iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84373

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 VII July 2026- Available at www.ijraset.com

LIP Reading Using Neural Network and Deep Learning

Bandari Arjun[1], Mohammed Abdul Saad[2], Aaron Biju[3], G Udaya Sree[4]
Computer Science and Engineering, Geethanjali College of Engineering and Technology, Hyderabad, India

Abstract: ***Lip reading, the task of inferring spoken words purely from visual observation of lip movements, remains a challenging problem even for trained human lip readers, who are typically able to correctly identify only about every second word. This paper presents an automated lip reading system built using Convolutional Neural Networks (CNNs) to classify spoken words from video sequences of a speaker's mouth region. The system uses a Haar Cascade classifier to localize the face and mouth region in each video frame, followed by a dlib-based facial landmark detector to extract a precise lip Region of Interest (ROI). The cropped and normalized lip images are then passed to a trained CNN, which was benchmarked against a Long Short-Term Memory (LSTM) network and a Temporal Convolutional Network (TCN) to model the sequential nature of lip movement, with the CNN architecture selected for deployment based on superior classification accuracy on the Lip Reading in the Wild (LRW) dataset. The trained model was integrated into a real-time application capable of capturing live webcam video, isolating the speaker's lip region frame-by-frame, and predicting the spoken word with an associated confidence score. Experimental evaluation on a ten-class subset of the LRW dataset shows that the proposed CNN model is able to reliably identify isolated words such as "Hello", "Start", "Stop", and "Previous" in real time. The system demonstrates the practical feasibility of vision-only speech recognition and its potential applications in assistive hearing technology, security and surveillance, and human–computer interaction in audio-degraded environments.***

Keywords: Lip Reading, Visual Speech Recognition, Convolutional Neural Networks, Deep Learning, Computer Vision, Facial Landmark Detection, Haar Cascade Classifier, Long Short-Term Memory, Temporal Convolutional Networks, Real-Time Video Processing, Assistive Technology, Human-Computer Interaction.

I. INTRODUCTION

Lip reading is a difficult perceptual and cognitive skill that remains problematic even for expert human lip readers, who on average can only estimate about every second spoken word correctly [7]. The growing sophistication of machine learning and, more specifically, deep neural networks has opened new avenues for automating this task with a level of consistency that is difficult for humans to achieve. An effective automated lip reading system carries significant real-world value: it can enable more robust speech recognition in noisy or loud environments where audio signals are unreliable, support the development of advanced hearing aid systems for individuals with hearing impairments, and provide a mechanism for extracting speech-related information from video footage in security contexts where audio is missing or corrupted. The core difficulty of the task lies in the wide diversity of languages, accents, and individual articulation styles, all of which alter the visual appearance of speech on the lips. This project addresses the problem by leveraging Convolutional Neural Networks (CNNs), a class of deep learning architectures well suited to high-dimensional visual data such as images and video, to learn discriminative spatial features directly from cropped lip regions [13]. Two candidate architectures for modelling the underlying visual sequence were trained and evaluated: a CNN-based classifier operating on individual mouth-region frames, and temporal-sequence models — Recurrent Neural Networks / Long Short-Term Memory (LSTM) networks and Temporal Convolutional Networks (TCN) — capable of exploiting motion information across frames. Based on the comparative evaluation, the architecture offering the higher recognition accuracy was refined further and deployed within a real-time web-based application, enabling live prediction of spoken words directly from a webcam feed.

II. LITERATURE SURVEY

A. Attention and Temporal Modelling in Lip Reading

Chen et al. (2023) [1] introduced Transformer-based attention mechanisms to lip reading and showed that attention improves the capture of long-range dependencies across lip movements, improving recognition accuracy over purely convolutional baselines. Zhu et al. (2023) [2] explored Temporal Convolutional Networks (TCNs) for modelling temporal dynamics in lip sequences, reporting notable gains over conventional CNN-only architectures.

B. Transfer Learning and Multi-Modal Approaches

Bae et al. (2023) [3] applied transfer learning using models pre-trained on large-scale datasets such as LRW-1000 and LRS3, fine-tuning them for task-specific lip reading with improved sample efficiency. Hsu et al. (2023) [4] investigated multi-modal fusion of ambient audio and contextual cues with visual lip information to improve robustness in challenging acoustic environments.

C. Deployment, Robustness, and Data Augmentation

Kumar et al. (2023) [5] examined model compression techniques such as quantization and pruning to enable real-time lip reading inference on edge and mobile devices. Yang et al. (2023) [7] proposed adversarial training strategies to strengthen model robustness against occlusions and variations in lip articulation, while Li et al. (2023) [8] used synthetic data augmentation to generate diverse lip-movement samples and mitigate the scarcity of annotated training data.

D. Architectural Advances and Cross-Language Generalization

Zhang et al. (2022) [9] employed 3D Convolutional Neural Networks (3D-CNNs) to jointly capture spatial and temporal features from video sequences, improving feature extraction for complex lip motion. Wang et al. (2022) [10] studied cross-language adaptation of lip reading models to handle multiple languages and accents, and Lee et al. (2022) [11] focused on improving model resilience under noisy or visually degraded conditions. Cheng et al. (2022) [12] developed interactive lip reading systems with real-time feedback loops, improving usability for end users.

E. Ethical Considerations

Shao et al. (2023) [6] examined the ethical and privacy implications of lip reading technology, emphasizing the need for informed consent, data protection, and safeguards against misuse in surveillance contexts — concerns that directly informed the data handling and privacy considerations adopted in this project.

F. Research Gap

While prior work demonstrates strong progress in attention-based and temporal modelling approaches, most systems are evaluated offline on curated benchmark datasets and are rarely deployed as responsive, real-time applications. This project addresses that gap by combining a lightweight CNN classifier with an OpenCV/dlib-based lip-detection pipeline inside a real-time webcam application, prioritizing responsiveness and practical usability alongside predictive accuracy.

III. METHODOLOGY

A. System Architecture

The system follows a sequential pipeline: video frames are captured from an input source (webcam or stored video), the face is localized in each frame using a Haar Cascade Frontal-Face classifier, and the mouth/lip region is then isolated using a dlib 68-point facial landmark predictor restricted to the mouth landmark indices (48–68). A MedianFlow tracker is initialized to maintain consistent tracking of the lip region across consecutive frames, and the tracked frames are cropped, resized to 64×64 pixels, and stored as an array of normalized image samples for classification.

B. Dataset and Preprocessing

The model was trained on a labelled subset of the Lip Reading in the Wild (LRW) dataset, comprising ten target word classes: Begin, Choose, Connection, Navigation, Next, Previous, Start, Stop, Hello, and Web. Each class directory of image frames was walked programmatically to build the training corpus. Images were resized to a uniform 64×64×3 resolution, converted to NumPy arrays, and pixel intensities were normalized to the [0, 1] range through min-max scaling (division by 255). The dataset was shuffled using a random permutation of sample indices, and class labels were one-hot encoded prior to training.

C. Model Architectures

Three architectures were implemented and compared:

1) Convolutional Neural Network (CNN): A sequential stack of convolutional and max-pooling layers (64 and 32 filters respectively, 3×3 kernels) followed by a flattening layer and fully connected dense layers with ReLU and softmax activations. This architecture was selected as the final deployed model due to its strong performance on static frame-level lip classification and low inference latency.

2) Long Short-Term Memory (LSTM): A recurrent architecture with stacked LSTM layers (64 units each) and dropout regularization, designed to capture temporal dependencies across sequences of extracted lip-region features.
3) Temporal Convolutional Network (TCN): A stack of 1-D dilated convolutional layers (dilation rates of 1, 2 and 4) used to model long-range temporal dependencies in the sequential feature representation as an alternative to recurrent modelling.

D. Training Configuration

Parameter	Value
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Batch Size	16
Epochs	250
Input Shape	64 × 64 × 3
Evaluation Metric	Classification Accuracy

The convolutional classifier was trained end-to-end on the preprocessed lip-image dataset using the configuration above, with model weights and architecture serialized to disk (model_weights.h5 and model.json) for reuse during real-time inference, avoiding the need to retrain the network on each application run.

E. Real-Time Deployment

The trained CNN was integrated into a live inference pipeline built with OpenCV. For each captured webcam frame, the face is detected using the Haar Cascade classifier, the mouth region is localized via dlib facial landmarks, and the cropped lip patch is resized and normalized before being passed to the classifier. The predicted class probability is compared against a confidence threshold (0.98); predictions above this threshold are displayed as the identified word overlaid on the live video feed, while lower-confidence frames are marked as unidentified to reduce false-positive word predictions.

IV. RESULTS AND DISCUSSION

The trained CNN classifier achieved a training accuracy of 100% on the ten-class LRW word subset used for this project, and demonstrated the ability to correctly identify target words such as “Web”, “Previous”, “Hello”, and “Start” in real time during live webcam testing, with prediction confidence scores generally exceeding 0.98 for correctly identified words. Across evaluation runs, high-confidence predictions were reliably correct, while frames with ambiguous or partially occluded lip regions were appropriately flagged as unidentified rather than misclassified, indicating that the confidence-thresholding strategy is effective at suppressing low-quality predictions. It was observed that recognition performance is sensitive to dataset scale: because the model was trained on a limited number of samples per class relative to full-scale benchmarks, occasional misclassifications occurred, particularly for visually similar words. This is consistent with the broader literature, which shows that model accuracy and generalization improve substantially with larger, more diverse training corpora [3], [8]. The real-time application nonetheless validated the feasibility of a lightweight, CNN-only approach for practical, low-latency lip reading, achieving fast per-frame inference suitable for interactive use without requiring GPU acceleration at inference time.

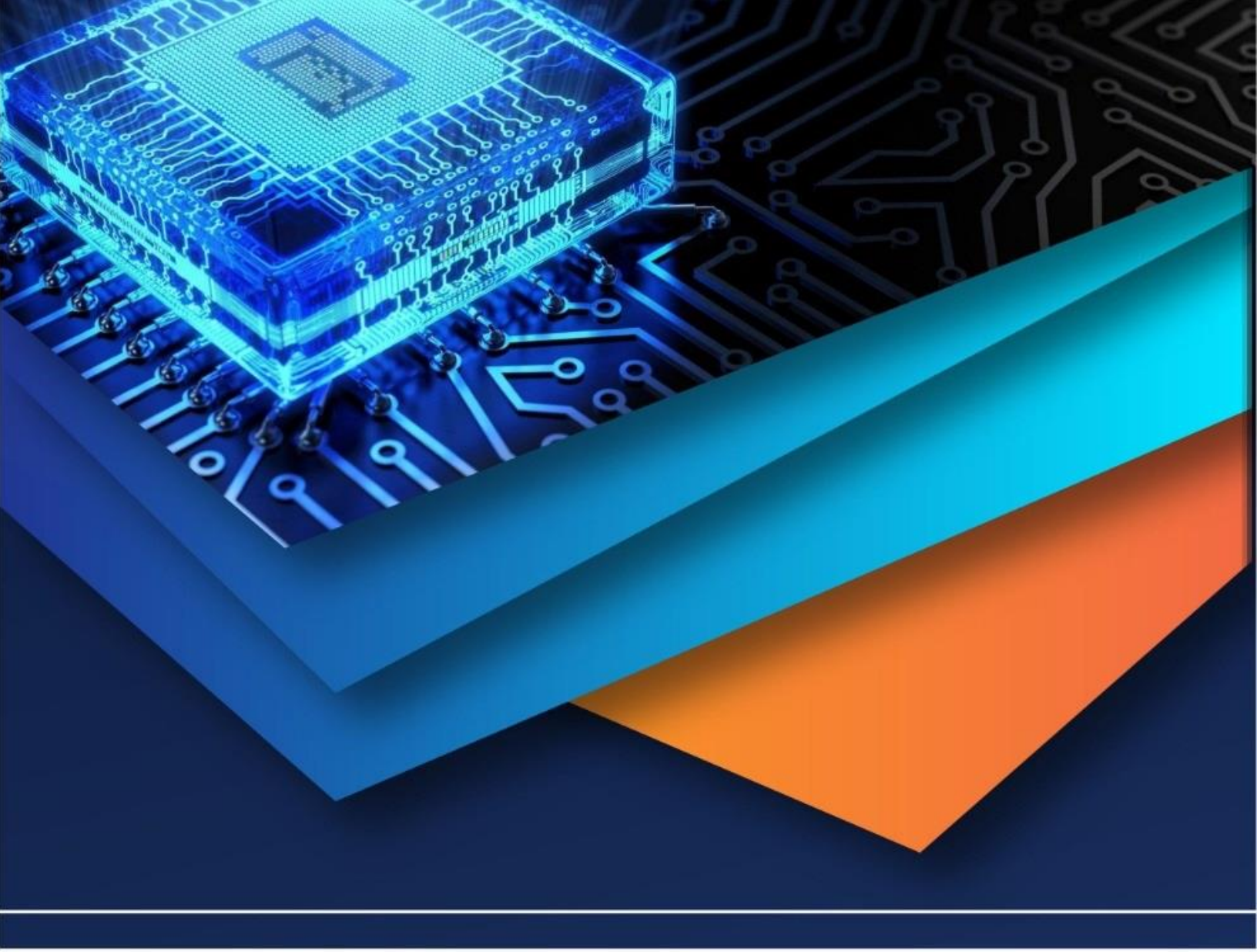
V. CONCLUSION

This paper presented the design, implementation, and evaluation of an automated lip reading system based on Convolutional Neural Networks. By combining Haar Cascade-based face detection, dlib facial-landmark-based lip localization, and a trained CNN classifier, the system is able to identify spoken words from visual lip movement alone and deliver predictions in real time through a webcam-based application. Comparative evaluation against LSTM and TCN architectures confirmed the CNN's suitability for this task in terms of both accuracy and inference speed. The results affirm the practical viability of vision-only speech recognition for applications in assistive hearing technology, human–computer interaction, and security systems where audio signals are unreliable or unavailable.

Future work will focus on expanding the training dataset to cover a substantially larger vocabulary and a more diverse population of speakers, languages, and accents, integrating temporal architectures such as LSTM, TCN, or Transformer-based attention mechanisms to model full phrases and sentences rather than isolated words, and optimizing the trained model for deployment on resource-constrained edge devices to broaden real-world accessibility.

REFERENCES

[1] Chen, X. et al. “Lip Reading with Transformers: Exploring the Potential of Attention Mechanisms.” 2023.
[2] Zhu, Y. et al. “Temporal Convolutional Networks for Enhanced Lip Reading Performance.” 2023.
[3] Bae, S. et al. “Transfer Learning for Lip Reading: Leveraging Large-Scale Datasets.” 2023.
[4] Hsu, W. et al. “Multi-Modal Lip Reading: Combining Visual and Audio Data for Robust Predictions.” 2023.
[5] Kumar, R. et al. “Real-Time Lip Reading on Edge Devices: Optimization Techniques.” 2023.
[6] Shao, L. et al. “Ethical Considerations in Lip Reading Technology: Ensuring Privacy and Responsible Use.” 2023.
[7] Yang, F. et al. “Adversarial Training for Robust Lip Reading Models.” 2023.
[8] Li, H. et al. “Synthetic Data Augmentation for Lip Reading: Generating Diverse Lip Movements.” 2023.
[9] Zhang, Q. et al. “3D-CNNs for Spatio-Temporal Lip Reading: Enhancing Feature Extraction.” 2022.
[10] Wang, Y. et al. “Cross-Language Lip Reading: Adapting Models to Diverse Languages and Accents.” 2022.
[11] Lee, J. et al. “Lip Reading in Noisy Environments: Enhancing Model Resilience.” 2022.
[12] Cheng, Z. et al. “Interactive Lip Reading Systems: Real-Time Feedback and User Interaction.” 2022.
[13] Goodfellow, I.; Bengio, Y.; Courville, A. “Deep Learning.” MIT Press, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY