International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Silent Speech Recognition Using AI and ML

K Swetha Sailaja*[1], BANDARI ARUN*[2], DANTURI ABHINAV*[3], AKKENAPALLY CHARAN*[4]

*[1] Assistant Professor of Department of CSE (AI & ML) Of ACE Engineering College, India.

*[2,3,4] Students of Department of CSE (AI & ML) Of ACE Engineering College, India.

Abstract

Lip movement interpretation for speech recognition is a cutting-edge technology that enables communication without sound, benefiting individuals with speech and hearing impairments and enhancing human-computer interaction. This project focuses on Visual Speech Recognition (VSR) by analysing lip movements and converting them into accurate text using deep learning models. The system processes video frames, extracting key lip features using OpenCV, Dlib, and MediaPipe, and is trained on datasets such as the GRID Corpus and LRS2 to achieve robust recognition. The technology has real-world applications in communication aids, accessibility tools, and silent interfaces in sensitive environments.

Introduction

Speech is a primary medium of communication, but individuals with speech disabilities face significant challenges. Silent Speech Recognition (SSR) presents an innovative approach to enable non-verbal communication using AI and ML. SSR systems focus on interpreting non-auditory inputs like facial muscle signals and lip movements to generate text or speech. With the advancement of deep learning, accurate interpretation of such inputs is now feasible. This project proposes a system that leverages these advancements to assist individuals with speech impairments and provide silent interfaces for secure or noise-free communication.

Background

Traditional Automatic Speech Recognition (ASR) systems rely heavily on audio input. In noisy environments or for individuals who cannot produce sound, these systems become ineffective. Silent Speech Interfaces (SSIs) aim to overcome these limitations by using non-acoustic features such as lip reading or facial EMG signals. Recent improvements in computer vision, deep learning, and real-time processing capabilities have enabled the development of systems that can achieve SSR with promising accuracy and latency.

Literature Review

Silent Speech Using Surface Electromyography (sEMG) – 2022

One notable study from 2022 focuses on silent speech recognition using surface electromyography (sEMG) signals. In this method, sEMG sensors are placed around the facial muscles to capture the subtle electrical signals generated during silent articulation. These signals are processed through Convolutional Neural Networks (CNNs), which extract spatial features corresponding to specific speech patterns. The model performs well when trained and tested on the same individual, demonstrating high accuracy in recognizing silently spoken words. However, it suffers from poor generalization across different users due to natural variations in muscle activity. Additionally, the requirement of specialized hardware (sEMG sensors) restricts its practical deployment in everyday applications.

Lip Reading with Deep Learning – Chung et al. (2018)

Chung and colleagues (2018) introduced a powerful visual speech recognition system that utilizes lip reading with deep learning. Their approach combines 3D Convolutional Neural Networks (CNNs) for extracting spatial and temporal features from lip movements with Long Short-Term Memory (LSTM) networks to handle sequence modeling. The model was trained on datasets like GRID and Lip Reading in the Wild (LRW), achieving over 90% word recognition accuracy under controlled conditions. Despite its impressive performance, the model is sensitive to external factors such as lighting, facial occlusions, and non-frontal face angles, which limit its effectiveness in real-world environments.

Deep Speech – Hannun et al. (2014)

Although not directly applicable to silent speech, Deep Speech by Hannun et al. (2014) remains a foundational work in end-to-end speech recognition. The system employs Recurrent Neural Networks (RNNs) with a Connectionist Temporal Classification (CTC) loss to transcribe spoken words from audio spectrograms. Trained on large-scale datasets like LibriSpeech, Deep Speech set a new benchmark in traditional audio-based speech recognition. However, since it relies entirely on acoustic signals, it cannot be used for silent or non-vocal speech input, underscoring the need for alternative modalities such as video or biosignals in SSR.

Whisper – OpenAI (2022)

Whisper, developed by OpenAI in 2022, is a highly robust and multilingual audio speech recognition model based on a Transformer architecture. It was trained on an extensive dataset containing over 680,000 hours of web-scraped multilingual audio. Whisper supports speech-to-text transcription, language identification, and translation, performing well even in noisy environments. Despite its sophistication, Whisper is exclusively audio-based and not suited for silent speech recognition. Nevertheless, its architecture demonstrates the potential of Transformer models for sequence processing, which could be adapted for visual or multimodal SSR systems in future research.

LipNet – Assael et al. (2016)

LipNet by Assael et al. (2016) was one of the first deep learning models to perform full-sentence lip reading using video input. It integrates spatiotemporal CNNs with Gated Recurrent Units (GRUs) and uses the CTC loss function to predict complete sentences directly from lip movement sequences. Trained on the GRID corpus—consisting of structured, fixed-vocabulary sentences—LipNet achieved over 93% accuracy in controlled testing scenarios. However, its effectiveness is limited by the constrained nature of the dataset, and it struggles to generalize to more natural or spontaneous speech patterns.

Comparison Table

Title	Input Type	Model Used	Dataset	Methodology	Findings
Silent Speech Using sEMG (2022)	sEMG signals	CNN	Custom Dataset	Sensors capture facial muscle activity; CNN extracts features and maps them to text	High accuracy on speaker-dependent data; poor cross-user performance

Lip Reading with Deep Learning (2018)	Video (lip movements)	3D CNN + LSTM	GRID, LRW	3D CNN for spatial-temporal features; LSTM for sequence modeling	90%+ accuracy in controlled settings; sensitive to real-world variability
Deep Speech (2014)	Audio	RNN + CTC	LibriSpeech	Uses spectrograms processed via RNN; CTC loss for sequence transcription	High accuracy for audio speech; not usable for silent speech
Whisper (OpenAI, 2022)	Audio	Transformer	680,000+ hrs of audio	Transformer model for multilingual ASR; end-to-end training with transcription and translation	Very robust and multilingual; not applicable to silent speech input
LipNet (2016)	Video (lip movements)	Spatiotemporal CNN + GRU + CTC	GRID	CNN captures video frames; GRU models sequences; CTC for full-sentence prediction	93%+ accuracy on GRID dataset; struggles with spontaneous and diverse real-world speech

Research Gaps

1.Speaker Dependency:

- Most SSR models are trained on specific speakers, making them less effective when tested on new individuals.
- Differences in lip shape, speaking style, and articulation patterns reduce generalization.

2. **Insufficient Dataset Diversity:**

- Available datasets like GRID and LRS2 lack diversity in ethnicity, age, and lighting conditions.
- Many datasets are small-scale or limited to specific languages (mostly English), hindering broader applicability.

3. **Visual Ambiguity (Homophenes):**

- Many phonemes (e.g., /b/ and /p/, /m/ and /n/) look nearly identical when spoken, which causes confusion during recognition.
- This visual similarity limits model accuracy, especially for short or single-word recognition.

4. **Occlusion and Obstruction:**

- Facial obstructions such as hands, masks, glasses, or facial hair can obscure lip movements.

- Changes in head pose, camera angle, or partial face visibility degrade performance significantly.

5. **Lighting and Environmental Sensitivity:**
   - Variations in lighting conditions (e.g., natural vs. artificial light, shadows) affect lip visibility and model performance.
   - SSR systems often lack robustness to real-world visual noise.
6. **Real-Time and Low-Power Inference Limitations:**
   - Achieving real-time SSR on edge devices (e.g., mobile phones or AR glasses) is computationally demanding.
   - Many models are not optimized for latency or power consumption, making them impractical for wearable use.
7. **Limited Multilingual and Multimodal SSR Research:**
   - Most SSR research focuses on English, with very limited support for other languages or multilingual environments.
   - Integration with other modalities like EMG or infrared imaging is still underdeveloped.
8. **Lack of Contextual and Semantic Understanding:**
   - Current models often fail to incorporate linguistic context, which could help in resolving homophones and increasing accuracy.
   - There is minimal use of language models or grammar rules in silent speech systems.
9. **Ethical and Privacy Considerations:**
   - Silent speech systems involving video or EMG capture raise concerns about user consent, data security, and misuse.
   - There is a need for privacy-preserving machine learning techniques in SSR.
10. **Limited Real-World Deployment:**
    - Few systems have transitioned from research to deployment in healthcare, accessibility, or consumer tech.
    - Lack of user-centered design and usability testing hinders practical adoption.

Proposed System

To address these gaps, a proposed system focuses on Silent Speech Recognition (SSR) using advanced Artificial Intelligence (AI) and Machine Learning (ML) techniques. It is designed to interpret speech without relying on audible input by analyzing visual cues such as lip movements and optionally integrating Electromyography (EMG) signals from facial muscles. Real-time video input is captured through webcams or smartphone cameras, and preprocessing techniques involving OpenCV, Dlib, and Mediapipe are used for lip detection and facial landmark extraction. Convolutional Neural Networks (CNNs) are employed to extract meaningful spatial features from each video frame, which are then processed using Long Short-Term Memory (LSTM) or Transformer architectures to model the temporal sequence of speech patterns. The system is trained using large-scale visual speech datasets such as GRID Corpus and LRS2 to ensure robust

performance and speaker-independent generalization. Finally, the processed data is converted into textual output or synthesized into speech using Natural Language Processing (NLP) and Text-to-Speech (TTS) modules. This system is especially beneficial in environments where sound-based communication is not feasible, and it provides an accessible solution for individuals with speech impairments. It also holds promise for use in security-sensitive and noise-restricted environments, offering real-time, accurate, and adaptive silent communication technology.

Conclusion and Future Scope

This literature survey outlines the current landscape of Silent Speech Recognition and highlights both achievements and challenges in the domain. The findings show that while deep learning approaches like CNNs, LSTMs, and Transformers have enhanced performance, most models still suffer from speaker dependency and data limitations. The proposed project aims to overcome these limitations by integrating visual-only features (e.g., lip movements) through real-time computer vision tools such as OpenCV, Dlib, and Mediapipe. By eliminating the reliance on audio and enhancing model generalization, this system becomes more adaptable across users and environments. Future scope includes creating a large-scale, diverse dataset for silent speech, advancing real-time processing capabilities for mobile and wearable platforms, and exploring multimodal fusion techniques to further improve accuracy and robustness.

References

1. Kumar, A., & Raj, R. (2022). Silent Speech Recognition Using Surface Electromyography and Deep Learning. *IEEE Access*, 10, 30482–30491.

2. Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Lip Reading Sentences in the Wild. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

3. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. *OpenAI*.

5. Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5), 2421–2424.

6. Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep Audio-Visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.