

SilentSpeech AI - A Vision Based Communication System For Mute Individuals

Tejas Mestry
Computer Science Engineering
(Data Science)
Vidyavardhini's College Of Engineering
& Technology
(University of Mumbai)
Vasai, India
tejas.234796105@vcet.edu.in

Sukesh Kotian
Computer Science Engineering
(Data Science)
Vidyavardhini's College Of Engineering
& Technology
(University of Mumbai)
Vasai, India
sukesh.234656101@vcet.edu.in

Chinmayee Mohapatra
Computer Science Engineering
(Data Science)
Vidyavardhini's College Of Engineering
& Technology
(University of Mumbai)
Vasai, India
chinmayee.234876201@vcet.edu.in

Nivedha Raut
Computer Science Engineering
(Data Science)
Vidyavardhini's College Of Engineering
& Technology
(University of Mumbai)
Vasai, India
nivesha.raut@vcet.edu.in

Abstract—Speech recognition has become a vital component of modern human-computer interaction, enabling efficient and natural communication between users and machines. However, conventional speech recognition systems rely primarily on acoustic signals, which limits their performance in noisy environments and restricts accessibility for individuals with speech or hearing impairments. To overcome these challenges, this work proposes an AI-based Silent Speech Lip Reader For Mute Individuals capable of interpreting spoken language through visual information derived from lip and facial movements. The proposed framework supports both English and Hindi languages, broadening its applicability in multilingual contexts. The system captures video input, isolates the lip region, and preprocesses the frames to extract meaningful spatio-temporal features using Deep learning architectures, including Convolutional Neural Networks (CNNs) and sequence modeling networks such as Long Short-Term Memory (LSTM). These extracted features are then decoded into textual output in the corresponding language, followed by conversion into audible speech using a Text-to-Speech (TTS) engine. The developed model aims to enhance communication accessibility, facilitate silent human-computer interaction, and contribute to the advancement of visual speech recognition for low-resource languages. Future directions include expanding the Hindi dataset, incorporating code-mixed speech understanding, and optimizing the system for real-time deployment on edge and mobile platforms.

Keywords— Deep learning, CNNs, LSTM, TTS

I. INTRODUCTION

Human speech is one of the most natural and efficient ways of communication. With rapid progress in artificial intelligence (AI) and machine learning (ML), Automatic Speech Recognition (ASR) has become essential in modern human-computer interaction. These systems are widely used in applications such as virtual assistants, transcription services, and voice-controlled devices. However, their performance declines in noisy environments or when used by individuals with speech or hearing impairments, as ASR systems depend heavily on audio input quality.

To overcome these limitations, Visual Speech Recognition (VSR), or lip reading, has emerged as a promising solution. Instead of processing audio signals, VSR

systems interpret speech by analyzing lip and facial movements. This approach enables communication in silent or noisy environments and improves accessibility for users unable to produce audible speech. Deep learning advancements, particularly in Convolutional Neural Networks (CNNs) and sequence learning models such as Long Short-Term Memory (LSTM) and Transformers, have greatly enhanced the ability to extract spatio-temporal features from video data for accurate visual speech understanding.

This study presents an *AI-based Silent Speech Lip Reader For Mute Individuals* capable of converting lip movements into text and speech in English and Hindi. The proposed system captures video input, detects and preprocesses the lip region, extracts meaningful features, and generates corresponding text output. A Text-to-Speech (TTS) module then converts the text into audible speech, enabling silent-to-spoken communication for accessibility and human-computer interaction applications

II. PROBLEM STATEMENT

In noisy or silent environments and for people with speech or hearing impairments audio-based speech recognition fails or cannot be used. The task is to build a vision-based system that reads lip and facial movements from video and converts them into text and then speech, supporting English and Hindi.

III. LITERATURE SURVEY

The study of lip reading through Artificial Intelligence (AI) and computer vision has gained significant attention due to its ability to assist individuals with speech and hearing impairments. In recent years, several researchers have focused on improving the accuracy, robustness, and real-time performance of visual speech recognition systems. This section reviews notable studies, emphasizing methodologies, datasets, and results that contribute to advancing human-computer communication using visual cues.

Martinez et al. (2025) proposed an end-to-end Spanish lip-reading model using CNN and Transformer architectures. Tested under diverse recording conditions, it achieved 56.6% Word Error Rate (WER) on the TEDx Spanish dataset and 44.6% on MOSEAS. The work demonstrated effective

continuous lip reading but lacked a practical communication interface. [3]

Innocente and Bicego (2025) designed a deep learning-based lip-reading model for assisting vocal-impaired patients. Using a dataset of 25 Italian words related to basic communication, their model achieved 96.4% accuracy. However, it was restricted to a small, predefined vocabulary. [7]

Vekkota et al. (2025) implemented a feature fusion approach combining CNNs such as ResNet, VGG16, and Inception-V3 with LSTM layers. On the MIRACL VC1 dataset, the system achieved 75% accuracy in speaker-dependent settings, demonstrating the potential of hybrid deep learning. [8]

Aripin and Setiawan (2024) developed an Indonesian lip-reading model using Face Mesh and Long-Term Recurrent Convolutional Networks (LRCN). The system achieved accuracies of 95.4% for words, 95.6% for phrases, and 90.6% on the MIRACL dataset. However, it was limited to controlled lighting and frontal face positions. [4]

Dubey and Sharma (2024) proposed a smart-glass-based system that recognizes lip movements and displays corresponding text on the glasses. The design showed innovation in wearable AI but lacked detailed accuracy results and was expensive for large-scale adoption. [6]

Prajwal K. R. et al. (2024) from the University of Oxford presented “Speech Recognition Models Are Strong Lip-Readers,” showing that cross-modal learning enables silent video-to-text transcription. Using the LRS3 dataset, they achieved a 24.3% WER, setting a benchmark but requiring high computational resources. [10]

Vural et al. (2022) built a Turkish lip-reading system using Bi-LSTM models to improve temporal sequence learning. Their method achieved 84.5% accuracy for words and 88.55% for sentences but lacked multilingual or real-time support. [2]

Deshmukh and Gawali (2021) implemented a vision-based lip-reading model using deep learning with live camera input. The model achieved 85% accuracy for digit recognition (0–9), though its vocabulary remained limited. [5]

Neeraja et al. (2021) introduced a deep learning-based system for mute drivers that converted lip movements into directional commands. The system achieved 88% accuracy but was restricted to predefined control words. [9]

Kumar et al. (2020) developed an LSTM-based model for recognizing Hindi lip movements in Devanagari script. Their system achieved 77% accuracy for 3 words, 35% for 10 words, and 20% for sentences, highlighting the difficulty of visual speech recognition in complex languages. [1]

In summary, research in AI-based lip reading continues to advance through CNN, LSTM, and Transformer models. Although high accuracies are observed in controlled settings, most systems face challenges in handling diverse speakers, lighting variations, and low-resource languages. Insights from these studies form the foundation for our proposed *AI-based Silent Speech Lip Reader For Mute Individuals*, aimed at developing a real-time, multilingual, and accessible communication framework.

IV. METHODOLOGY

The proposed *Silent Speech Lip Reader For Mute Individuals* system translates visual speech information, specifically lip and facial movements, into textual and corresponding audible outputs for English and Hindi

languages. The overall framework consists of five primary stages: data acquisition, preprocessing, feature extraction and sequence modeling, text-to-speech synthesis, and model training and optimization. The overall pipeline is illustrated in Fig. 1

A. Algorithm and Process Design

• Data Acquisition

For English visual speech recognition, the Lip Reading Sentences 2 (LRS2) dataset is used, containing thousands of video clips with aligned text transcriptions across diverse speakers and conditions. For Hindi, a small custom dataset was created using recordings from native speakers, ensuring synchronized video-text pairs. This bilingual setup supports evaluation across high- and low-resource languages.

• Preprocessing and Lip Detection

Each video is divided into frames, and facial landmarks are detected using MediaPipe Face Mesh. The lip region is cropped, resized to 96×96 pixels, and converted to grayscale. Frames are normalized and temporally aligned to remove background noise and ensure uniform input quality.

• Feature Extraction and Sequence Modeling

A ResNet-18 backbone extracts spatial lip features from consecutive frames, capturing shape and articulation patterns. The extracted features are fed into a Bidirectional Long Short-Term Memory (Bi-LSTM) encoder to model temporal dependencies. This hybrid setup effectively learns both spatial and temporal dynamics, producing the predicted text sequence.

• Text-to-Speech (TTS) Conversion

The generated text is transformed into audible speech using a Tacotron 2-based TTS engine, producing natural-sounding output in English and Hindi.

• Model Training and Optimization

The model is trained end-to-end using the Connectionist Temporal Classification (CTC) loss and optimized with the Adam optimizer. Data augmentation techniques such as brightness variation and frame jittering improve generalization. Evaluation metrics include Word Error Rate (WER) and Character Accuracy.

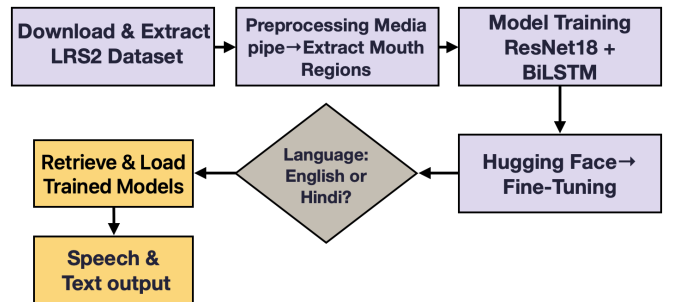


Fig. 1. Workflow of proposed system

B. Details of Hardware & Software

Hardware Specifications

- System Type: x64-based processor, 64-bit operating system
- Processor: Intel Core i5 / i7 (or equivalent AMD Ryzen)

- Memory (RAM): 8 GB or higher (recommended 16 GB for model training)
- Storage: 1 TB Hard Disk / 256 GB SSD
- GPU: NVIDIA GPU with CUDA support (e.g., RTX 2060 or higher)
- Display: Minimum 1080p resolution for dataset visualization and preprocessing tasks

Software Specifications

- Operating System: The implementation is platform-independent and can run on Windows 10/11, Linux (Ubuntu 20.04+), or macOS.
- Programming Language: Python 3.10+ is used as the core programming language for data preprocessing, deep learning, and model training.
- Integrated Development Environment (IDE): The project can be executed on IDEs such as Google Colab, Jupyter Notebook, PyCharm, or Visual Studio Code.
- Deep Learning Frameworks and Libraries:
 1. TensorFlow / PyTorch – for deep learning model design and training
 2. OpenCV – for video frame extraction and image processing
 3. MediaPipe – for facial and lip landmark detection
 4. NumPy – for numerical and array operations
 5. Pandas – for dataset handling and preprocessing
 6. Matplotlib – for data visualization
 7. Hugging Face Transformers – for model fine-tuning and multilingual support

The system primarily relies on Python-based deep learning frameworks and GPU acceleration to train the ResNet-18 + Bi-LSTM/Transformer architecture effectively. All experiments were conducted on Google Colab and local GPU-enabled machines.

V. RESULT

Epoch 14/100
 Train Loss: 1.7838
 Val Loss: 2.7391 | WER: 111.71% | CER: 73.73%

--- Sample Predictions (validation) ---

GT: so we need you to help us in our revival campaign
 Pred: st wats in to o ins hae a te ine was fer

GT: or into vodka and adding a bit of honey
 Pred: is tn ies tha tes pots

GT: those should have rooted
 Pred: t soe fr hereysy,

Fig. 2. Prediction for Epoch 14

Epoch 15/100
 Train Loss: 1.6885
 Val Loss: 2.8275 | WER: 109.97% | CER: 74.03%

--- Sample Predictions (validation) ---

GT: so we need you to help us in our revival campaign
 Pred: its wad in in to ion hae a ae an was ter

GT: or into vodka and adding a bit of honey
 Pred: i in ies tae aeis por wets

GT: those should have rooted
 Pred: ths so fthry

Fig. 3. Prediction for Epoch 15

VI CONCLUSION & FUTURE WORK

This work presents an AI-based Silent Speech Lip Reader designed to convert visual lip movements into corresponding text and audible speech for English and Hindi. The system combines ResNet-18 for spatial feature extraction with Bi-LSTM/Transformer models for temporal sequence learning, supported by a Text-to-Speech (TTS) module for natural audio output. Experiments conducted on the LRS2 and custom Hindi datasets demonstrate the framework's adaptability to multilingual environments. The proposed system shows promising potential in accessibility, assistive communication, and silent human-computer interaction, paving the way for more inclusive and intelligent speech technologies.

Future Work

1. Dataset Expansion: Increasing the size and diversity of the Hindi dataset to improve recognition accuracy.
2. Code-Mixed Speech Support: Extending the system to handle bilingual and code-switched communication common in real-world contexts.
3. Model Optimization: Reducing computational complexity for real-time inference on mobile and edge devices.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Prof. Nivedha Raut for her invaluable guidance, motivation, and

continuous support throughout the duration of this work. The authors also extend their appreciation to the faculty members of the Department of Computer Science and Engineering (Data Science) for their valuable suggestions and encouragement during the Mini Project. The support and cooperation provided by the departmental staff are also gratefully acknowledged. Finally, the authors thank their parents for their constant encouragement and support during this endeavor

REFERENCES

- [1] R. Kumar, K. Kapoor, S. Srivastava, A. Kumar, and P. Kumar, "LSTM Based Lip Reading Approach for Devanagari Script," *International Journal of Research in Engineering and Technology (IJRET)*, vol. 9, no. 2, pp. 45–50, 2020.
- [2] E. Vural, E. Çelikkán, and M. E. Çelebi, "Turkish Lip-Reading using Bi-LSTM and Deep Learning Models," *IEEE Access*, vol. 10, pp. 54132–54140, 2022.
- [3] A. Martinez, J. Crego, and M. R. Costa-jussà, "Evaluation of End-to-End Continuous Spanish Lipreading in Different Data Conditions," *Journal of Intelligent Systems and Applications*, vol. 15, no. 3, pp. 120–129, 2025.
- [4] Aripin and A. Setiawan, "Indonesian Lip-Reading Detection and Recognition Based on Lip Shape Using Face Mesh and Long-Term Recurrent Convolutional Network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 1, pp. 89–97, 2024.
- [5] N. Deshmukh and A. Ahire, "Vision-Based Lip Reading System using Deep Learning," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 9, no. 5, pp. 1025–1031, 2021.
- [6] S. Dubey and D. Sharma, "Establishing Communication Through Lip Reading With the Aid of Machine Learning," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, vol. 10, no. 2, pp. 87–94, 2024.
- [7] C. Innocente, M. Boemio, and M. Bicego, "Deep Learning-Based Lip Reading for Vocal Impaired Patient Rehabilitation," *Procedia Computer Science*, vol. 227, pp. 115–122, 2025.
- [8] S. Vekkota, D. Kaushik, and A. Mandapati, "Enhanced Lip Reading Using Deep Model Feature Fusion: A Study on the MIRACL VC1 Dataset," *International Journal of Computer Applications*, vol. 182, no. 4, pp. 15–22, 2025.
- [9] K. Neeraja, K. S. Rao, and G. Praneeth, "Deep Learning-Based Lip Movement Technique for Mute," *Proceedings of the IEEE International Conference on Communication and Electronics Systems (ICCES)*, pp. 1391–1397, 2021.
- [10] P. K. R., T. Afouras, and A. Zisserman, "Speech Recognition Models Are Strong Lip-Readers," *Interspeech Conference Proceedings*, University of Oxford, pp. 278–282, 2024.