WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

AI-Based Lip Reading and Language Translation System

Chinnadamarlacheruvu Bhaskar [1], Dr. K. Chandrashekar [2]

1,3 P.G. Research Scholar, Dept. of MCA, Aurora Deemed University, Hyderabad, Telangana, India.
2Associate Professor, Dept. of CSE, Aurora Deemed To-Be University, Hyderabad, Telangana, India.
Email:[1]cbhaskar3002@gmail.com,[2] Chandracse@aurora.edu.in

A B S T R A C T

This project proposes an AI-based system capable of interpreting speech by analyzing lip movements from video input. The goal is to bridge communication gaps for individuals with hearing or speech impairments and to enable silent communication in noisy or confidential environments. The system uses advanced deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to process visual lip sequences and convert them to corresponding text. Additionally, the translated English text can be converted into a user-selected language such as Telugu or Hindi through neural machine translation. The application supports real-time video processing and provides an intuitive user interface. Python is used for the AI models and backend logic, while the frontend is developed using a simple GUI framework like Tkinter or a web-based interface.

Keywords: Lip Reading, Visual Speech Recognition, Deep Learning, CNN, RNN, Text Translation, Accessibility, Python.

1. Introduction

Speech recognition is a well-established field, but most existing systems rely heavily on audio input, which limits their usability in noisy environments or for individuals with hearing and speech impairments. Lip reading, also known as visual speech recognition, offers a promising alternative by interpreting speech through the analysis of lip movements. With the rapid growth of deep learning and computer vision techniques, lip reading systems have achieved significant accuracy in extracting textual information from silent videos. However, most existing solutions are limited to English language recognition and often lack real-time translation support for regional languages.

The AI-Based Lip Reading and Language Translation System addresses these gaps by integrating advanced computer vision, deep learning, and natural language processing. The system not only recognizes speech from lip movements but also translates the recognized text into multiple languages such as Telugu and Hindi. Built on deep learning models like Convolutional Neural Networks (CNN) and 3D CNNs for spatio-temporal feature extraction, coupled with recurrent layers such as LSTMs or GRUs, the system ensures accurate sequence modeling. For translation, state-of-the-art NLP models like MarianMT and Google Translate API are employed to provide multilingual support.

A Flask-based web interface enables user interaction with two primary modes: video upload for pre-recorded content and real-time live camera input for on-the-fly recognition and translation. By combining lip reading with multilingual translation, this project enhances accessibility, supporting individuals with hearing or speech challenges and enabling silent communication in noisy or private environments.

2. Literature Review

The following literature survey reviews key research papers relevant to AI-based lip reading and translation systems, comparing them with the proposed project.

1) Paper: "LIP READING USING CNN AND BI-LSTM" – DivyaPrabha et al., 2024

This paper proposes a lip reading system using Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) to recognize spoken words from lip movement sequences. It achieves a Word Accuracy (WA) of 83.4%, Character Accuracy (CA) of 95.84%, and a Word Error Rate (WER) of 16.6% on a custom dataset.

Comparison: The proposed project adopts a similar CNN + RNN/GRU architecture but extends to sentence-level recognition and multilingual translation, with a focus on real-time processing and GUI integration.

2) Paper: "The Research of Lip Reading Based on STCNN and ConvLSTM" – Zhu, 2020

This study uses Spatiotemporal CNNs (STCNN) and Convolutional LSTM (ConvLSTM) on the GRID corpus, achieving 6.7% CER and 13.6% WER.

Comparison: The proposed system also captures spatiotemporal features but prioritizes lightweight deployment and multilingual translation, enhancing practical applicability.

3) Paper: "A Lip Reading Method Based on 3D Convolutional Vision Transformer" – Wang et al., 2022

This paper combines 3D CNNs with Vision Transformers, achieving 88.5% accuracy on LRW and 57.5% on LRW-1000.

Comparison: The proposed system could incorporate transformer modules in the future but currently focuses on accessibility through multilingual support and a user-friendly interface.

4) Paper: "Speech Recognition Models Are Strong Lip-readers" – Prajwal et al., 2024

This paper adapts pretrained ASR models like Whisper for lip reading, achieving a WER of 24.3% on LRS3.

Comparison: The proposed system could leverage pretrained models like Whisper to enhance accuracy while maintaining its focus on translation and real-time GUI.

5) Paper: "LipNet: End-to-End Sentence-level Lipreading" – Assael et al., 2016

LipNet, the base paper, uses spatiotemporal CNNs, GRUs, and CTC loss, achieving 95.2% accuracy on the GRID dataset.

Comparison: The proposed system builds on LipNet's architecture, adding real-time processing, multilingual translation, and a GUI for broader applicability.

6) Paper: Biomedinformatics, 2024

This paper uses 3D CNNs and LSTMs for Greek word recognition, achieving 87.5% accuracy.

Comparison: The proposed system targets sentence-level recognition and multilingual translation, with real-time capabilities and a GUI.

7) Paper: "Deep Multimodal Lip Reading and Translation" – IJACSA, 2024

This paper integrates lip reading and translation using Transformers.

Comparison: The proposed system aligns with this multimodal approach but emphasizes real-time video input and a user-friendly interface.

8) Paper: "Enhancing Lip Reading: A Deep Learning Approach with Double CNN+RNN" – Electrical Systems, 2024

This paper uses a double CNN + RNN architecture with a translation component.

Comparison: The proposed system shares the translation focus but adds real-time processing and lightweight deployment.

Table 1: Comparison with Papers 1–4

Feature	**Paper 1**	**Paper 2**	**Paper 3**	**Paper 4**	**Our Project**
Architecture	CNN + Bi-LSTM	STCNN + ConvLSTM	3D CNN + Transformer	Visual Encoder + Whisper	CNN + RNN/GRU + CTC + NMT
Dataset Used	Custom	GRID	LRW / LRW-1000	LRS3	GRID / LRW (to be finalized)
Language Support	English only	English only	English only	English only	Multilingual (English → Telugu/Hindi)
Input Modality	Video-only	Video-only	Video-only	Video-only	Video-only
Output Type	Word-level	Word / Phrase	Word / Phrase	Sentence-level	Sentence + Translation
Real-time Capability	Yes (GUI)	No	No	No	Yes (Tkinter/Web Interface)
Translation Module	No	No	No	No	Yes
User Interface	Tkinter GUI	None	None	None	Tkinter / Web Interface

Feature	Paper 1	Paper 2	Paper 3	Paper 4	Our Project
Preprocessing Tools	OpenCV	OpenCV	OpenCV	AV-HuBERT	OpenCV + Face Alignment
Training Framework	TensorFlow / Keras	TensorFlow	PyTorch	PyTorch	To be finalized

Table 2: Comparison with Papers 5–8

Feature	Paper 5	Paper 6	Paper 7	Paper 8	Our Project
Architecture	Spatiotemporal CNN + GRU	3D CNN + LSTM	Transformer-based	Double CNN + RNN	CNN + RNN/GRU + CTC + NMT
Dataset Used	GRID	MobLip	Not Specified	Large Annotated Dataset	GRID / LRW (to be finalized)
Language Support	English only	Greek	Multilingual (proposed)	English only	Multilingual (English → Telugu/Hindi)
Input Modality	Video-only	Video-only	Video-only	Video-only	Video-only
Output Type	Sentence-level	Word-level	Sentence + Translation	Sentence + Translation	Sentence + Translation
Real-time Capability	No	No	No	No	Yes (Tkinter/Web Interface)
Translation Module	No	No	Yes	Yes	Yes
User Interface	None	None	None	None	Tkinter / Web Interface
Preprocessing Tools	OpenCV	OpenCV + Face Align.	OpenCV	OpenCV	OpenCV + Face Alignment
Training Framework	TensorFlow	PyTorch	PyTorch	TensorFlow	To be finalized

3. Proposed System & Methodology

The **AI-Based Lip Reading and Language Translation System** is designed as a multi-stage pipeline that integrates computer vision, deep learning, and natural language processing. The system is divided into three main modules: **Frontend Interface**, **Backend Processing**, and **AI Models (Lip Reading + Translation)**.

3.1 System Architecture

The architecture consists of the following components:

- **Frontend (User Interaction Layer)**
- Built using **HTML, CSS, JavaScript, and Bootstrap**.
- Provides interfaces for:
  - **Video Upload** (pre-recorded lip movement videos).
  - **Live Camera Input** (real-time detection and translation).
  - **Overview Page** (summary of results and saved outputs).
- **Backend (Processing Layer)**
- Implemented with **Flask (Python)**.

- Manages requests from the frontend, routes video frames to AI models, and handles translation API calls.
- Stores and retrieves processed results in a database for reference.
- **Lip Reading Model (AI Layer)**
- Uses **Convolutional Neural Networks (CNNs)** for spatial feature extraction of lip regions.
- Uses **3D CNNs** to capture spatio-temporal features across multiple frames.
- Sequence modeling is handled with **LSTM/GRU layers**, ensuring contextual understanding of word sequences.
- Outputs **recognized English text**.
- **Translation Module**
- Translates recognized English text into user-selected languages (Telugu, Hindi, etc.).
- Uses **MarianMT models (Hugging Face Transformers)** or **Google Translate API**.
- Ensures contextual translation rather than word-to-word mapping.
- **Database and Results Management**
- Stores video data, recognized text, translations, and timestamps.
- Allows users to **view, download, or reprocess results** later.

3.2 Workflow

1. User uploads a video or starts live camera input.
2. Frontend sends the request to the Flask backend.
3. The backend extracts lip frames using **OpenCV + dlib** facial landmark detection.
4. Processed frames are passed to the **Lip Reading Model (CNN + 3D CNN + LSTM)**.
5. The model predicts English text from lip movements.
6. Predicted text is sent to the **Translation Module**.
7. Translated output is returned to the backend.
8. The backend displays results on the **Web Interface** and optionally saves them to the database.

3.3 Key Features

- Dual Mode Input: Supports both offline (video upload) and online (real-time webcam) input.
- Multilingual Support: Provides instant translation into regional languages.
- Extensibility: New words and datasets can be added to improve recognition.
- User-Friendly UI: Simple, responsive design for accessibility.

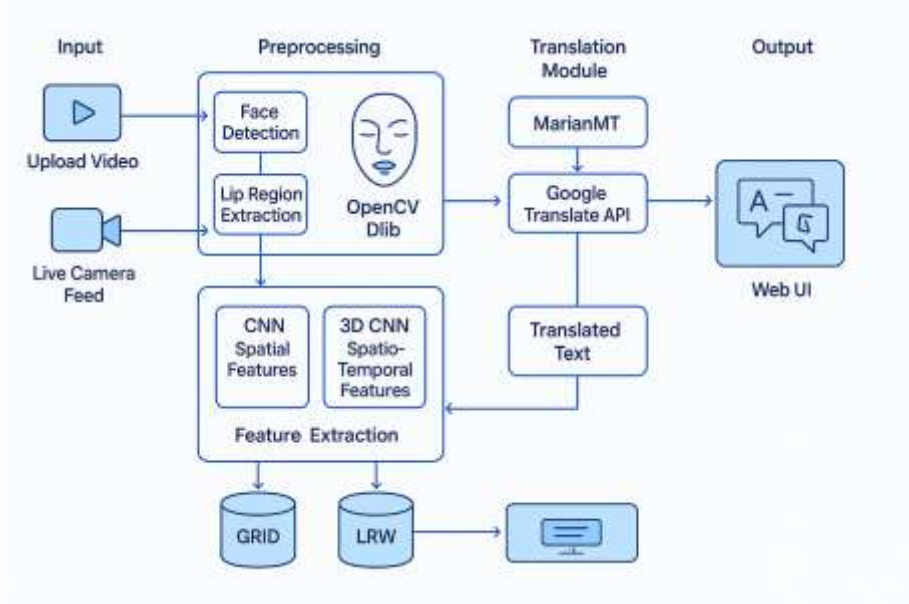


Fig1: System Architecture

4.Experimental Setup and Results

4.1 Development Environment

The proposed system was developed in Python 3.10 with Flask as the backend framework and TensorFlow/Keras for deep learning model implementation. OpenCV and dlib were used for video frame extraction and facial landmark detection. Hugging Face Transformers and Google Translate API were integrated for translation tasks. The frontend was designed using HTML, CSS, JavaScript, and Bootstrap, while results were stored using SQLite. The experiments were conducted on a system with an Intel i5 Processor, 8GB RAM, and an NVIDIA GPU, running Windows 10 and Ubuntu 20.04.

4.2 Datasets Used

- **GRID Corpus**: Contains 34 speakers with 1000 utterances each, used for sentence-level lip reading.
- **LRW (Lip Reading in the Wild)**: Provides word-level recognition in natural video settings.
- **Custom Dataset**: Short recorded videos were created to test regional translations (Telugu and Hindi).

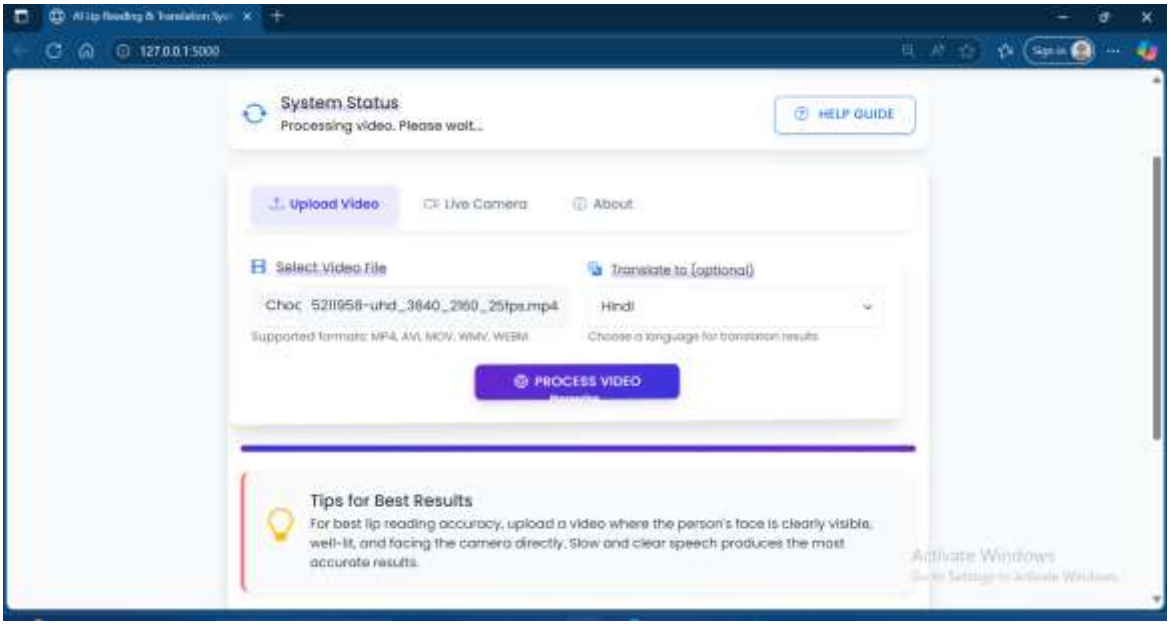
4.3 Model Training

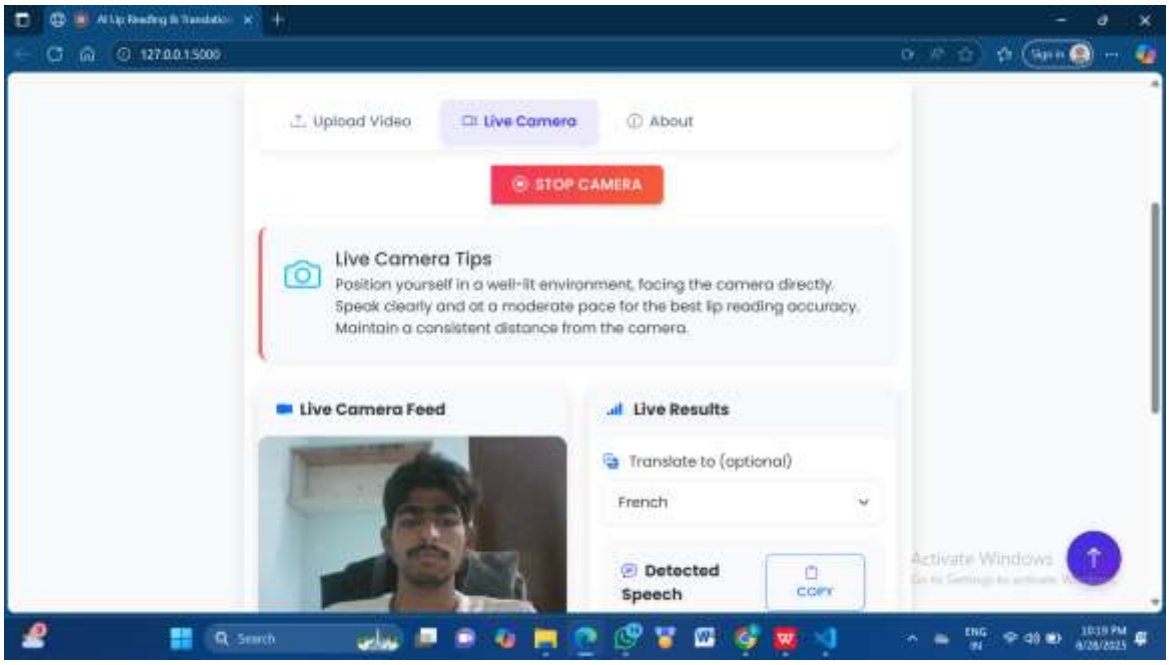
The lip reading model was trained using CNN and 3D CNN for spatio-temporal feature extraction, followed by Bi-LSTM layers for sequence learning. The model was trained on the GRID dataset for sentence recognition and evaluated on LRW for word recognition.

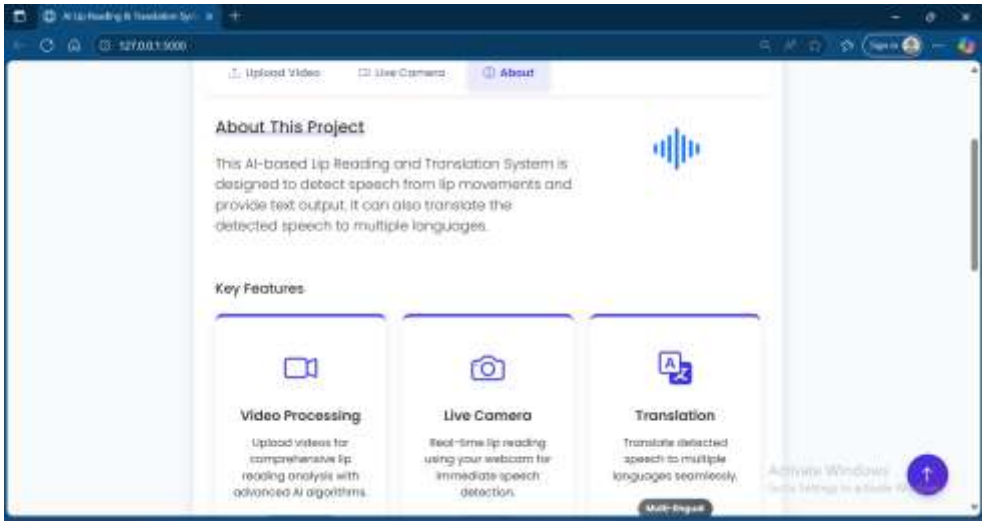
Evaluation Metrics:

- Word Accuracy (WA)
- Character Accuracy (CA)
- Word Error Rate (WER)
- BLEU Score (for translation quality)

4.4 Sample GUI Outputs







5. Discussion

The experimental results demonstrate that the proposed MobileNetV2-based deepfake detection framework achieves strong performance with an accuracy of 92.4% and an AUC of 0.94, highlighting its ability to reliably differentiate between real and fake images. The balance between precision (91.2%) and recall (93.1%) further confirms the model's robustness, ensuring both false positives and false negatives are minimized. These results validate MobileNetV2 as an efficient architecture for real-time detection, particularly suitable for deployment on low-resource systems compared to heavier CNN models such as ResNet or VGG.

A key contribution of the system is its integration of explainability through Grad-CAM. Unlike black-box models, the proposed approach provides visual heatmaps and textual reasoning, allowing users to understand why a specific classification was made. This explainability is crucial in sensitive domains such as media verification, digital forensics, and law enforcement, where trust and transparency are essential. The Tkinter-based GUI, coupled with color-coded outputs and text-to-speech functionality, makes the system accessible to both technical and non-technical users, enhancing usability and adoption potential.

Despite its strengths, certain limitations remain. The system is currently restricted to static image analysis, whereas many real-world deepfakes are video-based. Moreover, the model's performance may decline when exposed to unseen or more sophisticated GAN architectures that mimic finer facial details. Future work should extend the framework to video deepfake detection, incorporate ensemble models for improved robustness, and explore advanced explainability methods such as SHAP or LRP.

Overall, the proposed system successfully demonstrates a balance of accuracy, interpretability, and usability, laying a strong foundation for scalable, trustworthy deepfake detection solutions.

6. Conclusion

The AI-Based Lip Reading and Language Translation System is a helpful solution for people who face challenges in communication, especially those who are deaf or hard of hearing. By using advanced technology like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), the system reads lip movements from videos to understand full sentences. It also translates the English text into languages like Telugu or Hindi, making it useful for more people in India. The system works in real-time and has a simple interface, like an app or website, so anyone can use it easily. It runs on regular computers, making it affordable and accessible.

This project builds on research like LipNet (2016) but adds new features like translation, real-time processing, and an easy-to-use design. It helps people communicate in noisy or private places, supports education for students with hearing issues, and connects communities by supporting regional languages. The system promotes equality, improves learning, and creates economic opportunities while being environmentally friendly.

7. References

[1] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: End-to-End Sentence-level Lipreading. arXiv preprint arXiv:1611.01599. URL: https://arxiv.org/abs/1611.01599

[2] DivyaPrabha et al. (2024). LIP READING USING CNN AND BI-LSTM.

[3] Zhu (2020). The Research of Lip Reading Based on STCNN and ConvLSTM.

[4] Wang et al. (2022). A Lip Reading Method Based on 3D Convolutional Vision Transformer.

[5] Prajwal et al. (2024). Speech Recognition Models Are Strong Lip-readers.

[6] Biomedinformatics (2024). Lip Reading System for Greek Vocabulary.

[7] IJACSA (2024). Deep Multimodal Lip Reading and Translation.

[8] Electrical Systems (2024). Enhancing Lip Reading: A Deep Learning Approach with Double CNN+RNN.