International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI for Silent Speech Recognition (Lip Reading AI)

Dhvaneeth P Banakar[1], Ganavi M[2], Gurukiran S G[3], GaganDeep T V[4], Jagath M D[5]

[1,3,4,5]UG student, Dept. of CS&E, Jawaharlal Nehru New College of Engineering (JNNCE), Visvesvaraya Technological University (VTU), Karnataka, India.
[2]Associate Professor, Dept. of CS&E, Jawaharlal Nehru New College of Engineering (JNNCE), Visvesvaraya Technological University (VTU), Karnataka, India.

Abstract
Silent Speech Recognition using artificial intelligence has made a progressive technology for interpreting spoken language without relying on acoustic signals. This work influence the planning and building of a lip-reading system that leverages computer view and deep learning mechanism to translate visual speech cues into text. The suggested model integrates spatial and sequential neural networks to learn visual and periodic dynamics of lip movements, enabling accurate recognition even in noisy or sound-restricted environments. Beyond assisting individuals with speech or hearing impairments, the system demonstrates applicability in secure communication, healthcare, and defense, where silent or noise-independent interaction is essential. Experimental evaluation highlights the solidity of the model in handling variations in speakers, lighting, and video quality, with promising results in real-time recognition of isolated words and continuous speech. The study underscores the potential of SSR to enhance accessibility, privacy, and inclusivity in human–computer interaction, while paving the way for future improvements through multilingual support, multimodal integration, and deployment on portable devices.

Keywords: Silent Speech Recognition, Convolutional Neural Networks, Recurrent Neural Networks, Spatiotemporal Feature Extraction, Noise Independent Communication.

1. Introduction

Artificial Intelligence (AI) has revolutionized the way humans interact with machines by enabling systems to understand, interpret, and respond to complex inputs. One of the emerging fields in this domain is Silent Speech Recognition (SSR), also known as Lip-Reading AI, which focuses on interpreting speech by analyzing a speaker's lip movements without relying on audio signals. This technology integrates visual computing and advanced machine learning models, such as Convolutional Neural Networks and Recurrent Neural Networks, to decode visual cues and translate them into textual or spoken output. By analyzing spatiotemporal patterns in facial movements, these systems can effectively perform speech recognition in silent or noisy environments. The demand for silent communication systems has grown rapidly with advancements in human computer interaction and assistive technology. Unlike traditional audio-based recognition systems that struggle in noisy or privacy-sensitive environments, AI-driven lip reading offers a noise-independent and privacy-preserving solution. The integration of spatiotemporal feature extraction,

facial landmark detection, and sequence modeling enables these systems to deliver real-time, accurate speech interpretation. As a outcome, silent speech recognition has turned a promising field with transformative potential across multiple domains.

A. Background

In our fast technology world, developing an efficient intelligent system of interpreting speech through lip movements captured via video input. It leverages modern deep learning tools like CNNs for spatial information extraction and RNNs for temporal sequence learning. The ambition following this work arises from the limitations of traditional acoustic-based speech recognition systems, which typically fail high background noise in the environment or restricted audio use. By focusing on visual information processing, this project aims to bridge the communication gap for individuals with speech or hearing impairments.

The role of CNNs & RNNs holds significant importance in creating inclusive and adaptive communication systems. It enables individuals with speech or hearing disabilities to effectively interact more with digital devices and other people. Additionally, in settings where audio communication is not feasible—such as military operations, confidential workplaces, or noisy industrial environments—AI-based silent speech recognition provides an effective alternative. The system also contributes to improved privacy, accessibility, and efficiency in human–machine communication, making it a crucial step toward next-generation intelligent interfaces.

B. Applications

AI-based Silent Speech Recognition has wide-ranging applications across various domains. In assistive technology, it can be used to help people with speech or hearing impairments communicate seamlessly through visual speech input. In defense and security, it enables silent command recognition where verbal communication may compromise safety. In healthcare, it can helps patients who have lost their voice due to surgery or illness. Additionally, the mechanism can be applied in augmented and virtual reality (AR/VR) environments, surveillance systems, and hands-free device control, offering new ways to interact with digital systems in both silent and noisy conditions.

2. Literature Survey

The advancement of Artificial Intelligence (AI) and Deep Learning (DL) has considerably reforms the field of speech and facial recognition, paving the way for Silent Speech Recognition (SSR) or lip-reading systems. These systems interpret human speech using only visual cues—such as lip, tongue, and facial movements—without relying on audio signals. Lip reading serves as a crucial communication technology in noisy, privacy-sensitive, or silent environments and provides an assistive interface for individuals with speech or hearing impairments. The integration of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures has allowed computers to extract spatiotemporal features from facial videos, enabling accurate mapping of visual data to textual or phonetic outputs. Recent research has made an improving lip region extraction, enhancing video frame quality, and refining classification models to increase the robustness and generalization of visual speech recognition systems across speakers, languages, and environmental conditions.

Lip Reading Using Deep Learning [1], Cao and Yan introduces LipReader++, a novel Neural Computing platform that blends 3D Convolutional Neural Networks (3D CNNs) with Transformer architectures to improve visual speech recognition. The model is intended to capture both the spatial and temporal dynamics of lip movements, enabling it to recognize spoken words from visual cues alone with high

accuracy. By incorporating data augmentation and regularization techniques, LipReader++ demonstrates strong generalization across diverse datasets, including GRID and LRS2, achieving an accuracy of 93% on controlled conditions and 78.5% in real-world scenarios. By comparing with the earlier approaches such as LipNet and WAS, the model considerably reduces word error rates while maintaining robustness under variations in lighting, speech speed, and speaker diversity. The research highlights the practical potential of visual speech recognition systems for applications in assistive technology, education, and secure communication, while also identify future directions such as multimodal integration and improvements under challenging conditions.

LipReader++ delivers superior performance in both reliability and effectiveness while maintaining scalability for real-time applications. Its results confirm that deep visual models can successfully interpret speech without any audio input, offering major implications for assistive technologies**,** real-time communication systems, and privacy-preserving interfaces. The authors determined their approach establishes a strong foundation for future developments in multimodal speech recognition, with potential extensions in combining visual, auditory, and contextual cues to further enhance the robustness and adaptability of AI-driven lip-reading systems.

Enhancement of Human Face Mask Detection Performance by Using Ensemble Learning Models [2], Xinyi Gao**,** Minh Nguyen**,** and Wei Qi Yan presents an advanced deep learning-based approach to accurately detect face masks, dealing with the global need of public health safety during pandemics. The authors introduce a hybrid model called Swin+YOLOv8, which integrates the Swin Transformer and YOLOv8 frameworks to improve detection accuracy, precision, and mean average precision (mAP). The Swin Transformer component leverages hierarchical vision transformers and shifted window self-attention mechanisms for robust feature extraction, while YOLOv8 serves as the detection head for fast and reliable classification. The model was trained on a labeled mask dataset containing over 4,000 facial images and tested using mAP50 and mAP50–95 metrics. Experimental results demonstrated that Swin+YOLOv8 achieved 96.1% accuracy and an mAP of 0.962, outperforming baseline models like YOLOv7, YOLOv8n, and YOLOv8n+DCNv2.

Building on this foundational knowledge, the study also provides Trough comparison with otherpopular deep learning methods, including YOLOv3, YOLOv4, SSD, and RetinaFace, highlighting how Transformer-based architectures outperform traditional CNN models in complex visual recognition tasks. Model demonstrated superior generalization capability in handling occlusions**,** varying face angles, and lighting conditions. The authors deduce the proposed ensemble model effectively balances detection accuracy and computational efficiency, offering a scalable and practical solution for real-time face mask detection. This work contributes significantly to the advancement of AI-driven public health surveillance systems and underscores the potential of hybrid deep learning architectures in enhancing the robustness of object detection applications.

HMM Based Audio Visual Speech Recognition Integrating Geomertic and Apperance Based Visual Features [3], Michael T. Chan presents a hybrid approach for enhancing speech recognition performance by combining audio and visual (lip movement) information. The study unveils a new visual attribute extraction framework that combines both geometric-based features like lip height and width—and appearance-based features derived from pixel intensity values around the lip region. Using a contour-based lip-tracking algorithm, the system accurately identifies the region of interest (ROI) for feature extraction, compensating for variations in scale, translation, and illumination. The proposed feature representation is designed to improve recognition accuracy by capturing complementary spatial and appearance

information of the lips, teeth, and tongue.

Test-based evaluations were demonstrated using Hidden Markov Models (HMMs) for both visual-only and audio-visual speech recognition tasks involving isolated digits. Results demonstrated that the hybrid feature combination (geometric + pixel-based) achieved the highest recognition accuracy of 98% in visual-only mode and maintained strong performance in noisy audio environments. In audio-visual conditions, the bimodal recognizer consistently outperformed audio-only systems under various signal-to-noise ratios, confirming the robustness of the hybrid approach. The study concludes that integrating geometric and appearance-based visual features with HMM-based models significantly enhances recognition accuracy and noise robustness, establishing a foundation for real-time audio-visual speech recognition systems that can operate effectively in complex and noisy communication environments.

Face Detection and Recognition from Distance on Deep Learning [4], Hui Wang and Wei Qi Yan offer a thorough and insightful review on improving the robustness of human face recognition systems in diverse distances, angles, and lighting conditions using deep learning techniques. The authors employ a Convolutional Neural Network (CNN) framework inspired by the Single Shot MultiBox Detector (SSD) architecture to achieve real-time face detection and recognition. The model integrates multiple convolutional and pooling layers followed by dual-layer Multilayer Perceptrons (MLPs) for classification and localization tasks. To enhance accuracy and reduce overfitting, the study applies data augmentation techniques such as random cropping, rotation, and scaling, combined with dropout regularization. The system was trained on a diverse dataset consisting of faces captured from different distances and environments, leading to improved adaptability to real-world variations.

The progressive results demonstrate the proposed model achieved an precision of 90.18% in recognizing human faces at various distances, overtaking traditional machine learning types that struggle with occlusion, facial expressions, and environmental noise. The study highlights that the proportion of the face area in an image directly influences recognition accuracy, with optimal performance observed when the face occupies around 40% of the frame. Furthermore, the research emphasizes that deep learning-based detection models significantly surpass conventional approaches in terms of precision and resilience, even under partial occlusion or complex lighting. The authors conclude that optimizing CNN-based architectures, reducing training dataset size, and exploring hybrid neural network models could further enhance the efficiency and scalability of real-time distance-based face recognition systems.

LipNet: Sentence Level LipReading [5], Yannis M. Assael and his colleagues offer an in-depth look at the journey of end-to-end deep learning model capable of performing sentence-level lipreading directly from video frames. Unlike previous approaches limited to isolated word or phoneme classification, LipNet maps continuous sequences of video frames to textual sentences using an integrated architecture combining Spatiotemporal Convolutional Neural Networks (STCNNs), Bidirectional Long Short-Term Memory (Bi-LSTM) networks, and the Connectionist Temporal Classification (CTC) loss. This combination enables the model to capture both spatial and temporal visual features of lip movements while handling variable-length input and output sequences without requiring pre-segmented data. Trained on the GRID corpus, LipNet achieved a remarkable 93.4% word-level accuracy and 6.6% Word Error Rate.

Additionally, the authors authors also apply saliency visualization techniques to interpret the network's attention across frames, showing that it focuses on phonologically significant regions such as the lips and tongue. Further phoneme-level analysis using viseme confusion matrices confirmed that most recognition errors occurred within similar visual phoneme groups, underscoring the model's linguistic consistency. The research establishes a foundation for future audio-visual speech recognition systems, suggesting that

combining LipNet with audio input could further enhance robustness in noisy or silent environments. The study marks a major milestone in the field of visual speech recognition, paving the way for applications in assistive communication**,** silent dictation**,** and secure voice interfaces.

Facial Emotion Recognition Using Ensemble Learning [6], GuanQun Xu and Wei Qi Yan aim to enhance recognition accuracy without increasing computational complexity by training multiple mini-Xception models each specialized for a specific emotion—and combining their outputs through a confidence-based weighted voting mechanism. This ensemble transforms the multi-class classification task into multiple binary sub-tasks, allowing each expert model to focus on distinguishing one emotion from all others. The study employs the FER2013 dataset and uses extensive data augmentation techniques, such as SMOTE oversampling, brightness and contrast adjustments, and elastic deformations, to address class imbalance and improve model generalization. Each expert model achieved around 85% binary accuracy, while the overall ensemble attained 78% multi-class accuracy, demonstrating robust performance across diverse facial expressions.

Additionally, authors review using curves shows that the ensemble effectively enhances precision and recall through confidence-based weighting, particularly improving recognition for *fear* and *neutral* emotions. However, the study notes the constraints, including computational overhead and difficulty in capturing subtle or overlapping emotional states due to binary classification design. Overall, the paper underscores the efficiency of ensemble learning in balancing model robustness.

Face Detection and Recognition from Distance Based on Deep Learning [7], Hui Wang and Wei Qi Yan dive into a robust framework for recognizing human faces captured at varying distances and angles using deep learning techniques. The authors emphasize that traditional face recognition algorithms are often unreliable under conditions such as occlusion, illumination changes, or facial variations, and propose a Convolutional Neural Network (CNN) model inspired by the Single Shot MultiBox Detector (SSD) architecture to overcome these challenges. The model applies pattern extracting layer with both large and small kernels for comprehensive spatial feature extraction, coupled with dropout regularization to avoid overfitting. A structure of dual Multilayer Perceptron (MLP) is used to handle classification and localization tasks simultaneously. Data enrichment techniques such as random cropping, rotation, and scaling were applied to enhance dataset diversity, allowing model to seek effectively from limited samples. Additionally, progressive results demonstrate the proposed model achieved a recognition accuracy of 90.18% on test videos, successfully identifying faces across different distances, sizes, and lighting conditions. The authors highlight the correlation between the face-to-frame area ratio and recognition accuracy, noting optimal results when the face occupied approximately 40% of the image.

An Automatic Lip Reading for Short Sentences Using Deep Learning Nets [8], Maha A. Rajab and Kadhim M. Hashim dive into a deep learning-based system for recognizing and classifying short English sentences from video input without relying on audio. The study highlights the growing significance of lip reading in environments where audio data is unavailable or unreliable, such as hearing-impaired users or noisy surroundings. The proposed system uses the Viola-Jones algorithm for face detection, followed by 68-point facial landmark detection to extract the lip region accurately. To address issues like facial hair and lighting variation, the system enhances contrast in the lip region before feeding it into two deep learning models AlexNet and VGG-16 for classification. The dataset comprises 39 participants (32 males and 7 females) each repeating ten short sentences five times, resulting in over 130,000 video frames for training and testing.

Additionally, experimental evaluation based on metrics like accuracy, precision, recall, and specificity

demonstrates that AlexNet achieved 90.00% accuracy, outperforming VGG-16, which achieved 82.34% accuracy. The authors suggest future improvements by designing specialized deep networks for lip movements and integrating multimodal inputs (audio and visual) to enhance recognition performance. Overall, the work contributes a practical and efficient deep learning framework for short-sentence lip-reading systems, demonstrating strong potential in assistive communication and silent speech recognition.

3. Summary of Literature Survey

The rapid progress in deep learning-based lip reading, focusing primarily on improving accuracy, robustness, and real-time performance. Early approaches, such as HMM-based audio-visual recognition (Chan, 2003), combined geometric and appearance-based features to improve recognition under noisy conditions. Later, deep CNN models like LipNet achieved sentence-level lip reading by jointly learning spatial and temporal dependencies, surpassing human lipreaders in accuracy. Subsequent studies introduced hybrid models combining CNNs, RNNs, and Transformers, enabling more accurate spatiotemporal feature extraction and improving recognition rates for continuous speech. Other works, such as those by Wang & Yan (2020) and Xu & Yan (2021), applied ensemble and distance-based learning for facial and emotion recognition, demonstrating how deep architectures generalize across varying facial angles, lighting conditions, and expressions. These studies collectively emphasize the growing effectiveness of end-to-end deep learning pipelines for real-world speech and facial analysis tasks.

Moreover, researchers proposed refined preprocessing pipelines including face detection using Viola-Jones, landmark extraction, and contrast enhancement to address challenges like occlusion from mustaches and facial hair. Comparative analyses of models like AlexNet, VGG-16, ResNet, and Bi-LSTM reveal that AlexNet achieved the best performance (90% accuracy) in short-sentence lip reading. Similarly, LipReader++ (Cao & Yan, 2022) achieved over 93% accuracy using CNN-Transformer fusion, while ensemble CNN models improved emotion detection reliability. Collectively, these studies illustrate that the integration of visual feature engineering, hybrid deep learning models, and ensemble optimization has significantly advanced the state of automatic lip reading and facial recognition. The field continues to move toward multimodal and transformer-based frameworks capable of supporting real-time, speaker-independent, and language-adaptive silent speech recognition systems.

4. Table of Summary

Table 1.1 Literature Survey

SI No.	Authors	Research Focus	Remarks
[1]	Yue Cao and Wei Qi Yan [1]	Deep learning–based lipreading (LipReader++).	Achieves high accuracy and robustness, surpassing existing models.
[2]	Xinyi Gao, Minh Nguyen, and Wei Qi Yan [2]	Deep learning–based face mask detection (Swin+YOLOv8).	Achieved 96.1% accuracy and high robustness, outperforming existing models.
[3]	Michael T. Chan [3]	Audio-visual speech recognition using hybrid visual cues.	Achieved 98% accuracy and outperformed audio-only models in noise.
[4]	Hui Wang and Wei	Real Time Face Detection and	Achieved 90.18% accuracy

	Qi Yan [4]	recognize the varying distances	with robust performance under occlusion and lighting changes.
[5]	Yannis M. Assael, Shillingford, Shimon Whiteson, Nando de Freitas	Deep learning–based sentence-level lip reading	Achieved 93.4% accuracy, outperforming humans and earlier models.
[6]	GuanQun Xu and Wei Qi Yan [6]	Enhancing lightweight facial emotion recognition accuracy.	Achieved 78% accuracy with strong results for key emotions.
[7]	Hui Wang and Wei Qi Yan [7]	Real-time face detection and recognition from varying distances.	Achieved 90.18% accuracy with strong robustness.
[8]	Maha A. Rajab and Kadhim M. Hashim [8]	Deep learning–based lip reading for short sentences.	AlexNet achieved 90% accuracy, outperforming VGG-16.

5. Conclusion

The literature review underscores a Silent Speech Recognition (SSR) using Lip-Reading AI highlights the significant advancements and persistent challenges in developing systems that can interpret speech from visual cues alone. Despite progress in deep learning architectures such as CNNs, RNNs, and transformers, the problem of inter-speaker variability, lighting inconsistencies, and homophones remains a considerable obstacle. However, the increasing availability of datasets, improved preprocessing methods, and powerful computational tools are helping bridge these gaps, bringing SSR closer to practical, real-world applications.

The transformative Future of Artificial Intelligence in enabling Silent Speech Recognition (SSR) through lip-reading technologies. By integrating deep learning models such as Convolutional Neural Networks and Recurrent Neural Networks, the study demonstrates how visual features of lip movements can be effectively mapped to textual outputs, achieving high recognition accuracy even in noise-sensitive or audio-restricted environments. The system not only advances speech recognition technology but also offers a significant contribution toward improving communication accessibility for individuals with speech and hearing impairments.

The formulated model establishes a foundation for realistic applications across diverse fields, including healthcare, defense, and human–computer interaction, where privacy and silence are critical. Its robustness against speaker variability, lighting conditions, and visual inconsistencies reflects the progress in vision-based recognition systems. However, despite these advancements, there remain challenges such as optimizing model generalization, expanding language diversity, and enhancing real-time processing capabilities. Addressing these issues will be crucial to improving the scalability and efficiency of SSR systems.

6. References

1. Mitali, VK. Mbise, “The role of it professional certifications in instructors’ teaching quality,” The

International Journal of Education and Development using Information and Communication Technology, vol. 17, no. 1, pp. 176–187, 2021.

2. S. Shi, Y. Wang, C. Zou, and Y. Tian, "AES RSA-SM2 algorithm against man-in-the-middle attack in IEC 60870-5-104 protocol," Journal of Computer and Communications, vol. 10, no. 1, pp. 27–41, 2022.
3. E. W. Lee and G. A. Seomun, "Structural model of the healthcare information security behavior of nurses applying protection motivation theory," International Journal of Environmental Research and Public Health, vol. 18, no. 4, p. 2084, 2021.
4. J. Chen, F. Zhao, and H. Xing, "Research on security of mobile communication information transmission based on heterogeneous network," International Journal of Network Security, vol. 22, no. 1, pp. 145–149, 2020.
5. C. Han, X. Yang, and W. Hu, "Chaotic reconfigurable ZCMTprecoder for OFDM data encryption and PAPR reduction,"Optics Communications, vol. 405, no. 2, pp. 12–16, 2023.
6. S. Chen and Z. H. O. N. G. Xian-xin, "Research of cipher chipcore for sensor data encryption[J]," IEEE Sensors Journal, vol. 16, no. 12, 2024.
7. O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in Proc. British Machine Vision Conf. (BMVC), 2015, pp. 1–12.
8. Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 2892–2900.
9. M. Turk and A. Pentland, "Eigenfaces for recognition," J. Cognit. Neurosci., vol. 3, no. 1, pp. 71–86, Jan. 1991.
10. V. Suresh, S. C. Dumpa, C. D. Vankayala, H. Aduri, and J. Rapa, "Facial recognition attendance system using Python and OpenCV," in Proc. Int. Conf. Comput. Sci. Appl., 2023.
11. S. Sawhney, K. Kacker, S. Jain, S. N. Singh, and R. Garg, "Real-time smart attendance system using face recognition techniques," in Proc. Int. Conf. Comput. Intell. Data Sci. (ICCIDS), 2019, pp. 263–268.
12. TensorFlow Developers, "TensorFlow Lite," 2024. [Online]: https://www.tensorflow.org/lite/ [12] H. Kaur and D. Soni, "Automated attendance system using machine learning and face recognition," Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 3, no. 1, pp. 1–2018.