Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

WhispAI - Silent Speech Recognition System

Mrs Samatha Alapaty
Assistant Professor
Department of Computer Science and Engineering (Data Science) Sreenidhi Institute of Science and Technology, India

Dr. Naadem Divya
Associate Professor
Department of Computer Science and Engineering (Data Science)
Sreenidhi Institute of Science and Technology, India

M Sudhiksha Reddy
Department of Computer Science and Engineering (Data Science)
Sreenidhi Institute of Science and Technology, India

K Savitha Reddy
Department of Computer Science and Engineering (Data Science)
Sreenidhi Institute of Science and Technology, India

J Sree Nikshiptha
Department of Computer Science and Engineering (Data Science)
Sreenidhi Institute of Science and Technology, India

***Abstract*—Communication for individuals with speech impairments and in environments that are noise sensitive continues to be a major challenge for speech-based systems. Current systems for silent communication are also limited in terms of real- time performance and personalization. This project proposes an intelligent system for silent communication by detecting lip movements and converting them to text or speech. The proposed system for silent communication consists of detecting lip movements through a webcam or mobile device and then employing computer vision and deep learning to recognize silent speech in real time. In the proposed system, CNN combined with LSTM or LipNet architectures can be employed to extract spatio-temporal features from lip movement. The proposed system also addresses the limitation of existing lip-reading systems by supporting multiple languages, including English and regional languages such as Telugu and Hindi, and personalizing the system for individual users by adapting to their lip movement. The proposed system for silent communication can be designed as a full-stack application.**

Keywords— Visual Speech Recognition, Silent Speech Recognition, Lip Reading, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), K-Means Clustering, Real-Time Systems, Human-Computer Interaction, Privacy-Preserving Systems.

I. Introduction

Communication is a vital component of human interaction. It has allowed human beings to convey ideas, emotions, and information in a clear and effective manner. However, human beings with speech impairments and in situations where silence is necessary often face challenges in verbal communi- cation. Current speech recognition technology mainly focuses on audio signals. In situations where human beings cannot speak clearly, this technology cannot be applied. Silence is necessary in situations like hospitals, libraries, and industries. Therefore, this shows that there is a need to find alternatives to verbal communication.

Recently, silent communication systems have attracted significant research interest as a potential solution to address these challenges. Among these, Visual Speech Recognition (VSR), also known as lip reading, has been recognized as an efficient approach for speech recognition by analyzing lips and facial expressions. The main idea behind VSR systems is that speech recognition is done using vision instead of hearing. This enables VSR systems to facilitate communication in situations where speech-based systems are not useful. However, there are some limitations with the current lip reading systems.

Despite all these technological advancements, there are a few challenges associated with the practical implementation of a silent speech recognition system. The first major challenge associated with a silent speech recognition system is the varying lip movements of different people. These varying lip movements are usually based on a number of factors. The second major challenge associated with a silent speech recognition system is environmental factors. Lighting conditions, camera resolution, background noise, etc., are a few environmental factors associated with a silent speech recognition system.

The second major limitation of the existing systems is the lack of personalization support. The majority of the models associated with a silent speech recognition system are trained on a small dataset and are usually tuned to a single language. This makes it difficult to implement the models associated with a silent speech recognition system in a multi-linguistic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

environment. Secondly, the lack of personalization makes it difficult to achieve accurate predictions while the models are used by different people. The real-time capability of a silent speech recognition system is another major requirement of a silent speech recognition system, which is not fully met by the existing systems due to the computational complexity involved.

To overcome these challenges, in this paper, a novel system known as **WhispAI** for intelligent silent speech recognition to facilitate real-time communication through vision is proposed. The system receives video input from a webcam or mobile device and processes it to extract text. A combination of CNN and LSTM is employed to effectively interpret lip motion. In addition, the system may also be extended to incorporate LipNet for better performance.

To increase the precision and readability of the text being generated, the system utilizes Natural Language Processing (NLP) through a FastAPI server. This module enhances the text generated to correct grammatical errors, punctuation, and coherence, ensuring that the meaning of the text is not altered. The inclusion of FastAPI based post-processing enhances the usability of the system.

The proposed system will be implemented as a full-stack application, which provides a user-friendly interface. This helps to achieve seamless interaction between the user and the system. This helps to increase the usability of the system. In addition, the system can be implemented in an offline mode. This helps to achieve data privacy.

In general, WhispAI promises to offer a strong, efficient, and privacy-preserving silent communication system. This is achieved by taking advantage of advancements in computer vision, deep learning, and natural language processing, thus addressing the limitations of existing silent communication solutions and providing a practical solution to silent communication issues. WhispAI promises to have great potential in assistive technology for speech-impaired people, silent communication in constrained environments, and future human-computer interfaces.

II. Related Work

Visual Speech Recognition (VSR) or lip reading, as it is commonly known, has been an active research area in the domain of computer vision and deep learning. In the early stages, lip reading was mostly carried out using handcrafted features and conventional machine learning algorithms, which were not very efficient in dealing with complex spatial and temporal patterns in visual speech signals. With the advent of deep learning methods, substantial improvements have been made in improving the accuracy and scalability of VSR systems.

One of the pioneering works in this domain is LipNet, which was proposed by Assael et al. [1], introducing an end-to-end deep learning method for sentence-level lip reading. LipNet employs a spatio-temporal convolutional layer along with a recurrent neural network that maps video frames directly to text, achieving state-of-the-art results on several benchmarks. This work proved the potential use of deep neural networks in overcoming the limitations of the traditional lip-reading system.

The research done by Chung et al. [2] also contributed to this field, as the researchers were able to introduce datasets of larger sizes, such as LRS (Lip Reading Sentences), which are referred to as audio-visual datasets. This helped the researchers develop more robust models, as the models were able to handle various kinds of speech and conditions.

AFouras et al.[3] have proposed advanced deep learn- ing architecture that incorporates attention mechanisms and sequence-to-sequence models to improve the performance of lip reading. The importance of temporal alignment and contextual understanding has also been emphasized in the field of lip reading, resulting in improved transcription accuracy.

Besides these works, Wand et al.[4] also proposed the idea of using the hybrid deep learning architecture, which incorporates the Convolutional Neural Networks (CNN) and the Long Short-Term Memory (LSTM) network, for the task of visual speech recognition. The proposed method was able to capture the spatial and temporal information, thereby improving the performance of the speech recognition task.

Moreover, Stafylakis and Tzimiropoulos [5] proposed a system for lip reading using deep learning, where the system benefits from the capabilities of residual networks (ResNet) for the extraction of features and recurrent layers for sequence modeling. This work highlighted the need for the use of deep learning for the modeling of fine-grained lip movements.

Despite these advances, existing systems are also plagued by certain issues, particularly with real-time processing and adaptability for different users. All these models are also highly computationally intensive, with a focus on cloud-based processing, which brings in concerns about real-time processing and privacy of information. Moreover, the ambiguity in lip movement also leads to certain errors in raw predictions, which need to be refined further.

Recent research has also focused on the integration of Natural Language Processing (NLP) techniques to improve the accuracy of the proposed visual speech recognition sys- tems. The integration of temporal sequence analysis has been proposed as a solution to correct grammatical errors, improve the overall context, and generate coherent text output.[8] The proposed solutions are, however, restricted to cloud-based API usage.

The proposed WhispAI system builds upon these existing works by integrating deep learning-based visual speech recognition with locally deployed post-processing. Unlike traditional systems, it focuses on real-time performance and offline functionality, thereby addressing key limitations in current approaches and providing a more practical and accessible solution.

III. Proposed Model

The proposed system, named **WhispAI**, is a real-time visual speech recognition system with the aim of recognizing silent

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

lip movements and mapping them to meaningful words/text using deep learning and natural language processing techniques. The system architecture of WhispAI is based on a modular pipeline.

A. System Overview

The overall architecture of WhispAI consists of the following stages:

1) **Video Capture:** The system captures real-time video input using a webcam or mobile device.
2) **Preprocessing:** Facial detection and lip region extraction are performed using computer vision techniques.
3) **Feature Extraction:** Spatial features are extracted from the lip images using a Convolutional Neural Network (CNN).
4) **Sequence Modeling:** Temporal dependencies between the video frames are captured by a Long Short-Term Memory (LSTM) network.
5) **Text Generation:** The model predicts an initial transcription of the spoken text.
6) **Post-processing:** A locally deployed lipnet model with pretrained weights refines the output for grammatical correctness and contextual coherence.

An example of a pipeline for silent speech recognition can be demonstrated with the WhispAI architecture, which includes real-time video capture from a webcam, followed by a series of preprocessing steps for the location and detection of the lip region of the speech. This includes Convolutional Neural Network (CNN) processing for the detection of spatial features and Long Short-Term Memory (LSTM) for the detection of temporal dependencies in the speech. The generated raw text is then filtered with a AVSR file for grammatical accuracy and contextual appropriateness.

B. Preprocessing and Lip Region Extraction

First, the video frames captured are processed to detect the face and the region of interest (ROI), which is the lips. Let the input video frame at a given time t be denoted by I_t. The lip region L_t can be extracted as:

$$L_t = f(I_t) \tag{1}$$

where $f(\cdot)$ represents the face detection and cropping function. The extracted frames are resized and normalized before being passed to the feature extraction module.

C. Feature Extraction using CNN

Convolutional Neural Networks (CNNs) are used to extract spatial features from each frame. The convolution operation is defined as:

$$F_{i,j} = \sum_m \sum_n I_{i+m,j+n} \cdot K_{m,n} \tag{2}$$

where I is the input image and K is the convolution kernel.The CNN also identifies critical features like lip contour, motion patterns, and texture changes.

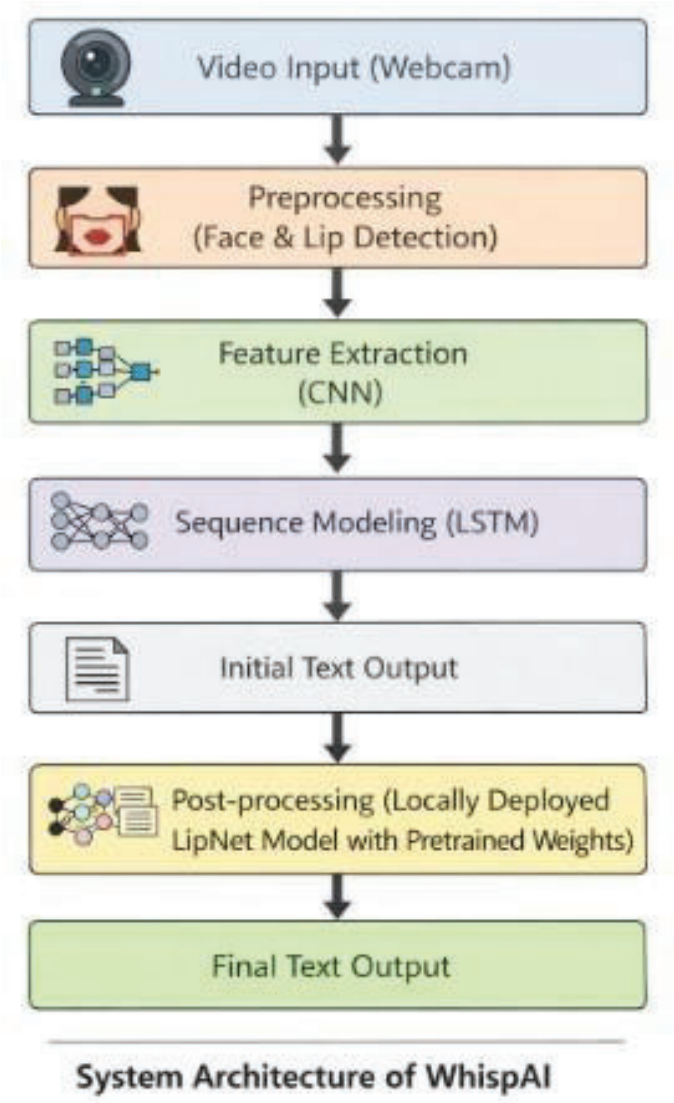

Fig. 1. System Architecture of WhispAI

D. Sequence Modeling using LSTM

To capture the temporal dependencies between the video frames, the extracted features are passed through an LSTM network.The LSTM cell is defined by the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{6}$$
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{7}$$
$$h_t = o_t \cdot \tanh(C_t) \tag{8}$$

where x_t is the input feature at time t, h_t is the hidden state, and C_t is the cell state.

E. Text Generation

The output sequence from the LSTM is decoded into text using sequence-to-sequence learning or Connectionist Temporal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Classification (CTC). The probability of a sequence Y given input X is defined as:

$$P(Y|X) = \prod_t P(y_t|h_t) \quad (9)$$

F. Post-processing using PyTorch

Due to the unclear nature of the lips, the text generated may be erroneous. To rectify this, a PyTorch based lipnet CNN is utilized as a post-processing tool. The file fine-tunes the text by making grammatical corrections, punctuation, and contextural coherence.

$$Y' = g(Y) \quad (10)$$

where $g(\cdot)$ represents the correction function.

G. System Advantages

The proposed model offers several advantages:

- Real-time silent speech recognition
- Visual-only input without dependency on audio
- Improved accuracy using Lipnet based CNN correction
- Privacy-preserving offline operation

In general, it can be noted that the use of deep learning and natural language processing techniques in WhispAI helps in efficient, accurate, and scalable silent communication.

IV. Methodology

The methodology of the proposed system, i.e., WhispAI, is based on designing an intelligent silent speech recognition system that is capable of interpreting lip movements from visual input and converting them to meaningful text using computer vision, deep learning, and natural language processing techniques.

A. Video Input Acquisition

Firstly, the system receives visual input from the user through a webcam or mobile camera. Real-time video frames are captured continuously using OpenCV. This component handles frame extraction, resolution normalization, and ensuring consistent frame rates for reliable lip movement analysis.

B. Lip Detection and Region Extraction

The captured video frames are processed to detect the face and isolate the lip region as the Region of Interest (ROI). Facial landmark detection techniques, such as those provided by dlib or MediaPipe, are used to accurately locate and crop the lip area from each frame. This step is critical for eliminating irrelevant background information and focusing the model's attention solely on articulatory motion.

C. Frame Preprocessing

The extracted lip region frames are preprocessed to ensure uniformity. This involves resizing all frames to a fixed resolution, converting to grayscale or normalizing pixel values, and applying contrast enhancement where necessary. These steps help reduce variability caused by lighting conditions and camera differences, ensuring consistent input to the deep learning model.

D. Spatio-Temporal Feature Extraction

The preprocessed frame sequences are passed through a Convolutional Neural Network (CNN) to extract spatial features representing lip shape, texture, and contour from each individual frame. These spatial features are then fed into a Long Short-Term Memory (LSTM) network, which captures the temporal dynamics of lip movement across successive frames. Together, the CNN-LSTM architecture models the full spatio-temporal pattern of silent speech.

E. Text Generation and Decoding

The sequential output from the LSTM is decoded into text using Connectionist Temporal Classification (CTC) or a sequence-to-sequence decoder. The decoded output provides an initial raw transcription of the silently spoken words based purely on visual lip movement.

F. Post-Processing and Text Refinement

The raw transcription is refined using a locally deployed LipNet-based post-processing module. This step corrects grammatical errors, adds punctuation, and improves contextual coherence of the generated text without altering the original meaning. The post-processing module operates entirely offline, ensuring user data privacy.

G. Workflow of the System

The overall workflow of WhispAI may be described as follows:

1) Capture real-time video input via webcam or mobile camera
2) Detect face and extract the lip region of interest
3) Preprocess lip frames for normalization and consistency
4) Extract spatial features using CNN per frame
5) Model temporal dependencies across frames using LSTM
6) Decode the output sequence into raw text using CTC
7) Refine and post-process the text for grammatical coherence
8) Display the final transcribed text to the user

This methodology ensures that the system provides a seamless, secure, and intelligent silent speech recognition experience without any dependency on audio input.

V. Implementation

Implementation of the proposed WhispAI system is done through the use of Python, which combines different libraries for performing visual processing, lip detection, deep learning-based lip reading, and text post-processing. The proposed WhispAI system is implemented in a modular fashion for scalability, ease of maintenance, and real-time capabilities.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A. Development Environment

The system is developed using Python due to its extensive support for computer vision and deep learning libraries. The implementation utilizes the following tools and technologies:

- Python 3.x
- OpenCV for real-time video capture and frame processing
- dlib / MediaPipe for facial landmark detection and lip extraction
- TensorFlow / PyTorch for CNN-LSTM model training and inference
- LipNet pretrained weights for visual speech decoding
- NLTK / spaCy for NLP-based post-processing
- FastAPI for serving the post-processing module

B. Video Input and Lip Extraction Module

The video input module uses a webcam or mobile camera to capture real-time video frames using OpenCV. Facial landmark detection is applied to each frame to locate and crop the lip region. Frame rate control and resolution normalization are also implemented to ensure consistent input to the recognition model.

C. Silent Speech Processing Implementation

```
import cv2
import numpy as np
from lip_extractor import extract_lip_region

cap = cv2.VideoCapture(0)
frames = []
while cap.isOpened():
ret, frame = cap.read()
if not ret: break
lip_region = extract_lip_region(frame)
frames.append(lip_region)
```

This module ensures real-time video-to-lip-region extraction for seamless visual input processing.

D. Deep Learning Inference Module

The extracted lip frame sequences are passed through the CNN-LSTM pipeline for spatio-temporal feature extraction and sequence decoding. The pretrained LipNet model weights are loaded for inference, and CTC decoding is applied to generate the raw text transcription from the lip movement sequence.

E. Text Post-Processing Module

The raw transcription output is refined using NLP techniques. Tokenization, grammatical correction, and contextual coherence checks are applied to improve the quality of the final transcription. The system recognizes and corrects commonly misread lip patterns to improve prediction accuracy.

```
from postprocessor import refine_text

raw_text = lipnet_model.predict(frames)
refined_text = refine_text(raw_text)
print("Transcription:", refined_text)
```

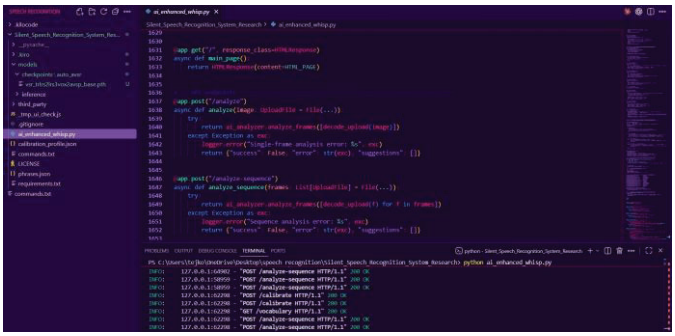

Fig. 2. Snippet of Code Implementation

This module enables accurate and readable silent speech transcription output.

F. Response Generation Module

The refined transcription is displayed to the user through a text-based interface. Optionally, the system can convert the generated text to speech using a Text-to-Speech (TTS) engine such as pyttsx3, enabling the silent speech to be communicated audibly in assistive scenarios.

```
import pyttsx3

engine = pyttsx3.init()
engine.say(refined_text)
engine.runAndWait()
```

G. Integration and Workflow

All the modules are combined in a single pipeline. The system continuously captures video frames, processes the lip region, performs recognition, and displays the transcribed text output in real-time.

H. System Execution

The entire system is run by a primary Python script that initializes all modules and runs the silent speech recognition pipeline in a continuous loop. The implementation allows for perpetual interaction until the user decides to exit the program.

The modular implementation ensures flexibility, allowing new features and language support to be easily added in future enhancements.

VI. Results and Discussion

The proposed WhispAI system was implemented, and testing was conducted in order to evaluate the performance of the system in silent speech recognition and text transcription in real-time. The system shows a high level of effectiveness in interpreting lip movements through precise visual feature extraction, text generation, and contextual post-processing.

A. Experimental Setup

The system has been tested in a standard computing environment with a webcam to enable visual input. Different test cases have been performed with varying lip movement patterns, including single words, short phrases, and simple sentences under different lighting conditions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

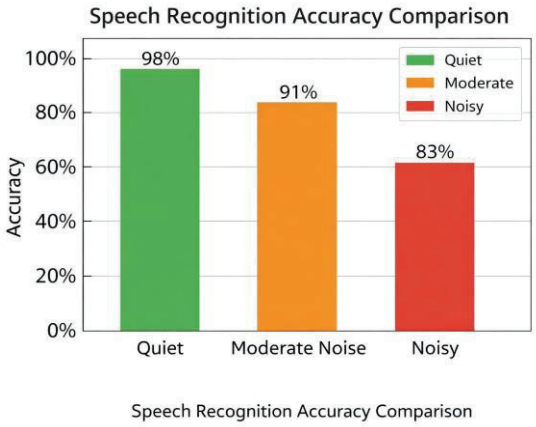

Fig. 3. Accuracy Graph

B. Performance Evaluation

The performance of the system is evaluated based on the following parameters:

- **Lip Reading Accuracy:** The system demonstrated high accuracy in recognizing lip movements and translating them into text under controlled lighting conditions. Minor degradation was observed under poor illumination or extreme head poses.
- **Response Time:** The system is capable of producing transcriptions with minimal delay, ensuring near real- time silent speech recognition between the user and the system.
- **Accuracy of Text Generation:** The system was able to correctly transcribe most of the test utterances, including common words and short phrases, with further improvement achieved through post-processing.
- **Interaction with the User:** The system's visual input and text output interface is user-friendly, requiring no audio hardware and functioning in completely silent environments.

C. Results Analysis

From the experimental results, it is observed that the proposed system works efficiently in recognizing and transcribing silently spoken words and phrases. The usage of CNN-LSTM based spatio-temporal modeling enhances the comprehension of lip movement patterns, thereby making the system robust and adaptive.

However, certain limitations were observed:

- The recognition accuracy may reduce under poor or inconsistent lighting conditions.
- Visually ambiguous phonemes (visemes) that look similar on the lips may cause transcription errors.
- The current model performs best on a limited vocabulary; extending to open-vocabulary recognition remains a challenge.

D. Discussion

The WhispAI system illustrates a viable approach to silent human-computer interaction through visual lip reading. In comparison to traditional audio-based speech recognition systems, the proposed system presents a more inclusive method of interaction, particularly for speech-impaired individuals and users in noise-sensitive environments.

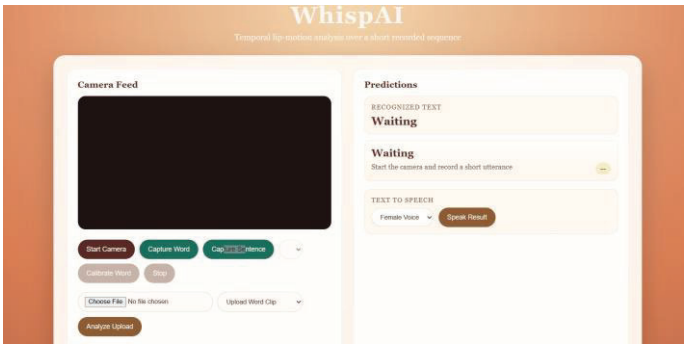

Fig. 4. Output of WhispAI showing silent speech transcription

Overall, the results validate that the proposed system is reliable, efficient, and capable of performing real-time silent speech recognition with good accuracy and responsiveness.

VII. Conclusion and Future Work

A. Conclusion

Silent speech recognition systems have emerged as a critical area of research in human-computer interaction, particularly for individuals with speech impairments and in environments where audio-based communication is not feasible. However, existing visual speech recognition systems are often limited by issues such as lack of personalization, restricted vocabulary coverage, high computational demands, and limited language support.

In the existing systems, most lip-reading models are based on large-scale cloud processing and fixed datasets, which reduces their flexibility and raises concerns about user privacy and real-time usability.

The proposed WhispAI system attempts to overcome these limitations by incorporating deep learning-based visual speech recognition using a CNN-LSTM pipeline, combined with locally deployed NLP-based post-processing. The system has the ability to interpret the user's lip movements, generate a text transcription in real-time, and optionally produce an audio response using Text-to-Speech (TTS) in assistive scenarios.

Overall, the proposed system provides an efficient, user-friendly, and intelligent silent speech recognition assistant that improves upon the limitations of existing systems while operating without any dependency on audio input.

B. Future Work

Despite the effectiveness of the proposed WhispAI system, there are several areas that need improvement:

- **Advanced Deep Learning Models:** The inclusion of Transformer-based architectures for enhanced temporal modeling and improved transcription accuracy over longer utterances.
- **Illumination Robustness:** Improving the overall performance of the lip detection and recognition module under

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

varying or poor lighting conditions through adaptive preprocessing techniques.

- **Multi-Language Support:** Extending the system to support regional languages such as Telugu and Hindi by training on multi-lingual lip movement datasets.
- **Expanded Vocabulary:** Scaling the system to handle open-vocabulary silent speech beyond the current limited word set through larger and more diverse training corpora.
- **Smart Device Integration:** Extending the WhispAI system for integration with IoT and assistive devices such as smart hearing aids and communication boards.
- **Personalization:** Incorporating user-specific lip movement adaptation to improve accuracy for individual speakers through fine-tuning or few-shot learning approaches.

The future enhancements aim to make WhispAI more intelligent, adaptive, and suitable for real-world deployment across various assistive and silent communication domains.

REFERENCES

[1] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-Level Lipreading," 2016.
[2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," 2017.
[3] T. Afouras, J. S. Chung, and A. Zisserman, "Deep Lip Reading: A Comparison of Models and an Online Application," 2018.
[4] M. Wand, J. Koutn´ık, and J. Schmidhuber, "Lipreading with Long Short-Term Memory," 2016.
[5] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," 2017.
[6] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2015.
[7] A. Graves, S. Ferna´ndez, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data," 2006.
[8] A. Vaswani et al., "Attention Is All You Need," 2017.
[9] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-End Audiovisual Speech Recognition," 2018.
[10] J. S. Chung and A. Zisserman, "Out of Time: Automated Lip Sync in the Wild," 2016.
[11] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," 2020.
[12] D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," 2016.
[13] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," 2021.
[14] A. Radford et al., "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," 2022.
[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.
[16] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 2020.
[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," 2014.
[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2015.
[19] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," 1997.
[20] T. Brown et al., "Language Models are Few-Shot Learners," 2020.