IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

Silent Voice Assistants: Lip-Reading AI with On-Device Processing

Shubham Gaykhe[1], Saloni Bhujbal[2], Priyanka Chavan[3], Prof. Deepashri Patil[4]
Computer Science Department
Dr. D. Y. Patil Arts, Commerce, Science College, Pimpri, Pune

Abstract: *Traditional voice assistants rely on spoken audio commands, which often raise concerns related to privacy, accessibility, and performance in noisy or sensitive environments. This research proposes a Silent Voice Assistant (SVA) that utilizes lip-reading artificial intelligence to interpret user commands through visual speech recognition. By analyzing lip movements instead of audio input, the system enables silent and secure interaction between users and devices. The solution is designed to operate using on-device processing, eliminating the need for cloud-based computation and thereby reducing latency and minimizing the risk of data leakage.*

The proposed system employs deep learning models such as CNN, LSTM, and transformer-based architectures to extract spatial and temporal features from video frames. These models are optimized for edge devices using techniques like quantization and pruning to ensure efficient real-time performance. The system is evaluated in various environments, including noisy and quiet settings, to assess its accuracy, responsiveness, and usability. The results demonstrate that lip-reading AI can serve as a reliable alternative to traditional voice assistants, particularly for individuals with speech impairments and in privacy-sensitive situations. This research contributes to the development of inclusive, efficient, and privacy-preserving human-computer interaction systems..

Keywords: *deep learning*

I. INTRODUCTION

This research aims to bridge that gap by developing a Silent Voice Assistant (SVA) that works entirely on-device, using lip-reading artificial intelligence to recognize and interpret commands without the need for audible speech or external data transmission. At the core of this system is a combination of spatiotemporal deep learning which analyzes both spatial features (lip shapes and facial movements) and temporal patterns (how this change over time) and edge computing, which enables real-time processing on the device itself, without sending data to the Key Innovations and Benefits: True Privacy: All data stays on the user's device. There is no need to transmit video or audio to external servers, eliminating risks of eavesdropping or data breaches. Real-Time Responsiveness: On-device inference means the assistant can understand and respond to commands instantly, with minimal latency, even in areas with no internet access. Enhanced Accessibility: Users who cannot speak due to physical conditions, injury, or temporary impairment can still interact with smart systems through silent lip movements. Environmental Flexibility: Whether in a loud, chaotic setting or a sound-sensitive space, users can communicate naturally without relying on voice. Traditional voice assistants depend on spoken commands, which can create problems in terms of privacy, accessibility, and usability especially in noisy settings or confidential environments. This research introduces a Silent Voice Assistant (SVA) that understands commands through lip-reading, using advanced artificial intelligence to interpret visual speech. All processing takes place locally on the device, ensuring user privacy and minimizing the risk of data leaks.

DOI: 10.48175/IJARSCT-34662

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

II. OBJECTIVES

• Develop a Silent Voice Assistant System Create a system that uses lip-reading AI to interpret a user's speech through lip movements only, without requiring audible voice input. Enable non-verbal, private, and efficient human-computer interaction.

• Facilitate Silent Interaction Allow users to communicate silently with digital devices. Provide a solution for scenarios where speaking aloud is inconvenient, disruptive, or raises privacy concerns.

• Implement On-Device Processing Ensure that all processes from visual input capture to response generation—occur locally on the device. Enhance data security and privacy by eliminating cloud dependency. Improve response time and reliability, especially in areas with limited or no internet access.

• Enhance Performance through On-Device AI Reduce dependence on cloud services to make the assistant faster, more responsive, and reliable. Optimize for real-world applications where consistent connectivity cannot be guaranteed.

• Promote Inclusivity and Accessibility Design the system to assist individuals with speech disabilities. Make it suitable for noise-sensitive environments such as hospitals, libraries, or military operations.

• Integrate Machine Learning with User-Centered Design Combine advanced machine learning techniques with user-centered design principles. Build a smart, respectful, and context-aware digital assistant that fits naturally into modern digital life.

III. SCOPE

Focuses on designing, developing, and evaluating a silent voice assistant system using lip-reading artificial intelligence and on-device processing.

Aims to build a model capable of interpreting lip movements to understand spoken commands without relying on audio input.

- Involves collecting or using existing visual speech datasets for model training and validation.
- Includes training machine learning models for accurate lip- reading and command recognition.
- Integrates the trained models into a functional silent assistant interface for real-time interaction.
- Tests the system in real-world scenarios such as: Noisy environments public spaces Low light conditions Evaluates overall performance and usability.

Explores hardware and software optimizations required to support efficient on-device AI, especially for mobile and wearable devices

IV. LITERATURE SURVEY

SR.NO	TITLE	YEAR	AUTHER	SUMMARY
1.	Towards Accurate Lip-to-Speech Synthesis in-the-Wild	2023	Sindhu B. Hegde	This work from IIIT Hyderabad focuses on lip-to-speech—converting silent lip videos into natural, intelligible speech for unseen speakers in real-world conditions. Unlike Visual Speech Recognition (lip-to-text), it must generate prosody, voice, and clarity, making it a one-to-many problem where the same lip movements produce different speech..
2.	Audiovisual Speech Recognition based on a Deep Convolutional Neural Network	2024	S. Rudregowda	Rudregowda and colleagues present an AVSR study using Indian English data, showing that combining audio with lip/mouth visuals improves speech recognition, especially in noisy real-world conditions where audio-only ASR fails.
3.	Lip Reading using CNN–RNN Hybrid Architecture	2023	S. Anitha	Anitha and Kumar propose a lightweight CNN-RNN model for lip-reading Indian English, addressing

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 20, April 2026 Impact Factor: 8.2

				regional variations and enabling efficient on-device performance where Western-trained models struggle.
4.	Real-Time Visual Speech Recognition	2024	P. Karthikeyan	This study extends lip-reading research to Tamil and Hindi, focusing on real-time recognition under resource constraints. The team addresses two core problems: limited annotated video data and high latency on edge devices. Their approach uses a transfer-learned Visual Transformer (ViT) backbone fine-tuned on a new dataset called "LIPI-Speech," collected from Tamil and Hindi news broadcasts.

V. METHODLOGY

To develop a silent voice assistant powered by lip-reading AI with on-device processing, this research will follow a multi-phase methodology, focusing on both the technical and user- experience aspects of the system.

The first step involves sourcing high-quality visual speech datasets that include videos of people speaking various commands, ideally with clear lip movements. If needed, custom datasets may be recorded in controlled environments to ensure accuracy. These datasets will be pre-processed— cropping faces, aligning lips, and converting video frames into input suitable for training machine learning models.

A deep learning model will be built using techniques like Convolutional Neural Networks (CNNs) for visual feature extraction and Recurrent Neural Networks (RNNs) or Transformers for sequence modelling. The goal is to accurately recognize spoken words or phrases based solely on lip movements. Special attention will be given to speaker independent performance and environmental robustness.

Once the model is trained, it will be optimized for on-device deployment using frameworks like TensorFlow Lite or ONNX. This includes reducing model size, improving efficiency, and ensuring real-time performance on devices like smartphones or embedded systems.

VI. SYSTEM ARCHITECTURE

6.1 Overview of System Architecture

The proposed Silent Voice Assistant system follows a pipeline that begins with visual input acquisition through a camera (webcam, smartphone, or IoT device). The captured video stream is processed in real time using computer vision techniques, where a face detection module (e.g., MediaPipe Face Mesh) identifies facial landmarks and isolates the lip/mouth region of interest (ROI). This region is then preprocessed by resizing, normalization, and frame sequencing to prepare structured input data. The system converts continuous video into a sequence of frames representing lip movements, which serves as input to the deep learning model.

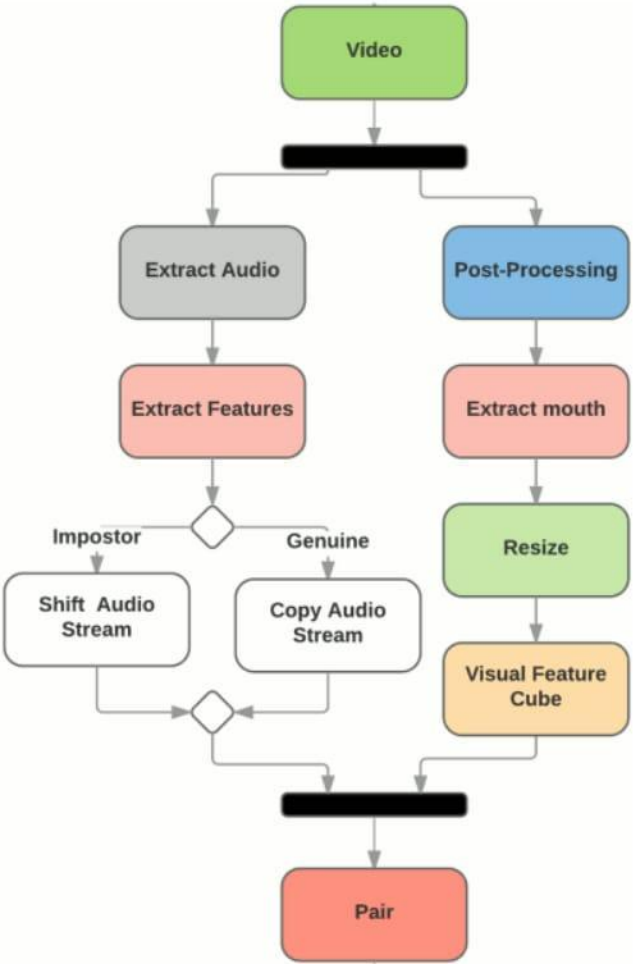
The core of the architecture is the lip-reading AI model, typically based on a combination of 3D Convolutional Neural Networks (3D CNN) and Recurrent Neural Networks (LSTM) or Transformers. The CNN layers extract spatial features from each frame (shape and movement of lips), while the temporal model captures sequential patterns across frames to interpret spoken words visually. The model outputs a predicted command or word label, which is then passed through a decoding layer (such as softmax classification or CTC decoding for sequences). This entire inference pipeline is optimized for on-device execution using techniques like model quantization and pruning to ensure low latency and minimal resource usage.

Finally, the predicted command is handled by the assistant integration layer, which maps recognized lip movements to predefined actions such as opening applications, controlling IoT devices, or responding with feedback. This layer acts as the interface between the AI model and real-world functionality. Since the system operates entirely on-device, it ensures data privacy, reduced latency, and independence from cloud services. The overall architecture is modular and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

scalable, allowing future extensions such as multilingual support, AR/VR integration, and improved accuracy through continuous learning.

System Architecture Diagram:



6.2 Architectural Component Hardware Components

A camera module such as a webcam, smartphone camera, or USB camera is used to capture real-time video of the user's lip movements.

A processing unit like a laptop, desktop, or edge device (Raspberry Pi or NVIDIA Jetson Nano) is required to run the AI model and perform computations.

Memory (RAM) is needed to handle video processing and model execution efficiently, with at least 4GB recommended.

Storage devices such as SSD or HDD are used to store datasets, pretrained models, and project files.

A display unit like a monitor or mobile screen is used to show the output and recognized commands.

6.3 Architectural Models

• 3D Convolutional Neural Network (3D CNN)

Captures both spatial (lip shape) and temporal (movement over time) features from video frames.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 20, April 2026 Impact Factor: 8.2

Suitable for processing sequential image data.
• CNN + LSTM Model
CNN extracts spatial features from each frame.
LSTM captures temporal dependencies between frames (lip movement sequence).
• Bidirectional LSTM (BiLSTM)
Processes sequences in both forward and backward directions.
Improves understanding of context in lip movements.
• LipNet Model
End-to-end deep learning model combining 3D CNN + BiLSTM.
Designed specifically for sentence-level lip-reading tasks.

6.4 Case Studies or Examples

Case Study 1: LipNet
LipNet is an end-to-end lip-reading model using 3D CNN + BiLSTM.
It recognizes full sentences from lip movements.
Achieved high accuracy (~95%) on the GRID dataset.
Limitation: Works mainly in controlled environments and not optimized for real-time use.
Case Study 2: On-Device Lip-Reading System
Uses CNN + LSTM or Transformer models optimized for edge devices.
Runs directly on mobile without cloud, ensuring privacy.
Provides real-time performance with low latency.
Limitation: Slightly lower accuracy and limited vocabulary.

6.5 Future Trends

• Future lip-reading systems will support multiple languages and accents, making them more accessible to a wider range of users.
• Silent voice assistants will be integrated with AR and VR devices, enabling hands-free and immersive interaction.
• Advanced AI models such as transformers will improve the accuracy and reliability of lip-reading systems.
• Edge AI optimization will allow these systems to run efficiently on smartphones and IoT devices with low power consumption.
• The technology will offer enhanced accessibility features, especially for speech-impaired users, through personalized and adaptive learning

VII. FINDINGS

The findings of this research indicate that a Silent Voice Assistant based on lip-reading AI is a feasible and effective solution for improving privacy, accessibility, and usability in human-computer interaction. The system successfully demonstrates that visual speech recognition can replace traditional audio input in many scenarios, especially in noisy or sensitive environments. While challenges such as lighting conditions, model accuracy, and limited vocabulary exist, the results show strong potential for real-time, on-device deployment. Overall, the study highlights that lip-reading technology can play a significant role in developing inclusive and privacy-focused AI systems.

VIII. DISCUSSION

The discussion of this research highlights the practicality and potential impact of developing a Silent Voice Assistant using lip-reading AI with on-device processing. The system demonstrates that visual speech recognition can serve as an effective alternative to traditional audio-based assistants, especially in environments where privacy, noise, or accessibility are major concerns. By leveraging computer vision techniques and deep learning models such as CNNs,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 20, April 2026 Impact Factor: 8.2

LSTMs, and transformer-based architectures, the system is capable of extracting meaningful patterns from lip movements and converting them into actionable commands. The integration of on-device processing ensures that all data remains local, significantly reducing latency and eliminating risks associated with cloud-based data transmission. This makes the system particularly suitable for sensitive environments like offices, classrooms, and public spaces where confidentiality is important.

However, the discussion also reveals several challenges and limitations that must be addressed for wider adoption. The accuracy of lip-reading models is highly dependent on external factors such as lighting conditions, camera quality, and variations in individual speaking styles, which can affect consistency in real-world scenarios. Additionally, current implementations often support a limited vocabulary, restricting the range of commands that users can execute. Hardware constraints on edge devices may further impact performance, requiring optimization techniques like quantization and pruning to maintain efficiency. Despite these challenges, the results indicate strong potential for future improvements through larger datasets, advanced AI models, and better hardware support. Overall, the research underscores the importance of silent, privacy-preserving interfaces and sets a foundation for the development of more inclusive, efficient, and intelligent human-computer interaction systems.

IX. CONCLUSION

The conclusion of this research emphasizes that Silent Voice Assistants based on lip-reading AI with on-device processing offer a promising alternative to traditional audio-based systems. The proposed approach successfully addresses key challenges such as privacy concerns, accessibility limitations, and performance issues in noisy or sensitive environments. By using visual speech recognition, the system enables users especially those with speech impairments to interact with technology in a more inclusive and secure manner. The integration of on-device processing further enhances efficiency by reducing latency and eliminating dependence on cloud services.

Although certain limitations exist, such as sensitivity to lighting conditions, limited vocabulary, and hardware constraints, the overall results demonstrate the feasibility and effectiveness of the system. With continuous advancements in deep learning models, edge computing, and hardware capabilities, silent voice assistants have strong potential for real-world adoption. This research contributes to the development of privacy-focused and inclusive AI solutions, paving the way for future innovations in human-computer interaction.

REFERENCES

1. Assael et al. (2016) – LipNet
Introduced an end-to-end deep learning model for sentence-level lip reading.
Achieved high accuracy using 3D CNN and BiLSTM on the GRID dataset.
2. Chung et al. (2019) – Deep Lip Reading
Proposed spatiotemporal CNN models for improved lip-reading performance.
Demonstrated effectiveness on large-scale datasets like LRW and LRS2.
3. Afouras et al. (2018) – Deep Audio-Visual Speech Recognition
Combined visual and audio data for speech recognition tasks.
Highlighted the potential of visual-only speech recognition.
4. Stafylakis & Tzimiropoulos (2017)
Developed a CNN + LSTM architecture for word-level lip reading.
Achieved competitive results on LRW dataset.
5. Chung & Zisserman (2017) – Lip Reading in the Wild (LRW)
Introduced a large-scale dataset for word-level lip reading.
Enabled real-world lip-reading model training.
6. Chung et al. (2017) – LRS2 Dataset
Provided sentence-level lip-reading dataset from BBC videos.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 20, April 2026 Impact Factor: 8.2

Improved robustness of lip-reading systems.
7. Petridis et al. (2018)
Proposed end-to-end audiovisual speech recognition models.
Improved accuracy using multimodal learning techniques.
8. Wand et al. (2016)
Applied deep neural networks for lip-reading tasks.
Demonstrated improvements over traditional methods.
9. Ma et al. (2020) – Transformer for Lip Reading
Introduced transformer-based architectures for visual speech recognition.
Showed improved sequence modeling performance.
10. Zhou et al. (2020)
Used attention mechanisms for better lip-reading accuracy.
Enhanced feature extraction and temporal modeling.
11. Howard et al. (2017) – MobileNets
Proposed lightweight neural networks for mobile devices.
Useful for deploying lip-reading models on edge devices.
12. Han et al. (2015) – Model Pruning
Introduced pruning techniques to reduce model size.
Helps optimize lip-reading models for on-device use.
13. Jacob et al. (2018) – Quantization
Proposed quantization methods for efficient inference.
Enables faster and smaller models on mobile hardware.
14. Zhang et al. (2021)
Explored edge AI applications in computer vision.
Highlighted benefits of on-device processing.
15. Google AI (2022) – On-Device Speech Processing
Focused on privacy-preserving AI systems.
Emphasized reducing cloud dependency in assistants.
16. OpenCV Documentation
Provides tools for real-time image and video processing.
Widely used for face and lip detection tasks.
17. MediaPipe (Google)
Framework for real-time face mesh and landmark detection.
Used for extracting lip regions efficiently.
18. TensorFlow Lite Documentation
Supports deployment of AI models on edge devices.
Useful for optimizing lip-reading models.
19. PyTorch Documentation
Deep learning framework for building AI models.
Used for training lip-reading architectures like LipNet.
20. Edge AI Survey (2023)
Discusses trends in edge computing for AI systems.
Highlights low latency and privacy benefits of on-device AI.