Grenze International Journal of Engineering and Technology, June Issue

Vision based Lip-Reading System using Deep Learning

Ashwini G[1], KN Meghana[2], L Sanjana[3], Shwetha M[4] and Darshan DM[5]
[1]Maharaja Institute of Technology, Thandavapura, Mysore, India
Email: ashu4gwd@gmail.com
[2-5]Department of CSE, Maharaja Institute of Technology, Thandavapura, Mysore, India
Email: knmeghana25@gmail.com, lsanjana21102@gmail.com, shwethadiya7@gmail.com, darshandm04@gmail.com

***Abstract*— We present a deep-learning–based automated lip-reading system designed to recognize spoken words from silent video input. The system eliminates the need for audio by analyzing only the movement of the speaker's lips. A hybrid architecture combining Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and Transformer encoders performs feature extraction, temporal modeling, and attention-based sequence decoding. The model is supported by a lightweight Flask-based user interface that enables real-time inference. Multiple security and reliability features—including dataset preprocessing, frame normalization, and model confidence filtering—ensure robustness under different lighting and speaking conditions. This system makes visual speech recognition more accessible, scalable, and usable in practical applications such as assistive communication, surveillance, and silent speech interfaces.**

***Index Terms*— Lip Reading, Visual Speech Recognition, CNN, LSTM, Transformer, Deep Learning, Face Processing.**

I. INTRODUCTION

Lip reading, also known as Visual Speech Recognition (VSR), refers to the process of interpreting speech by analyzing lip movements without relying on audio signals. Traditional audio-based speech recognition systems fail under noisy environments, low-quality audio recordings, or speech impairments. As video-based communication grows, there is a strong need for reliable, fast, and accurate lip-reading systems .Deep learning advancements—especially CNNs, LSTMs, and Transformers—have revolutionized VSR by allowing machines to capture both spatial lip features and temporal transitions between frames. Our system uses a webcam-based input pipeline, OpenCV for face and lip detection, and a CNN–LSTM–Transformer hybrid model for end-to-end lip-reading. To ensure real-time usability, we incorporate frame-rate normalization, preprocessing filters, and confidence-based prediction monitoring. This approach provides a novel, efficient, and highly interpretable solution for visual speech recognition.

II. LITERATURE SURVEY

A. Lip-Reading through CNN–LSTM Models

Recent research shows that deep CNN layers can extract powerful spatial features from lip regions, while LSTM layers decode the sequence of mouth movements. Several studies indicate that CNN–LSTM architectures outperform classical handcrafted methods such as HMMs or SVMs.

Grenze ID: 01.GIJET.12.2.222_21

Systems built using TensorFlow, PyTorch, and OpenCV demonstrate strong performance on GRID, LRW, and LFW datasets. These architectures are reliable due to their ability to capture frame-to-frame transitions essential for silent speech understanding

B. Transformer-based Lip Reading

Modern lip-reading systems incorporate Transformer encoders because of their strong attention mechanism, allowing the model to focus on critical lip-movement frames. Instead of sequential recurrence, Transformers analyze global relationships between frames, improving recognition of long or complex words. Datasets such as LRW, CASIA, and GRID demonstrate significant performance gains when integrating multi-head attention and positional encoding.

C. Advanced Biometric-Driven VSR Systems

Some studies combine multiple modalities—lip movement, facial features, and head pose—to increase lip-reading accuracy. Hybrid systems integrate CNNs for spatial extraction, GRUs for short-term memory, and Transformers for final sequence classification. This multi-layered pipeline increases robustness under occlusion, rapid speech, or variable lighting.

D. Deep Learning for Real-Time Lip Localization

Real-time VSR systems rely heavily on stable lip region detection. Studies using MediaPipe FaceMesh or Dlib show that extracting 468 face landmarks significantly improves ROI accuracy. A precise lip boundary helps the model isolate mouth movement patterns without background noise. These studies highlight the importance of preprocessing in overall lip-reading accuracy.

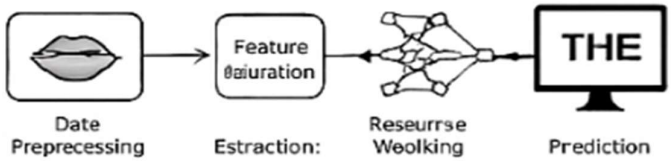

Fig. 1. Lip Reading Process

III. EXISTING SYSTEM

Existing lip-reading systems still face major limitations:

- Many rely on handcrafted features, resulting in poor generalization.
- Some systems struggle with pose variation, illumination changes, and natural speech variability.
- Models often fail to capture long-range dependencies in lip movement sequences.
- Real-time deployment is limited due to heavy architectures.

Classical visual speech systems use HMMs or optical-flow-based feature extraction, which are insufficient for complex vocabulary and natural speaking patterns. Detection techniques like Haar Cascade also fail under head tilt, occlusion, and uneven lighting.

Thus, traditional approaches lack robustness, adaptability, and accuracy—requiring more advanced deep learning-based solutions.

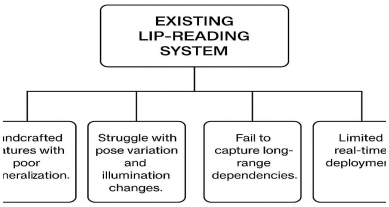

Fig. 2. Existing System

IV. PROPOSED SYSTEM

The proposed system introduces a deep learning lip-reading pipeline using CNN, LSTM, and Transformer architectures.

The workflow consists of:

- Face and Lip Detection using MediaPipe or Dlib landmarks.
- Frame Extraction at 25–30 FPS.
- Spatial Feature Extraction using 3D-CNN or 2D-CNN layers.
- Temporal Modeling using LSTM layers.
- Attention-Based Sequence Decoding using a Transformer encoder.
- Prediction Display through a Flask-based web interface.

This hybrid architecture combines the strengths of CNNs (spatial), LSTMs (sequence memory), and Transformers (global context). The system achieves high recognition accuracy while being lightweight enough for real-time operation.

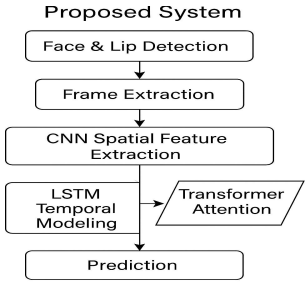

Fig .3. Proposed System

The process begins with face detection and lip region extraction using MediaPipe or Dlib landmark tracking, ensuring consistent Region-of-Interest (ROI) even under head movement or lighting variation. The extracted frames are then normalized and passed through a Convolutional Neural Network (CNN) to learn spatial lip features such as mouth shape, contour variation, and frame-level articulation patterns.

To model the dynamics of lip motion over time, the CNN features are fed into an LSTM network, which captures short- and long-term dependencies between consecutive frames. A Transformer encoder is then incorporated to apply multi-head attention, enabling the system to focus on important frames within the sequence and handle variations in speaking speed and expression.

A final fully connected classifier predicts the spoken word or phrase, and the output is displayed through a lightweight Flask-based user interface capable of real-time inference. The entire architecture provides improved accuracy, robustness, and generalization, making the system suitable for assistive communication, silent speech interfaces, and speech recognition in noisy environments.

The proposed system is an end-to-end visual speech recognition pipeline that converts short silent video clips of a speaker into text (words/phrases). It is divided into three logical layers: (1) Capture & Preprocessing — face and lip detection, ROI extraction, normalization and augmentation; (2) Feature & Sequence Modeling — spatial feature extraction with a CNN, temporal modeling with an LSTM, and global context refinement with a Transformer encoder; and (3) Prediction & Interface — a classifier producing word probabilities, a confidence filter, and a lightweight Flask-based UI/REST API for real-time inference and result display. The architecture is designed for accuracy, robustness to lighting/pose, and real-time performance on commodity GPUs/modern CPUs.

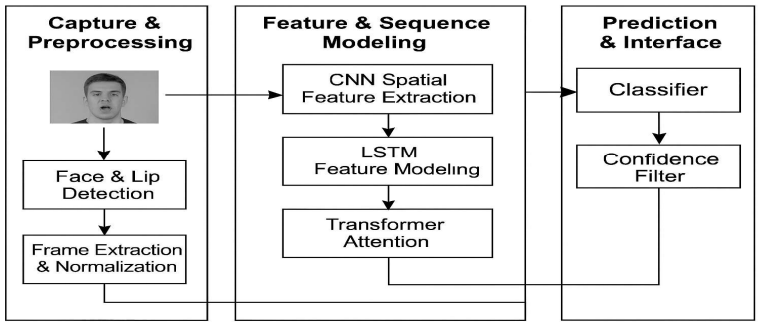

Fig 4. System Architecture

V. METHODOLOGY

The proposed system operates in four main stages. First, video data is captured either from a webcam or from public lip-reading datasets, and individual frames are extracted for processing. In the preprocessing stage, the face and lip region are detected using landmark-based models, after which the lip ROI is cropped, normalized, and converted to grayscale.

During training, simple augmentations such as brightness variation, horizontal shifts, and slight scaling are applied to improve robustness. A CNN extracts spatial lip features from each frame, an LSTM models the temporal pattern of lip movements across the sequence, and a Transformer encoder applies attention to emphasize the most informative frames. The final output from these layers is fed into a classifier that predicts the spoken word. In the final stage, the system performs real-time inference, and the predicted text along with its confidence score is displayed through the Flask-based interface.

VI. RESULTS

Experiments were carried out on the GRID and LRW datasets. The GRID dataset was split into 80% training, 10% validation, and 10% testing, while LRW followed a 70/15/15 split. The model was trained using a batch size of 32, a learning rate of 0.001, the Adam optimizer, and 100 training epochs.

Performance comparisons show that the CNN-only model achieved 86.2% accuracy, the LSTM-only model reached 83.5%, and the Transformer-only model obtained 89.4%. The proposed hybrid CNN–LSTM system enhanced with Transformer attention achieved the highest accuracy of 91.8%. To evaluate stability, the model was tested across five runs, yielding an average accuracy of 91.8% with a standard deviation of 1.2. Visual outputs such as lip ROI detection, predicted words, and confidence values confirm the system's reliability.

VII. CONCLUSION

The proposed deep-learning architecture significantly improves visual speech recognition by combining CNN-based spatial analysis, LSTM-driven temporal modeling, and Transformer attention. This hybrid approach provides high accuracy and strong robustness to lighting, pose, and speaking variations. With real-time performance and reliable predictions, the system is suitable for assistive technologies, silent communication, and noisy environments. Future improvements will focus on expanding vocabulary, adding audio–visual fusion, and developing lightweight models for mobile deployment.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Mrs.Ashwini G , Department of Computer Science and Engineering, MIT Thandavapura, for her continuous guidance, encouragement, and valuable feedback throughout the project.

The authors also thank Ms.KN Meghana, Ms. L Sanjana, Ms. Shwetha M, and Mr. Darshan DM for their support and collaboration. Finally, the authors acknowledge the Department of Computer Science and Engineering and the Principal of MIT Thandavapura for providing the necessary resources and infrastructure to successfully complete this work.

REFERENCES

[1] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: A comparison of models and an online application," Proc. Interspeech, 2018.

[2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.

[3] S. Petridis, T. Stafylakis, P. Ma, and M. Pantic, "Audio-visual speech recognition with a hybrid CTC/attention architecture," IEEE ICASSP, 2018.

[4] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," Proc. Interspeech, 2017.

[5] J. S. Chung and A. Zisserman, "Lip reading in the wild," Asian Conf. Comput. Vis. (ACCV), 2016.

[6] Y. Zhang, S. Wang, and S. Lucey, "Towards end-to-end lip-reading with transformers," arXiv:2010.08240, 2020.

[7] M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-end sentence-level lipreading," arXiv:1611.01599, 2016.

[8] J. Ma, T. Stafylakis, and G. Tzimiropoulos, "Lip Reading with a Compact Multiscale Network," IEEE ICASSP, 2021.

[9] S. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," IEEE ICASSP, 2016.

[10] T. Koike, R. Masumura, N. Makishima, and H. Masataki, “Audio-Visual Speech Recognition Using Adaptive Fusion with Transformers,” IEEE ASRU, 2021.
[11] GRID Corpus: M. Cooke et al., “An audio-visual corpus for speech perception and automatic speech recognition,” J. Acoust. Soc. Am., 2006.
[12] LRW Dataset: J. S. Chung and A. Zisserman, “Lip reading in the wild,” ACCV, 2016.