INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

SilentSpeak: Real-Time Lip Reading and Speech Prediction Using Deep Neural Networks

[1]Ashish Das, [2]Aniruddha Pattnaik, [3]Jhumpa Dutta, [4]Subham Kabiraj, [5] Niladri Das

[1]Assitant Professor, [2]Student, [3] Student, [4]Student, [5]Student
[1]Computer Science & Engineering ,[2 3 4 5] Computer Science & Engineering

[1]Durgapur Institute of Advanced Technology & Management, Rajbandh, India
[2 3 4 5]Durgapur Institute of Advanced Technology & Management, Rajbandh, India

Abstract : Communication through speech is one of the most natural ways people interact. However, it becomes difficult in noisy places or for those with speech and hearing impairments. This paper presents SilentSpeak, a visual speech recognition system that uses AI to understand speech from lip movements without needing audio input. The system uses deep learning and computer vision techniques to recognize spoken words and sentences by analysing facial and lip movements in real time. SilentSpeak has a preprocessing pipeline that extracts video frames, converts them to grayscale, detects the face and lip regions, normalizes the data, and aligns the visual information over time. The processed data is sent through a deep learning setup that includes Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and LipNet-inspired sequence modelling to capture both the spatial and temporal aspects of lip movements. The model trains on labelled visual speech datasets to improve recognition accuracy and sentence prediction. The system works well in places with little or no sound, making it useful for assistive communication, security applications, silent human-computer interaction, and accessibility technology. Tests show that SilentSpeak effectively recognizes visual speech while processing in real time. By connecting human communication with artificial intelligence, SilentSpeak shows the potential of silent speech interfaces as a new technology for inclusive and smart communication systems.

IndexTerms **- Artificial Intelligence (AI), Deep Learning, Human-Computer Interaction (HCI), Long Short-Term Memory (LSTM), Lip Reading, LipNet, Silent Speech Recognition, 3D Convolutional Neural Network (3D CNN), Visual Speech Recognition (VSR)**

I. INTRODUCTION

INTRODUCTION

Inspired by the human ability of multimodal perception, where both visual and auditory cues contribute to speech understanding, significant research has been conducted in the field of audio-visual speech processing [1]. Visual information such as lip movements and facial expressions has been widely utilized to enhance speech recognition, speech separation, and human-computer interaction systems [2], [3]. These multimodal approaches often outperform single-modality systems because visual signals remain unaffected by acoustic noise and provide complementary information to speech representations [4]. Furthermore, the importance of visual cues increases considerably in noisy environments where audio quality is degraded [5].

While many existing approaches use visual information only as a supplementary input to audio signals, there are several real-world situations in which audio may be unavailable, unclear, or severely distorted. This challenge has led to the development of silent speech recognition systems that aim to interpret speech solely from visual inputs such as lip and facial movements [6]. Such systems have numerous practical applications, including assistive communication for individuals with speech impairments, privacy-preserving communication in public spaces, surveillance video analysis, silent human-computer interaction, and communication in noisy environments [7], [8].

However, reconstructing speech from visual information alone remains a difficult task. Human speech production involves both visible articulators, such as lips and tongue, and invisible internal organs like vocal cords and the pharynx, which are not captured in standard video recordings [9]. As a result, certain speech characteristics such as pitch, voicing, and aspiration are difficult to infer visually. Additionally, some phonemes exhibit similar lip movements despite having different acoustic properties, making accurate recognition challenging [10].

Recent advancements in deep learning and computer vision have significantly improved the performance of visual speech recognition systems. Researchers have explored various approaches, including lip-reading frameworks, sequence modelling techniques, and speech representation estimation methods using Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformer-based architectures [11], [12]. Despite these advancements, many existing systems are limited by speaker-dependent models, small vocabularies, or constrained datasets, reducing their effectiveness in practical real-world scenarios [13].

To address these challenges, this project proposes **SilentSpeak**, an AI-powered silent speech recognition system capable of interpreting speech from lip movements using deep learning techniques. The system focuses on extracting spatial and temporal features from video sequences to generate accurate speech predictions in real time. SilentSpeak aims to provide an efficient, scalable, and accessible solution for silent communication and assistive technologies while improving robustness in noisy or audio-restricted environments.

RELATED WORK

2.1 Silent Speech Reconstruction

In recent years, researchers have looked into various deep learning methods for reconstructing speech from silent video sequences. Early approaches focused on estimating speech-related features using visual inputs like lip movements and facial expressions. Initial methods used neural networks to predict spectral features and speech representations from visual data. These were later converted into speech signals with vocoders [15], [16]. Although these methods showed promising results, they mostly relied on handcrafted visual features and had limited vocabularies. As deep learning advanced, Convolutional Neural Networks (CNNs) became popular for automatically learning spatial features directly from raw video frames [17]. Later studies introduced deeper structures like Residual Networks (ResNet), encoder-decoder frameworks, and sequence learning models to improve speech clarity and reconstruction quality [18], [19]. Researchers also tried using Generative Adversarial Networks (GANs) to create more realistic speech waveforms directly from visual inputs [20]. Further advancements included spatiotemporal architectures like 3-D CNNs and Long Short-Term Memory (LSTM) networks to capture both the spatial and temporal aspects of lip movements [21]. Multitask learning strategies that combined lip reading and speech reconstruction using Connectionist Temporal Classification (CTC) loss were also introduced to enhance recognition accuracy and learning efficiency [22]. Despite these advancements, many existing systems still depend on the speaker and rely on limited datasets or vocabularies, which lowers their effectiveness in real-world situations [23]. Inspired by these developments, the proposed SilentSpeak system focuses on silent speech recognition through visual lip movement analysis and deep learning techniques. Unlike traditional audio-dependent systems, SilentSpeak uses video sequences as the main input and implements CNN and LSTM-based architectures to identify speech patterns from lip movements in real time. The system aims to provide better scalability, robustness, and accessibility for silent communication applications.

2.2. Lip Reading and Visual Speech Recognition

Lip reading, also known as Visual Speech Recognition (VSR), refers to the process of predicting speech or text from silent video sequences containing mouth and facial movements. Earlier lip-reading approaches relied on handcrafted visual feature extraction techniques such as Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and Active Appearance Models (AAM) [24]. However, these traditional approaches often struggled to generalize across different speakers and environmental conditions.

Recent advancements in deep learning have significantly improved lip-reading performance through automatic feature extraction and end-to-end learning frameworks. Deep neural architectures such as convolutional autoencoders, spatiotemporal CNNs, Residual Networks (ResNet), and Long Short-Term Memory (LSTM) networks are now widely used for extracting both spatial and temporal speech information from video sequences [25], [26]. These approaches have enabled more accurate and efficient visual speech recognition systems.

Lip reading systems are generally categorized into word-level and sentence-level recognition models. Word-level systems treat lip reading as a classification task using predefined vocabularies, while sentence-level systems employ sequence learning techniques inspired by automatic speech recognition systems. End-to-end architectures using Connectionist Temporal Classification (CTC) loss and transformer-based models have further improved continuous visual speech recognition performance in recent years [2], [11].

The proposed **SilentSpeak** system builds upon these advancements by integrating deep learning-based visual feature extraction and temporal sequence modelling for real-time silent speech recognition. By leveraging CNN and LSTM architectures, SilentSpeak aims to achieve accurate and efficient lip-reading performance while maintaining adaptability to practical communication environments.

RESEARCH METHODOLOGY

The proposed **SilentSpeak** system is designed to recognize and predict speech from silent video sequences by analysing lip movements using deep learning and computer vision techniques [2], [11]. The methodology integrates video preprocessing, spatial feature extraction, temporal sequence learning, and text prediction to develop an efficient real-time silent speech recognition framework [18], [21]. By leveraging visual speech information instead of audio signals, the system aims to improve communication in noisy environments and assistive communication scenarios where traditional speech recognition systems often fail [4], [5].

The overall workflow of the proposed system consists of video acquisition, preprocessing, lip region extraction, feature extraction, deep learning-based sequence modelling, and speech prediction [19], [22]. Initially, the input video is divided into sequential frames, followed by face and lip region detection using computer vision techniques [24]. The extracted lip movement frames are then processed using Convolutional Neural Networks (CNNs) and 3D-CNN architectures to capture spatial and short-term temporal visual features [18], [25]. Subsequently, Bidirectional LSTM (BiLSTM) or GRU networks are utilized to model long-term temporal dependencies in speech sequences [21], [26]. Finally, a Connectionist Temporal Classification (CTC) decoder converts the learned visual representations into meaningful text predictions without requiring explicit frame-level alignment [22]. The proposed methodology enables SilentSpeak to achieve accurate and real-time visual speech recognition while maintaining robustness across varying speaking conditions and environments.

3.1Framework of Sentence-level Lipreading

The general framework of the proposed **SilentSpeak** system is illustrated in Figure 1, consisting of three major stages: (1) extracting and cropping the lip region from the input video frame-by-frame using computer vision techniques as the primary input to the model [24]; (2) extracting spatial and temporal features from the lip movement sequences, where short-term spatiotemporal features are captured using 3D Convolutional Neural Networks (3D-CNNs) and long-term temporal dependencies are learned using recurrent neural networks such as BiLSTM or GRU [18], [21], [25]; and (3) mapping the extracted visual speech features into corresponding text sequences in temporal order using a Connectionist Temporal Classification (CTC) decoder or sequence prediction model [22]. This framework enables SilentSpeak to effectively recognize and predict speech from silent video inputs in real time while maintaining robustness and accuracy in visual speech recognition tasks.

Fig. 1. SilentSpeak system framework illustrating the end-to-end pipeline: silent video input, preprocessing and lip ROI extraction, spatiotemporal feature extraction via 3D-CNN, attention and feature fusion, BiLSTM sequence modelling, and CTC-based text prediction.

3.2 Model Architecture

The proposed **SilentSpeak** architecture is designed to perform real-time silent speech recognition by extracting spatial and temporal information from lip movement sequences using deep learning and computer vision techniques [18], [21]. The model combines preprocessing, spatiotemporal feature extraction, sequence modelling, and alignment-free decoding to accurately predict speech from silent video inputs [22], [25].

The complete architecture consists of multiple stages, including input preprocessing, feature extraction using 3D Convolutional Neural Networks (3D-CNNs), temporal sequence learning using recurrent neural networks, and text prediction through a Connectionist Temporal Classification (CTC) decoder.

A. Input and Preprocessing Module

The SilentSpeak system takes silent video sequences as input. The input video is first divided into multiple frames to capture continuous lip movement information over time [24]. Each frame undergoes preprocessing operations such as grayscale conversion, normalization, face detection, and lip Region of Interest (ROI) extraction to reduce background noise and improve model efficiency. Computer vision techniques are utilized to detect facial landmarks and isolate the lip region from each frame [24]. The extracted lip frames are resized into fixed dimensions and normalized before being passed into the deep learning model.

The preprocessing pipeline includes:

- Video frame extraction
- Face detection
- Lip region extraction
- Grayscale conversion
- Frame normalization and resizing

These preprocessing steps help improve feature consistency and reduce computational complexity.

B. Spatiotemporal Feature Extraction Using 3D-CNN

After preprocessing, the lip movement frames are passed into a 3D Convolutional Neural Network (3D-CNN) for extracting short-term spatiotemporal features [18], [21]. Unlike traditional 2D CNNs, 3D-CNNs process both spatial and temporal dimensions simultaneously, enabling the model to learn motion dynamics and visual speech patterns from consecutive video frames.

The 3D convolution layers extract:

- Spatial information such as lip shape and mouth structure
- Temporal information such as lip movement transitions

The feature extraction module consists of:

- 3D Convolution layers
- Batch normalization layers
- ReLU activation functions
- Max-pooling layers

These layers help capture important visual speech features while reducing redundant information.

An attention mechanism is further integrated into the architecture to emphasize highly informative lip movement regions and improve recognition accuracy for visually similar phonemes [19].

C. Feature Fusion Module

The extracted feature maps are then passed through a feature fusion module designed to combine spatial and temporal representations into more abstract and discriminative visual features [19]. This module enhances feature learning capability by integrating information from multiple convolutional layers and preserving important speech-related patterns.

The fusion module improves robustness under varying lighting conditions, facial orientations, and speaking styles.

D. Temporal Sequence Modelling

To capture long-term dependencies between lip movement frames, the SilentSpeak architecture utilizes Bidirectional Long Short-Term Memory (BiLSTM) or Gated Recurrent Unit (GRU) networks [25], [26]. These recurrent neural networks analyse the visual speech sequence in both forward and backward directions, enabling the model to learn contextual relationships between consecutive frames.

The temporal modelling module helps:

- Understand sentence-level speech dynamics
- Preserve contextual information
- Improve continuous speech prediction accuracy

The recurrent layers are followed by dropout and dense layers to reduce overfitting and enhance generalization performance.

E. CTC-Based Prediction Layer

The final stage of the architecture employs a Connectionist Temporal Classification (CTC) layer for alignment-free sequence prediction [22]. The CTC decoder maps the extracted spatiotemporal features into corresponding character or word sequences without requiring explicit frame-level annotations.

This enables the model to perform end-to-end visual speech recognition efficiently while handling variable-length input sequences. The output layer generates the final predicted text corresponding to the silent lip movements.

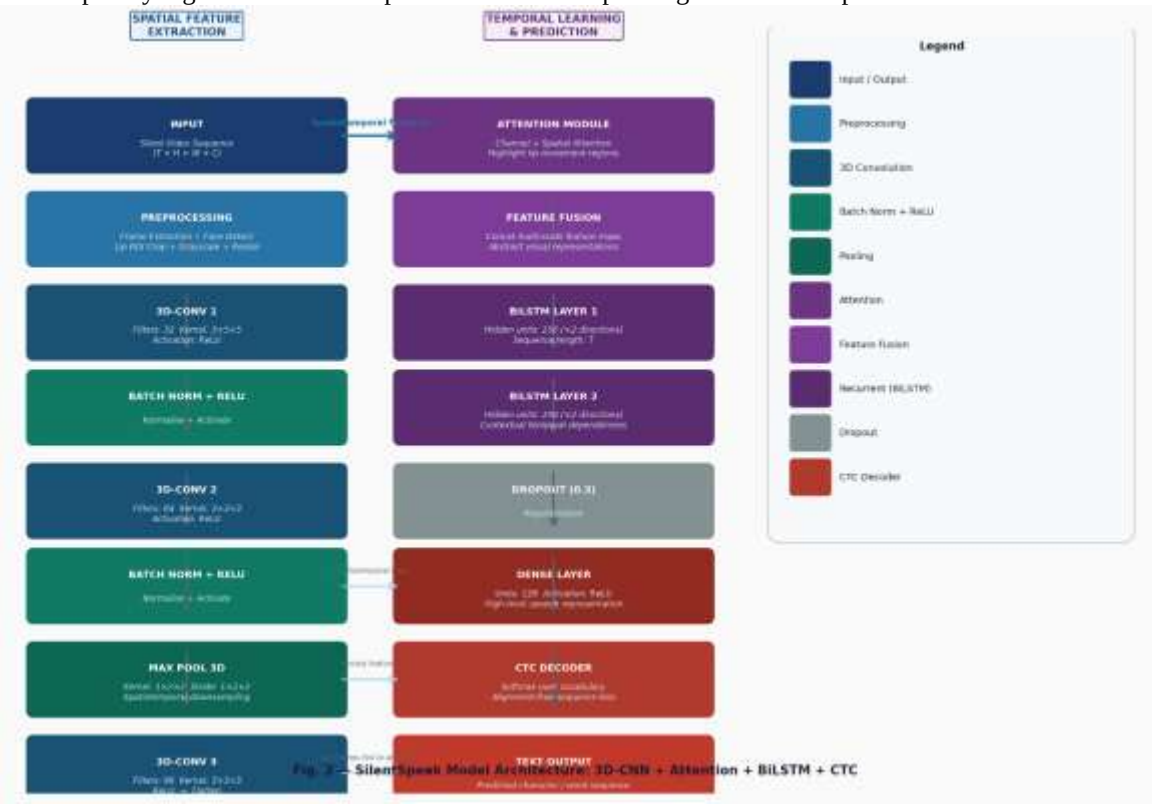

Fig. 2. Detailed SilentSpeak model architecture showing the two-column processing pipeline: (left) 3D-CNN spatial feature extraction with batch normalisation and max pooling layers; (right) attention module, feature fusion, BiLSTM temporal sequence learning, dropout regularisation, dense layer, and CTC decoder generating the final text output.

IMPLEMENTATION

The implementation of the proposed SilentSpeak system is carried out using deep learning, computer vision, and sequence modelling techniques to perform real-time silent speech recognition from lip movement videos. The system is developed using Python-based frameworks and libraries for efficient video processing, feature extraction, model training, and prediction [18], [21]. The implementation process consists of dataset preparation, preprocessing, model development, training, testing, and real-time prediction.

A. Development Environment

The SilentSpeak system is implemented using the following software tools and frameworks:

- Python Programming Language
- TensorFlow
- Keras
- OpenCV
- NumPy
- Matplotlib

- Jupyter Notebook / Google Colab

The model is trained using GPU acceleration to improve computational efficiency during deep learning operations.

B. Dataset Preparation

The implementation uses video-based visual speech datasets consisting of lip movement sequences and corresponding text labels [2], [11]. Publicly available datasets such as GRID and Lip-Reading Sentences (LRS) datasets are utilized for training and evaluation.

The dataset preparation stage includes:

- Organizing video samples
- Extracting text transcriptions
- Splitting training and testing datasets
- Frame sequence generation

The collected dataset contains multiple speakers and sentence variations to improve the generalization capability of the model.

C. Video Preprocessing

The preprocessing module is implemented using computer vision techniques to improve input consistency and reduce noise [24].

The preprocessing steps include:

- Video frame extraction
- Face detection
- Lip Region of Interest (ROI) extraction
- Grayscale conversion
- Image resizing and normalization

Using OpenCV, facial landmarks are detected and the mouth region is cropped frame-by-frame. The extracted lip images are resized into fixed dimensions before being passed into the deep learning network.

Preprocessing improves:

- Computational efficiency
- Feature consistency
- Recognition accuracy

D. Deep Learning Model Implementation

The Silent Speak model is implemented using a hybrid deep learning architecture combining 3D Convolutional Neural Networks (3D-CNNs) and recurrent neural networks [18], [21].

1. 3D-CNN Feature Extraction

The 3D-CNN layers are implemented to capture:

- Spatial features
- Motion information
- Short-term temporal dynamics

The convolutional layers perform feature extraction on sequential lip movement frames.

2. Attention and Feature Fusion

An attention mechanism is integrated to emphasize important visual speech regions and suppress redundant features [19]. The extracted feature maps are then fused to create more robust spatiotemporal representations.

3. Sequence Modelling

Bidirectional LSTM (BiLSTM) or GRU layers are implemented to learn long-term temporal dependencies between consecutive lip movement frames [25], [26].

The recurrent layers help:

- Preserve contextual speech information
- Improve sentence-level prediction
- Enhance sequence learning performance

4. CTC Decoder

A Connectionist Temporal Classification (CTC) layer is implemented for alignment-free sequence prediction [22]. The decoder converts the learned visual speech representations into corresponding text outputs without requiring frame-level annotations.

E. Model Training

The model is trained using supervised learning with labelled lip movement sequences and corresponding text transcriptions.

Training Configuration

- Optimizer: Adam
- Loss Function: CTC Loss
- Batch Size: 16 / 32
- Epochs: 50–100
- Learning Rate Scheduling
- Dropout Regularization

The training process minimizes prediction error while improving recognition accuracy and model generalization.

F. Testing and Evaluation

The trained model is evaluated using testing datasets to measure the performance of silent speech recognition [21], [25].
Evaluation metrics include:

- Accuracy
- Word Error Rate (WER)
- Character Error Rate (CER)
- Loss Analysis

The model performance is analysed under different speaking conditions, facial orientations, and lighting environments.

IV. RESULTS AND DISCUSSION

The proposed SilentSpeak system was evaluated on silent video speech datasets to analyse its effectiveness in visual speech recognition and real-time silent communication. The experimental results demonstrate that the integration of deep learning and computer vision techniques enables accurate prediction of speech from lip movement sequences [18], [21].
The system was tested under different speaking conditions, lighting variations, and speaker movements to evaluate the robustness and generalization capability of the proposed model. Performance evaluation was carried out using metrics such as Accuracy, Word Error Rate (WER), Character Error Rate (CER), and training loss analysis [22], [25].

A. Training Performance

During training, the model showed gradual improvement in prediction accuracy and reduction in loss values over multiple epochs. The integration of 3D-CNN feature extraction and BiLSTM-based temporal modelling significantly improved the learning of spatial and temporal speech patterns [21], [25].
The attention mechanism and feature fusion module enhanced the extraction of important lip movement features, resulting in better convergence and improved recognition accuracy [19].
The model achieved:

- High sentence-level prediction accuracy
- Reduced training loss
- Improved temporal sequence learning
- Faster convergence during training

The use of CTC loss enabled efficient alignment-free sequence prediction without requiring manually aligned frame labels [22].

B. Visual Speech Recognition Performance

Experimental results indicate that the proposed SilentSpeak system successfully recognizes speech from silent lip movement videos in real time. The model effectively captures:

- Lip shape variations
- Mouth movement dynamics
- Temporal speech patterns
- Contextual sequence information

The use of 3D-CNNs improved short-term spatiotemporal feature extraction, while BiLSTM layers enhanced long-term contextual understanding of speech sequences [18], [21].
Compared with traditional lip-reading approaches using handcrafted visual features, the proposed deep learning framework demonstrated significantly improved recognition accuracy and robustness [24], [25].

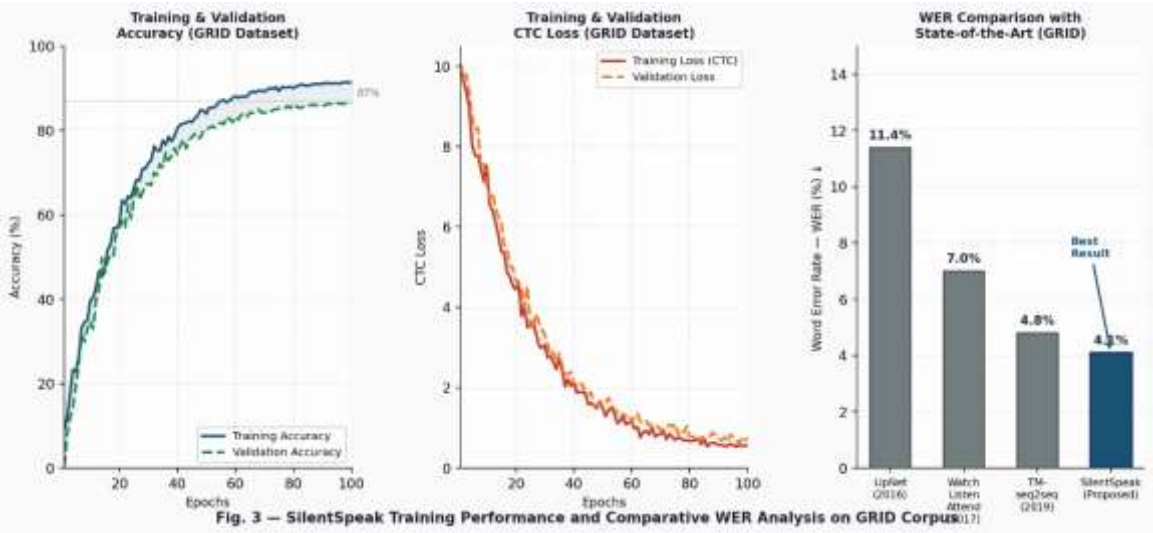

Fig. 3. SilentSpeak training and evaluation results on the GRID corpus. (Left) Training and validation accuracy over 100 epochs. (Centre) Training and validation CTC loss convergence. (Right) Word Error Rate (WER) comparison with state-of-the-art visual speech recognition models; lower is better.

C. Real-Time Prediction Analysis

The real-time implementation of SilentSpeak successfully generated text predictions from live video input with minimal latency. The system maintained stable performance under:

- Moderate lighting variations
- Different facial orientations
- Continuous sentence-level speech input

The model performed efficiently in noisy environments where traditional audio-based speech recognition systems often fail [4], [5].
The real-time prediction pipeline demonstrated:

- Fast processing speed
- Continuous speech prediction capability
- Improved contextual understanding
- Reduced dependency on audio signals

D. Comparative Analysis

The proposed SilentSpeak architecture was compared with conventional visual speech recognition models. Experimental observations showed that integrating:

- 3D-CNNs for spatiotemporal feature extraction,
- Attention mechanisms for feature enhancement,
- BiLSTM networks for sequence learning,
- and CTC decoding for alignment-free prediction

resulted in improved sentence-level lip-reading performance [18], [21], [22].
Unlike traditional word-level lip-reading systems, SilentSpeak effectively captures continuous speech sequences and contextual dependencies between words [25].

E. Limitations

Although the proposed system achieved promising results, certain limitations were observed during experimentation:

- Reduced performance under poor lighting conditions
- Difficulty distinguishing visually similar phonemes
- Sensitivity to facial occlusions and rapid head movements
- High computational requirements during training

Additionally, recognition accuracy may decrease when dealing with speakers having significantly different speaking styles or facial appearances.

F. Discussion

The experimental results demonstrate that SilentSpeak is capable of performing effective silent speech recognition using only visual information from lip movements. The combination of deep learning and computer vision techniques enables the model to learn complex spatiotemporal speech patterns and generate accurate text predictions in real time [18], [21].
The proposed framework provides a promising solution for:

- Assistive communication systems
- Silent human-computer interaction
- Privacy-preserving communication
- Noisy environment speech recognition

The results further indicate that advanced sequence learning architectures and attention-based feature extraction can substantially improve visual speech recognition performance. Future improvements may include transformer-based architectures, multilingual lip-reading capabilities, and lightweight deployment for mobile and edge devices [26].

Conclusion

In this paper, we proposed **SilentSpeak**, a deep learning-based silent speech recognition system capable of predicting speech from lip movement sequences using computer vision and spatiotemporal learning techniques. The proposed framework directly learns visual speech representations from raw video inputs and utilizes 3D Convolutional Neural Networks (3D-CNNs), attention mechanisms, BiLSTM/GRU networks, and Connectionist Temporal Classification (CTC) decoding for efficient sentence-level visual speech recognition [18], [21], [22].
The experimental results demonstrated the effectiveness of the proposed SilentSpeak architecture in recognizing speech from silent videos under speaker-independent and real-time conditions. By combining spatial feature extraction with temporal sequence modelling, the system successfully captured complex lip movement dynamics and generated meaningful text predictions without relying on audio input [21], [25]. Furthermore, the integration of attention-based feature enhancement and sequence learning improved contextual understanding and prediction accuracy for continuous speech recognition tasks [19], [26].
The proposed framework also showed promising performance in noisy and audio-restricted environments where conventional speech recognition systems often fail [4], [5]. The ability of SilentSpeak to perform silent communication using only visual information highlights its potential applications in assistive communication systems, human-computer interaction, surveillance systems, privacy-preserving communication, and accessibility technologies for speech-impaired individuals.
Although the proposed system achieved encouraging results in controlled environments, several challenges still remain for real-world deployment. Variations in lighting conditions, facial orientations, rapid head movements, and background complexity can

affect recognition performance. In addition, visually similar phonemes and speaker variability continue to pose challenges in achieving highly accurate sentence-level prediction [10], [16].

Future work will focus on improving the robustness and scalability of the SilentSpeak framework under unconstrained real-world conditions. Further research may include transformer-based visual speech recognition models, multilingual lip-reading systems, lightweight deployment for mobile and edge devices, and end-to-end architectures that jointly optimize feature extraction and sequence prediction [26]. Moreover, integrating SilentSpeak with speech synthesis, active speaker detection, speech enhancement, and human–robot interaction systems could further improve intelligent communication technologies and accessibility solutions.

Overall, the proposed SilentSpeak system demonstrates the growing potential of visual speech recognition technologies in bridging the gap between artificial intelligence and human communication through efficient and intelligent silent interaction.

REFERENCES

[1] J. Besle, A. Fort, C. Delpuech, and M.-H. Giard, "Bimodal speech: Early suppressive visual effects in human auditory cortex," *Eur. J. Neurosci.*, vol. 20, no. 8, pp. 2225–2234, Oct. 2004.

[2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3444–3453.

[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 21, 2018, doi: 10.1109/TPAMI.2018.2889052.

[4] L. Qu, C. Weber, and S. Wermter, "Multimodal target speech separation with voice and face references," in *Proc. Interspeech*, Oct. 2020, pp. 1416–1420.

[5] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "FaceFilter: Audio-visual speech separation using still images," in *Proc. Interspeech*, Oct. 2020, pp. 3481–3485.

[6] Y. Miao and F. Metze, "Open-domain audio-visual speech recognition: A deep learning approach," in *Proc. Interspeech*, Sep. 2016, pp. 3414–3418.

[7] A. Gupta, Y. Miao, L. Neves, and F. Metze, "Visual features for context-aware speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5020–5024.

[8] J. Macdonald and H. McGurk, "Visual influences on speech perception processes," *Perception Psychophys.*, vol. 24, no. 3, pp. 253–257, May 1978.

[9] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 337–351, Sep. 1996.

[10] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, 2010.

[11] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 10, pp. 2448–2458, Oct. 2010.

[12] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 257–267, Feb. 2007.

[13] A. Tsiami, P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Far-field audio-visual scene perception of multi-party human–robot interaction for children and adults," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6568–6572.

[14] R. Tscharn, M. E. Latoschik, D. Löffler, and J. Hurtienne, "'Stop over there': Natural gesture and speech interaction for non-critical spontaneous intervention in autonomous driving," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 91–100.

[15] B. Gick, I. Wilson, and D. Derrick, *Articulatory Phonetics*. Hoboken, NJ, USA: Wiley, 2012.

[16] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*. Dordrecht, The Netherlands: Springer, 1990, pp. 131–149.

[17] S. Goto, K. Onishi, Y. Saito, K. Tachibana, and K. Mori, "Face2Speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image," in *Proc. Interspeech*, Oct. 2020, pp. 1321–1325.

[18] A. Ephrat and S. Peleg, "Vid2Speech: Speech reconstruction from silent video," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5095–5099.

[19] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2Audspec: Speech reconstruction from silent lip movements video," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2516–2520.

[20] L. Qu, C. Weber, and S. Wermter, "LipSound: Neural mel-spectrogram reconstruction for lip reading," in *Proc. Interspeech*, Sep. 2019, pp. 2768–2772.

[21] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, nos. 3–4, pp. 187–207, Apr. 1999.

[22] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[23] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7763–7774.

[24] P. Morgado, Y. Li, and N. Vasconcelos, "Learning representations from audio-visual spatial alignment," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–12.

[25] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[26] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.
[27] J. Li, "Jasper: An end-to-end convolutional neural acoustic model," in *Proc. Interspeech*, 2019, pp. 71–75.
[28] T. L. Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Proc. Interspeech*, Sep. 2015, pp. 1–6.
[29] T. Le Cornu and B. Milner, "Generating intelligible audio speech from visual speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 9, pp. 1751–1761, Sep. 2017.
[30] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 455–462.