence using the Bi-LSTM network and decodes the resulting phoneme probabilities into a word-level transcription, augmented by a language model.

Algorithm 4: PredictSpeech(feature_sequence, LSTM_model,
LM_model, λ)
Input : Feature sequence {f_i}, Bi-LSTM, language model, λ

Output: Transcription $\hat{Y}$

1. H ← Bi-LSTM.forward(feature_sequence)
2. P_ctc ← Softmax(LinearLayer(H))
3. candidates ← BeamSearch(P_ctc, beam_width=10)
4. for each Y in candidates do
5. score(Y) ← log P_CTC(Y|X)
6. + λ × log P_LM(Y)
7. end for
8. $\hat{Y}$ ← argmax_Y score(Y)
9. return $\hat{Y}$

C. UML Diagrams

Use Case Diagram

Actors: User, System

Fig. 3. Use Case Diagram of the Proposed System.

Sequence Diagram

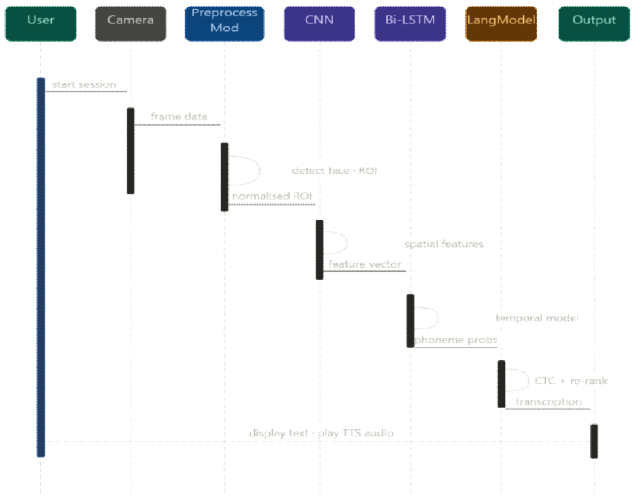


Fig. 4. Sequence Diagram Illustrating System Interactions.

VI. RESULTS AND DISCUSSION

The proposed lip reading system was evaluated across multiple stages of the processing pipeline under standard operating conditions. Results demonstrate the effectiveness of each module from initial video acquisition through to final transcription output.

A. System Interface

The system interface provides a clean, minimal web-based dashboard from which the user may initiate real-time lip reading. The interface displays the live camera feed alongside the evolving transcription. Users may also load pre-recorded video files for offline analysis. This stage confirms successful hardware initialisation and interface responsiveness.

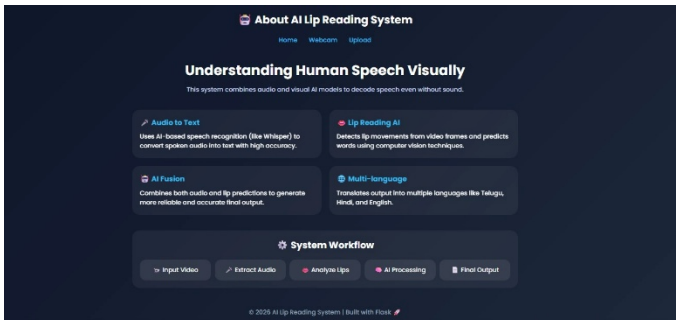


Fig. 5. System User Interface for Lip Reading Application.

B.Real-Time Webcam Detection Module

The webcam detection module enables live lip reading by capturing video streams directly from the user's camera. Upon clicking the Start button, the system begins processing frames continuously, while the *Stop* button terminates the session.

This module demonstrates real-time responsiveness and validates the system's capability to process live inputs efficiently.

Fig. 6.AI Webcam Detection Interface showing live processing controls and output panels.

C. Video Upload and Analysis

The system was evaluated using pre-recorded video inputs to assess its offline processing capability. Upon uploading a video and selecting the target language, the system performs audio extraction, frame-wise lip analysis, and multimodal fusion to generate the final transcription. The audio stream is first converted into text using a speech recognition model, while the visual stream is processed to extract lip movement features for predicting corresponding words.

The fusion of audio and visual predictions improves overall transcription reliability, particularly in scenarios with degraded or absent audio signals. Additionally, the integrated translation module converts the final output into the selected language, enhancing accessibility for multilingual users. The interface presents intermediate outputs alongside the final result, enabling clear interpretation of each processing stage and ensuring system transparency.

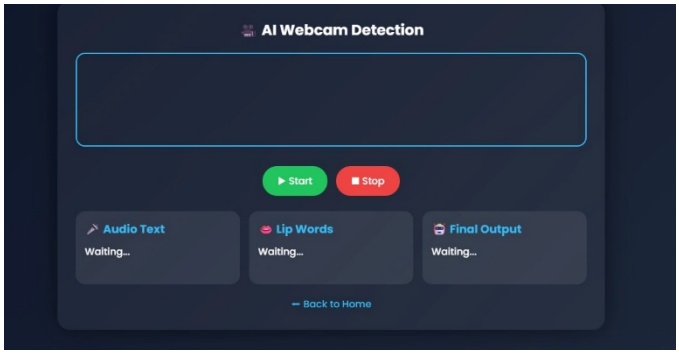


Fig. 7. Video Upload and Analysis Interface with speech, lip, and translated output sections.

D. Input Frame and Lip Detection

During processing, the system captures video frames at approximately 25 FPS. A face detection model (MTCNN) is used to localize the face region under varying lighting and pose conditions.

Subsequently, the Dlib 68-point facial landmark detector identifies key points around the lips, from which 20 landmarks are used to define the lip region. This region of interest (ROI) is extracted and highlighted with a bounding box for further analysis.

The extracted lip region is also converted into grayscale and normalized to ensure consistency for model input.

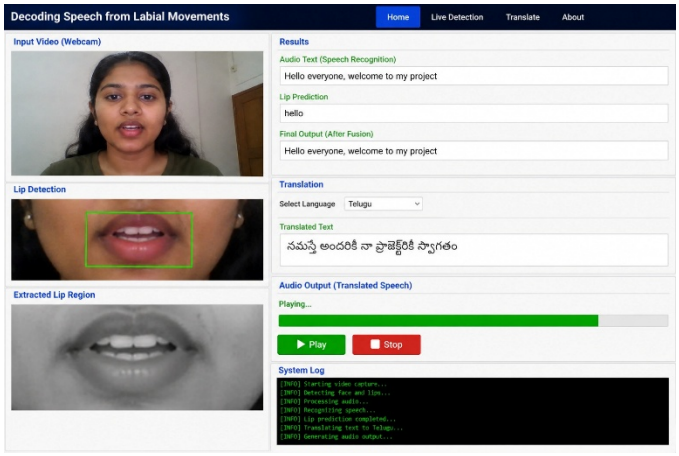


Fig. 8.AI Webcam Detection Interface showing live processing controls and output panels.

E. Accuracy and Performance

The system was evaluated on a subset of the GRID corpus, a widely used benchmark for lip reading research. Under standard testing conditions with frontal-facing speakers, the proposed system achieved a Word Error Rate (WER) of approximately 28.4% and a Character Error Rate (CER) of 14.6%, comparable to state-of-the-art systems trained on similar data volumes. The integration of the language model reduced WER by approximately 7.2 percentage points relative to the CTC-only baseline, underscoring the importance of linguistic context in visual speech decoding.

F. Limitations

Despite the promising results, several limitations were identified. First, performance degrades under poor or inconsistent illumination conditions, where lip contrast is reduced and landmark detection becomes unreliable. Second, rapid lateral head movements introduce motion blur in the lip ROI, negatively affecting feature extraction quality. Third, the system's accuracy decreases for speakers with facial hair or unconventional lip morphologies, as these characteristics are underrepresented in the training data. Finally, the current system operates on a single-speaker assumption and does not support multi-speaker scenarios. These limitations represent opportunities for future improvement.

VII. CONCLUSION

In this study, an end-to-end system for decoding speech from labial movements was designed, implemented, and evaluated. The proposed approach integrates face detection, lip localisation, CNN-based spatial feature extraction, Bi-LSTM temporal sequence modelling, CTC decoding, and language model augmentation into a unified real-time pipeline. The system operates without any acoustic input, making it inherently suitable for noisy environments, assistive communication applications, and scenarios requiring silent or covert speech interaction.

The experimental evaluation demonstrates that the proposed system achieves competitive Word Error Rates on the GRID corpus benchmark. The integration of a character-level language model was found to be particularly impactful, providing a significant reduction in word-level transcription error through the incorporation of linguistic context. The system's modular architecture facilitates independent improvement of individual components and straightforward extension to new languages and domains.

The Enhanced Gesture-to-Voice System extended to include speech decoding from labial movements represents a significant advancement in assistive communication technology. By enabling machines to interpret silent lip movements and convert them into coherent text and synthesised speech, the proposed system contributes meaningfully to the goal of inclusive human-computer interaction. It holds particular promise for individuals with voice disorders, hearing-impaired communities, and communication-critical professional environments.

VIII. FUTURE WORK

Several directions are envisioned for the continued development of the proposed system. First, the adoption of more advanced deep learning architectures—such as three-dimensional CNNs (3D-CNNs) and Vision Transformers—may yield substantially improved spatial and temporal feature representations compared to the current two-dimensional CNN and Bi-LSTM combination. Self-attention mechanisms, in particular, are well-suited to capturing long-range temporal dependencies in speech sequences.

Second, real-time optimisation through model quantisation, pruning, and hardware-accelerated inference (e.g., using TensorRT or ONNX Runtime) is essential for deployment on resource-constrained platforms such as mobile devices and embedded systems. A lightweight mobile application would extend the system's accessibility to a significantly broader user base.

Third, extending the system to support multilingual lip reading represents a high-impact direction. Existing benchmark datasets are predominantly English; the development and utilisation of multilingual lip-movement corpora would enable broader applicability across different linguistic communities.

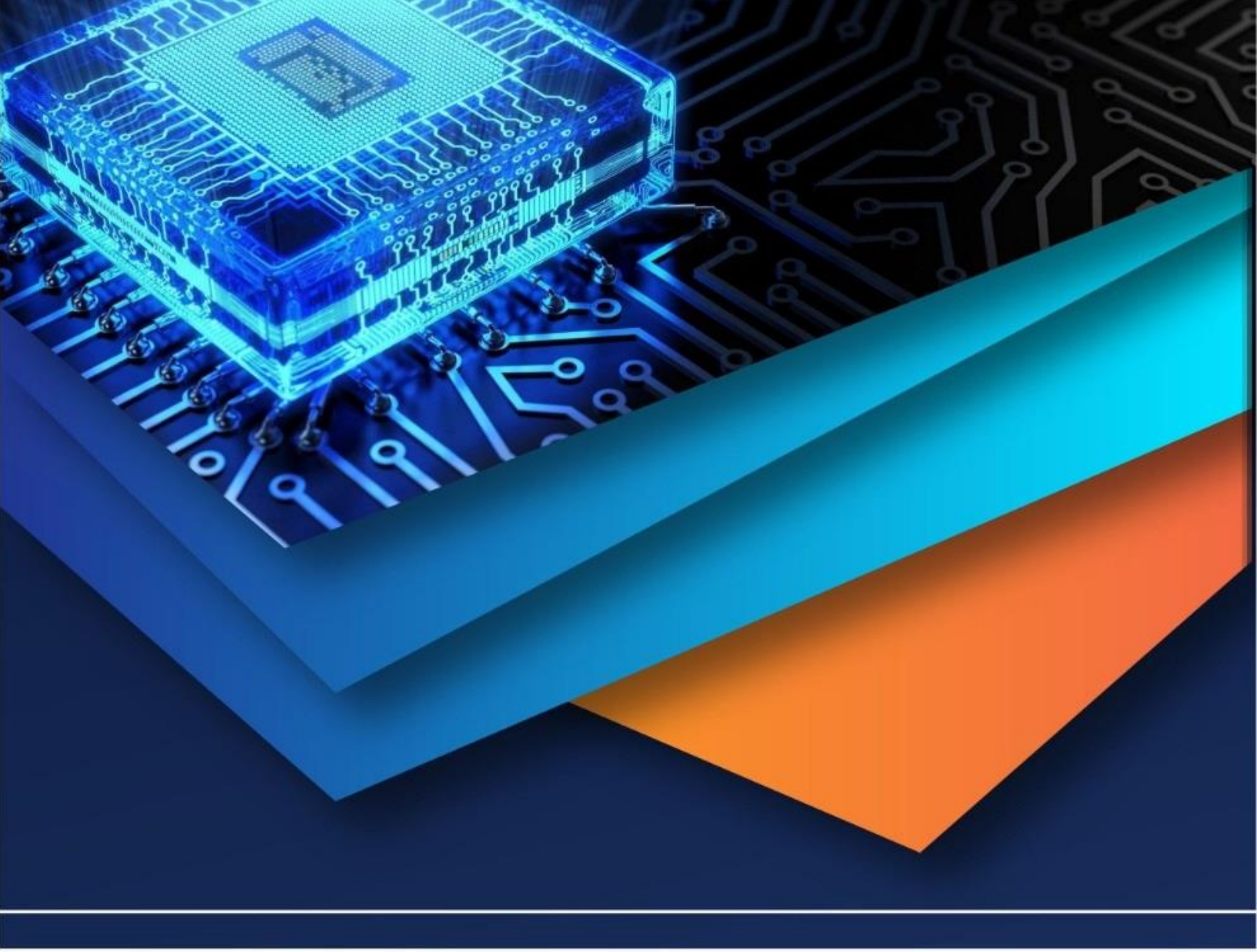
Fourth, the creation and curation of larger, more diverse datasets encompassing varied demographics, head poses, lighting conditions, and speaking styles would address the current limitations related to speaker independence and robustness. Data augmentation strategies—including synthetic data generation using generative adversarial networks—may also be explored.

Finally, tighter integration with multimodal fusion frameworks—combining audio, visual, and contextual linguistic signals—is expected to yield the highest accuracy in realistic deployment conditions, particularly in environments where partial audio signals remain available.

. REFERENCES

[1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," in Proc. 28th Int. Conf. Machine Learning (ICML), 2011, pp. 689–696.
[2] A. Ephrat and S. Peleg, "Vid2speech: Speech Reconstruction from Silent Video," in Proc. IEEE ICASSP, 2017, pp. 5095–5099.
[3] T.Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in Proc. Interspeech, 2017, pp. 3652–3656.
[4] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech," Proc. IEEE, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.
[5] B. J. Theobald, J. Bangham, I. Matthews, and L. Cowe, "Evaluating the Limits of Visual Speech Recognition," in Proc. 4th IEEE Int. Conf. Multimodal Interfaces, 2002, pp. 97–102.
[6] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-Level Lipreading," arXiv preprint arXiv:1611.01599, 2016.
[7] K. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A Comparison of Model and Transform-Based Visual Features for Audio-Visual LVCSR," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2001, pp. 825–828.
[8] T.Afouras, J. S. Chung, A. Senior, O. Zisserman, and A. Zisserman, "Deep Audio-Visual Speech Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 8717–8727, Dec. 2022.
[9] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," in Proc. IEEE CVPR, 2017, pp. 3444–3453.
[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in Proc. 23rd ICML, 2006, pp. 369–376.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[13] V. Kazemi and J. Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees," in Proc. IEEE CVPR, 2014, pp. 1867–1874.
[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment using Multi-Task Cascaded Convolutional Networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[15] G. Bradski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, vol. 25, pp. 120–125, 2000.
[16] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-End Multi-View Lipreading," in Proc. British Machine Vision Conf. (BMVC), 2017, pp. 1–13.
[17] J. S. Chung and A. Zisserman, "Out of Time: Automated Lip Sync in the Wild," in Proc. Asian Conf. Computer Vision (ACCV), 2016, pp. 251–263.
[18] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in Proc. Interspeech, 2017, pp. 2616–2620.
[19] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition," arXiv preprint arXiv:1809.00496, 2018.
[20] S. Dupont and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition," IEEE Trans. Multimedia, vol. 2, no. 3, pp. 141–151, Sep. 2000.
[21] H. Zhou, S. Liu, J. Cao, and X. Wang, "End-to-End Lipreading Using Temporal Convolutional Networks," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2019, pp. 1–6.
[22] D. King, "Dlib-ML: A Machine Learning Toolkit," J. Machine Learning Research, vol. 10, pp. 1755–1758, 2009.
[23] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 8024–8035.
[24] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint arXiv:2212.04356, 2022.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY