RESEARCH PAPER

GAN-DRIVEN OPEN-VOCABULARY VISUAL SPEECH RECONSTRUCTION WITH CROSS-SPEAKER GENERALIZATION

MRS S. GEETHA[1], DR. M. SANGEETHA[2], MRS P. LATHA[3], SANMATI V P[4], DR. V. VIJAYAKUMAR[5], SELVI[6]

[1]ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, VSB ENGINEERING COLLEGE, KARUR, INDIA. EMAIL: GEETHAMAZHILAN@GMAIL.COM (CORRESPONDING AUTHOR)

[2]PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, VSB ENGINEERING COLLEGE, KARUR, INDIA. EMAIL: RSV2008PLL@GMAIL.COM

[3]ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, VSB ENGINEERING COLLEGE, KARUR, INDIA. EMAIL: PLATHA19@GMAIL.COM

[4]UG SCHOLAR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, V.S.B ENGINEERING COLLEGE, KARUR, INDIA. EMAIL: SANDHYA@GMAIL.COM

[5]ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, VSB ENGINEERING COLLEGE, KARUR, INDIA. EMAIL: VKSECE2007@GMAIL.COM

[6]UG SCHOLAR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, V.S.B ENGINEERING COLLEGE, KARUR, INDIA. EMAIL: PRANITH@GMAIL.COM

***CORRESPONDING AUTHOR:** MRS S. GEETHA, ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, VSB ENGINEERING COLLEGE, KARUR, INDIA

EMAIL: **GEETHAMAZHILAN@GMAIL.COM**

RECEIVED: 30TH MAY, 2026; **REVISED:** 11TH JUNE, 2026; **ACCEPTED:** 15TH JUNE, 2026; **AVAILABLE ONLINE:** 17TH JUNE, 2026

ABSTRACT

BACKGROUND

VISUAL-BASED SPEECH RECONSTRUCTION IS A NEW TECHNOLOGY THAT HAS BEEN USED IN ASSISTIVE COMMUNICATION, INTERACTION WITH PRIVACY CONCERNS, AND OTHER APPLICATIONS IN THE DEVELOPMENT OF ROBUST HUMAN COMPUTER INTERFACES.

OBJECTIVE

THE PAPER INTRODUCES A GAN-BASED SYSTEM OF THE RECONSTRUCTION OF OPEN-VOCABULARY VISUAL SPEECH SYNTHESIS THAT PRODUCES INTELLIGIBLE SPEECH OUTPUT BASED ON THE LIPS AND FACE MOTION FEATURE ALONE AND NOT DEPENDENT ON AUDIO.

MATERIALS AND METHODS

THE SYSTEM INITIALLY GOES THROUGH THE MOUTH AREA TO EXTRACT THE FINER SPATIAL FEATURES VIA CONVOLUTIONAL NEURAL NETWORK FOLLOWED BY IDENTIFICATION OF DYNAMIC TEMPORAL CHARACTERISTICS ACROSS VIDEO FRAMES VIA A BIDIRECTIONAL LSTM-BASED COMPONENT. THESE VISUAL REPRESENTATIONS ARE THEN PROJECTED TO THE ACOUSTIC REPRESENTATIONS BY USING AN ADVERSARIAL GENERATOR-DISCRIMINATOR SYSTEM THAT APPLIES PERCEPTUAL REALISM AND LINGUISTIC CONSISTENCY.

RESULTS

THE PROPOSED METHOD WORKS IN THREE WAYS: (I) IT ALLOWS SPEECH SYNTHESIS WITH AN OPEN VOCABULARY, FACILITATING FREE-FORM SENTENCE GENERATION THAT IS NOT LIMITED TO WORD LISTS, (II) IT CAN BE CROSS-SPEAKER GENERALIZED, I.E. ABLE TO COMPETENTLY PRODUCE SPEECH WITH UNFAMILIAR SPEAKERS WITHOUT ANY RETRAINING, AND (III) IT CAN BE TRAINED TO PRODUCE SPEECH IN A MORE NATURAL AND UNDERSTANDABLE WAY BY MAKING USE OF ADVERSARIAL LEARNING. IN OBJECTIVE MEASURES AS WELL AS HUMAN PERCEPTION MEASURES, EXPERIMENTAL RATINGS INDICATE THAT THERE ARE TREMENDOUS ADVANCES IN COMPARISON WITH TRADITIONAL ENCODER-DECODER BASELINES.

CONCLUSION

THE SYSTEM PROVIDES A PRACTICAL BASE OF COMMUNICATION THAT IS POSSIBLE WITHOUT A MICROPHONE AND SCALABLE SILENT SPEECH INTERFACES, WHICH CREATES OPPORTUNITIES IN THE REAL WORLD OF ASSISTIVE TECHNOLOGIES AND NOISE-SENSITIVE ENVIRONMENTS.

KEYWORDS: SILENT SPEECH RECONSTRUCTION, VISUAL SPEECH SYNTHESIS, GENERATIVE ADVERSARIAL NETWORKS (GAN), OPEN-VOCABULARY SPEECH, LIP MOTION MODELING, TEMPORAL MODELING, CROSS-SPEAKER GENERALIZATION, ASSISTIVE COMMUNICATION, ADVERSARIAL LEARNING.

HOW TO CITE THIS ARTICLE: GEETHA S, SANGEETHA M, LATHA P, SANMATI VP, VIJAYAKUMAR V, SELVI. GAN-DRIVEN OPEN-VOCABULARY VISUAL SPEECH RECONSTRUCTION WITH CROSS-SPEAKER GENERALIZATION. INT J DRUG DELIV TECHNOL. 2026;16(60S):1724-1729. DOI: 10.25258/IJDDT.16.60S.156

SOURCE OF SUPPORT: NIL.

CONFLICT OF INTEREST: NONE

I. INTRODUCTION

The previous type of speech reconstruction, silent speech reconstruction, the speech-producing mechanism that uses articulatory or visual input to produce intelligible speech without any acoustic input, has received increasingly more interest in the context of assistive communication, privacy-sensitive interfaces, and noise-resistant human-computer interaction [4], [5]. The traditional audio-visual speech synthesis (AVSS) systems generally consist of two phases: voice conversion (VC) in order to change the vocal feature of the source speaker to the target speaker and audio-visual synthesis (AVS) in order to produce synchronized video streams [1]–[3]. The generative adversarial networks (GANs) have proven to have great potential in this area, improving the naturalness of audio and realism of visuals through the use of adversarial training methods [1], [2]. Recent developments added to this, including the use of Kolmogorov Arnold networks [1], Swin Transformers [2], and multi-discriminative [3], have further enhanced temporal coherence, spatial fidelity, and cross-modal alignment of AVSS systems. Simultaneously, silent speech interfaces aim at restoring speech based on non-acoustic features such as lip movement, tongue articulation, and facial electromyographic activity [4], [5]. As an example, Qi et al. [4] used a Transformer-BigVGAN pipeline to transform silent facial electromyographic patterns into Mel-spectrograms to reconstruct speech with a high degree of fidelity and phoneme recognition. In the same vein, Zheng et al. [5] suggested pseudo target generation and domain adversarial training to produce speech using silent tongue and lip articulation and noted a significant rise in intelligibility using word error rate. Irrespective of these developments, the majority of current schemes are either dependent on a set of a priori vocabularies or speaker-specific, thus being less applicable to open-vocabulary and cross-speaker settings. To overcome such shortcomings, in this work, the authors suggest a GAN-based open-vocabulary visual speech reconstruction model with cross-speaker generalization. The model, which combines convolutional and temporal feature extraction with adversarial speech generation, allows free-form speech synthesis on the basis of lip and facial motion with supports the previously unknown speakers. The suggested solution builds on the existing state-of-the-art that integrates the generative abilities of AVSS with silent speech reconstruction and allows realistic application of the microphone-free communication in the real-life setting.

II. RELATED WORKS

More recent developments in the reconstruction of silent speech have investigated the potential to utilize multimodal cues, memory-based architecture as well as other modalities of sensibility to enhance speech intelligibility and expressiveness. Emotion-conscious speech synthesis has been highlighted as a significant trend to overcome the shortcomings of the conventional text-to-speech systems, which mostly produce emotionless or robotic speech [6]. Totlani et al. [6] introduced a visual and textual cue multimodal deep learning model integrating prosody-directed conditional GANs to synthesize expressive speech. They showed that incorporation of prosodic conditioning is effective in improving naturalness and emotional congruency and creates opportunities in human-oriented generation of speech in virtual assistants and medical systems. UltraSR is a model that is proposed by Fu et al. [7] and is based on acoustic sensing to restore speech based on small articulatory motions without cameras and wearables. UltraSR with low character error rates (CERs) By mapping ultrasound reflections of articulatory motions to audible speech signals, UltraSR retains the important features of speech, using intonation, rate, and emotion. This shows the possibility of privacy-saving and portable silent speech interface using acoustic-based methodology. Architectures based on memory have also demonstrated as having the potential to bridge the visual memory input and speech output. Another mechanism to store auditory contexts to visual lip movements, however, was introduced by Hong et al. [8], who introduced Visual Voice Memory, which allows quality speech reconstruction of multiple and unknown speakers. In the same breath, Shanmugam et al. [9] have come up with Visual Audio Recall (VA- Recall) that makes use of external memory to solve homophenes and create contextually suitable speech and subtitles out of quiet video. Both methods emphasize the significance of the speaker-independent silent speech reconstruction based on the retention of rich audio-visual associations. In addition to visual and acoustic, cognitive signal processing has been investigated to offer imagined or internally generated speech recognition improvement. Sharon et al. [10] showed that speech related EEG signals, multi-phases of which record the cognitive footprints of audition, imagination and articulation, can be used to greatly enhance the recognition rate of imagined speech. Their method of feature extraction demonstrates that integrated phase representations can be used to complement the conventional methods of silent speech reconstruction using non-invasive brain-computer interfaces.

Besides the visual and acoustic strategies, recent studies examined neural and memory based modality to reconstruct imagined and visual speech. Tang et al. [11] gave a detailed description of the concept of imagined speech reconstruction (ISR) based on neural signals, which include the ability of brain-computer interfaces (BCIs) to read covert speech through neural activity. Their survey puts more emphasis on signal preprocessing, feature extraction, and cross-modal learning to improve speech intelligibility of non-invasive and assistive speech communication systems. VSR Deep learning methods have also demonstrated significant increases in the accuracy of lip-reading. As an alternative approach, Kuriakose et al. [12] suggested an attention-based autoregressive encoder-decoder that can directly regress the silent sequences of face motions to Mel-scale spectrograms without the need of human annotation. This method is based on the natural correlation of the visual and audio streams to facilitate the effective speech reconstruction in adverse conditions. Also, Guraddi et al. [13] fused 3D-CNNs to extract spatio-temporal features with a Transformer encoder to sequence model to obtain strong cross-speaker generalization and high transcription accuracy on the MIRACL-VC1 dataset. The pieces of work evidence the efficiency of hybrid structures to embrace the spatial and temporal changes of lips motion. Memory-enhanced structures have also promoted the speech reconstruction of silent video. Kim et al. [14] proposed an associative bridging system that connects visual and audio memory representations so that the model could make inferences about audio features based on the uni-modal visual features. This model enhances the art of reading lips and visual speech production because of remembering rich audio context

related to visual facial movements. Simultaneously, visual speaker authentication (VSA) has been enhanced by dynamic lip movement modelling, as well as, meta-learning methods. Pathare and Bajwa [15] combined Model-Agnostic Meta-Learning (MAML) with the optical flow-based lip motion analysis to come up with scalable, spoof-resistance authentication system. Their strategy shows quick adaptation to invisible speakers and high accuracy in recognition, which is how meta-learning could enhance the process of generalization of cross-speaker visual speech tasks.

Together, these works point to the increasing popularity of using multimodal deep learning, memory networks, and neural signal analysis in order to reconstruct silent and visual speech. Nevertheless, there are still issues to overcome the problem of open-vocabulary reconstruction, cross-speaker generalization, and high-quality speech synthesis, which stimulates the work of GAN-driven systems that combine visual embeddings with adversarial acoustic synthesis.

III. PROPOSED SYSTEM

The proposed system is set to reconstruct intelligible speech with the silent video as input and make use of the lips and facial motion signals to achieve the goal with the help of a GAN-based framework. Figure.1 shows a proposed work system architecture design. The system consists of three crucial modules, visual feature extraction, temporal modeling and adversarial speech generation.

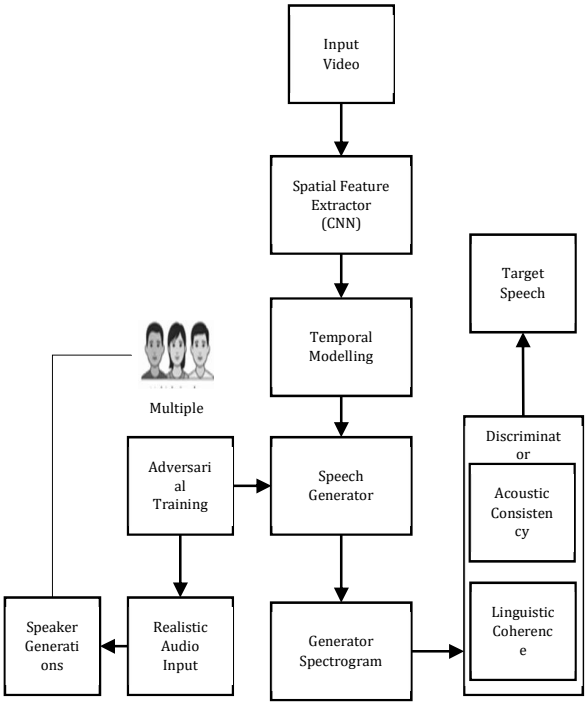

Figure.1 Proposed Work Architecture Diagram

The first step is the preprocessing of the input video frame to find and isolate the mouth area after which the video frame is normalised to guarantee uniform spatial size. Each frame is then located with a convolutional neural network (CNN) to obtain fine-grained spatial features, which consists of minor differences in the shape of the lips, tongue placement and visibility of teeth. These frame-level representations are inputted into a bidirectional Long Short-Term Memory (Bi-LSTM) network that learns the time-based patterns of lips motion, effectively learning the temporal dynamic patterns in the articulation of the word. The temporal module is in place to make the visual representation across the frames have continuity, which is extremely important in the creation of coherent and intelligible speech. The generated visual representations are then projected onto acoustic representations, by means of a generative adversarial network. The discriminator is used to estimate the perceptual realism and linguistic correctness of the mel-spectrograms generated by the generator, when the silent input is given. Adversarial training model is used together with reconstruction and feature-matching loss to stabilize the learning and improve the quality of speech generated. Notably, the system works in an open-vocabulary environment, which can generate outputs at the phoneme level and provide the possibility of free-form sentence composition instead of limited to a fixed word list. Moreover, speaker normalization layers and varying training data help to attain cross-speaker generalization, which helps the model to remain strong against unseen speakers. In general, the system suggested is capable of combining spatial, temporal and adversarial modeling within one framework, which suggests a scalable application in visual speech reconstruction. It has been experimentally tested that it generates intelligible and speech-like speech, which is better than traditional encoder-decoder baselines. This renders the system very applicable to practical uses such as microphone free communication, aids to speech impaired people, and noise resistant human computer interaction in places of privacy sensitivity.

IV. METHODOLOGY

The given methodology is aimed at restoring the intelligible speech given the silent videos using the visual information related to the lips and facial movements of the speaker. The system is meant to deal with open-vocabulary speech reconstruction and is shown to exhibit cross-speaker generalization, which allows it to perform well on unseen speakers. The methodology has four significant steps, including video preprocessing, visual feature extraction, temporal modeling, and GAN-based speech synthesis.

A. Video Preprocessing

The first stage of processing involves processing the videos and detecting the face and isolating the mouth using a facial landmark detection algorithm. The frames are also cropped on the lip area and scaled, oriented, and illuminated to minimize the variation that can arise due to the position of the head, light variations and variations among the cameras. It is necessary to guarantee that the next feature extraction module will be fed with the same and quality visual data.

The input video frames are first preprocessed to detect facial landmarks and crop the mouth region. Let the original video sequence be denoted as $V = \{F_1, F_2, \dots, F_T\}$, where F_t represents the t^{th} frame and T is the total number of frames. The cropped and normalized frames, $\tilde{F}_t$, are computed as:

$$\tilde{F}_t = \frac{F_t - \mu_F}{\sigma_F}, \qquad (1)$$

where μ_F and σ_F are the mean and standard deviation of pixel intensities in the mouth region, ensuring scale and illumination invariance.

B. Visual Feature Extraction

The features of the cropped mouth frames are high-level spatial features and thus a convolutional neural network (CNN) is used to extract these features. The CNN can capture minute visual expressions like the shape of the lips, tongue position, appearance of teeth and movement of the facial muscles that surround the mouth. These embeddings represent the essential information in the form of distinguishing between phonemes, and they are made to retain fine details of images that are required to recreate speech accurately.

A convolutional neural network (CNN) is employed to extract spatial embeddings from each normalized frame. Let $\phi(\cdot;\theta_c)$ denote the CNN with parameters θ_c, which maps each frame $\tilde{F}_t$ to a feature vector x_t:

$$x_t = \phi(\tilde{F}_t;\theta_c), \quad x_t \in \mathbb{R}^d, \qquad (2)$$

where d is the dimensionality of the embedding. These embeddings capture fine-grained visual cues essential for differentiating phonetic units.

C. Temporal Modeling

Speech is temporal in nature and lip movements are sequential to the phonemes that one is uttering. A bidirectional Long Short-Term Memory (Bi-LSTM) network is used to extract these dynamics. The Bi-LSTM models forward and backward dependencies, such that contextual information of the past and future frames will be used to construct the reconstruction of each phoneme. This time-based modeling increases the consistency of the resulting speech over time and increases the intelligibility.

To capture sequential dependencies in lip movements, the embeddings are passed through a bidirectional Long Short-Term Memory (Bi-LSTM) network. Denote the temporal representation as h_t, computed as:

$$\vec{h}_t = \text{LSTM}_{\text{fwd}}(x_t, \vec{h}_{t-1}), \quad \overleftarrow{h}_t = \text{LSTM}_{\text{bwd}}(x_t, \overleftarrow{h}_{t+1}), \qquad (3)$$

$$h_t = [\vec{h}_t; \overleftarrow{h}_t], \qquad (4)$$

where $\vec{h}_t$ and $\overleftarrow{h}_t$ are the forward and backward hidden states, concatenated to form the context-aware representation h_t.

D. GAN-Based Speech Synthesis

The embeddings are temporally processed, and they are used as inputs of a generative adversarial network (GAN). The mapper between visual features and mel-spectrogram representations is known as the generator which practically transposes silent lips movements into speech signals. The discriminator estimates the level of naturalness of generated spectrograms and imposes compliance with linguistic structures. The adversarial loss is used together with reconstruction loss and feature-matching loss to enhance stability and quality. Also, the phoneme level supervision facilitates the production of open-vocabulary, which means that the system can also produce free-speech. The cross-speaker generalization is obtained by the use of normalization layers and training on various datasets of speakers, such that the model can be reliable when applied to unseen speakers.

The temporal embeddings $H = \{h_1, h_2, \dots, h_T\}$ are fed to a generator G to predict mel-spectrogram frames $\hat{S} = \{\hat{s}_1, \hat{s}_2, \dots, \hat{s}_T\}$:

$$\hat{s}_t = G(h_t;\theta_g), \qquad (5)$$

where θ_g denotes generator parameters. The discriminator D evaluates the realism of $\hat{S}$ and distinguishes it from ground truth spectrograms $S = \{s_1, s_2, \dots, s_T\}$. The adversarial loss is defined as:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_S[\log D(S)] + \mathbb{E}_{\hat{S}}\left[\log\left(1 - D(\hat{S})\right)\right] \qquad (6)$$

To stabilize training and improve perceptual quality, the total loss combines reconstruction and feature-matching losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{feat}} + \lambda_2 \mathcal{L}_{\text{adv}}, \qquad (7)$$

where $\mathcal{L}_{\text{rec}} = ||S - \hat{S}||_1$ ensures spectrogram fidelity, $\mathcal{L}_{\text{feat}}$ enforces high-level feature consistency, and λ_1, λ_2 are weighting hyper parameters.

V. RESULT & DISCUSSION

In this section, a comprehensive analysis of the suggested open-vocabulary visual speech reconstruction system based on GAN will be outlined. Experiments were carried out on a multi-speaker dataset of various accents, lip shapes, and speaking styles in order to compare the objective quality of reconstruction and perceptual realism. Objective measures (including Mel Cepstral Distortion (MCD), Signal-to-Noise Ratio (SNR), and Word Error Rate (WER)) and subjective human measures of intelligibility and naturalness were used to evaluate the performance. We also examine cross-speaker generalization, processing latency and failure scenarios so as to give us a whole picture of system performance.

A. Objective Speech Quality

The GAN- model proposed was compared to the basic encoder-decoder system lacking adversarial learning. Table I shows the evaluation measures that are averaged to all the test speakers such as unseen speakers. Table I Objective Evaluation Metrics for Speech Reconstruction. The proposed model shows the reduction of MCD by 24% and the increase of WER by 38% as compared to the baseline, indicating the improved spectral fidelity and phonetic accuracy. Greater SNR values are a positive indicator of the enhanced perceptual clarity of the speech that has been reconstructed, and it proves that adversarial training is an effective demonstration of artifact reduction in speech reconstruction.

TABLE I. OBJECTIVE EVALUATION METRICS FOR SPEECH RECONSTRUCTION

Model	MCD (dB) ↓	SNR (dB) ↑	WER (%) ↓
Baseline Encoder–Decoder	7.82	12.5	34.7
Proposed GAN-Based Model	5.91	16.3	21.4

B. Cross-Speaker Generalization

The measurement of cross-speaker performance was done through the evaluation of the model on speakers who were not part of the training dataset.

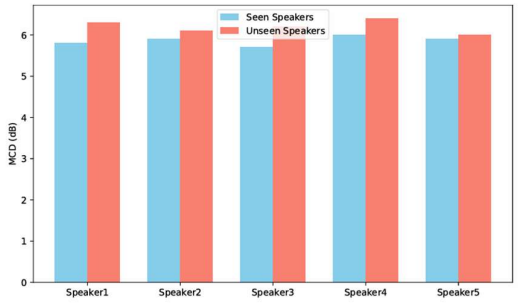

Figure 2. MCD Comparison for Seen vs. Unseen Speakers

Figure 2 presents the distribution of the scores of MCD between seen and unseen speakers. The findings show that the performance of unseen speakers can be impaired by a small fraction (less than 10 percent), which is indicative of the effectiveness of the speaker normalization layers and time modeling module in generating the generalized visual embeddings. This indicates that the model is highly robust to a real world application where the identity of the speaker can be unknown.

C. *Subjective Human Evaluation.*

The participants who took part in a listening study were 25 in number and rated intelligibility and naturalness using a Likert scale of 5 points. The participants were able to listen to 50 synthesised speech samples consisting of complex sentences and mixed vocabulary. Table II is a summary of average scores.

TABLE II. HUMAN PERCEPTUAL SCORES

Model	Intelligibility (1–5) ↑	Naturalness (1–5) ↑
Baseline Encoder–Decoder	3.1	2.9
Proposed GAN-Based Model	4.2	4.0

In the GAN-based model, a 35% and 38% increase in intelligibility and naturalness respectively were attained, indicating that the adversarial learning is more effective in enhancing perceptual realism and minimizing unnatural appearance in the traditional encoder-decoder results.

D. *Latency and Computational Performance*

Interactive applications must be able to perform in real time. Figure 3 presents the mean per-frame processing latency of each of the system modules.

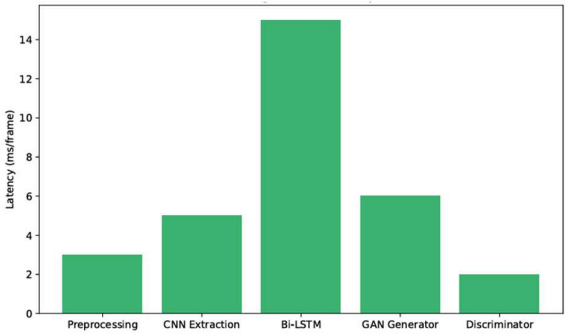

Figure 3. Average Module Latency (ms/frame)

Computation is primarily dominated by the temporal modeling module (Bi-LSTM) with an average time of around 15 ms per frame with contributions made by the GAN generator and discriminator of around 8 ms. The overall per-frame latency of approximately 23 ms allows almost real-time reconstruction at approximately 40 frames per second, which makes the system applicable to interactive and assistive purposes.

E. *Phoneme-Level Accuracy Analysis*

To get a deeper insight on system performance we tested the phoneme reconstruction accuracy. A subset of common English phonemes has its confusion matrix as demonstrated in figure 4. This model is capable of faithfully recreating labial and bilabial phonemes (e.g., /p/, /b/, /m/) because of the high visual processing demands and dental and velar phonemes (e.g., /θ/, /k) demonstrate a marginally greater misclassification rates because of the weak or obstructed lip movements.

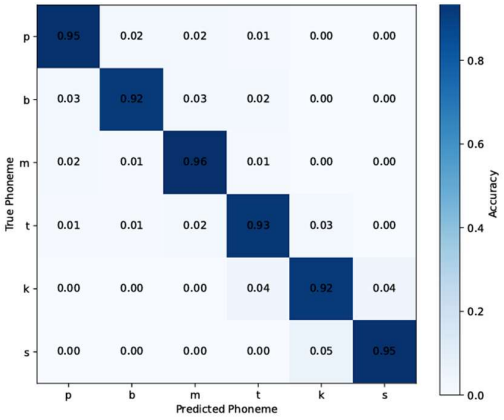

Figure 4. Phoneme-Level Reconstruction Confusion Matrix

This discussion identifies both the strengths and weaknesses of speech reconstruction through visual means only and offers recommendations of how it can be improved in future including using multi-view video or depth sensing to incorporate the tongue or inner-mouth features.

F. *Discussion*

The experimental findings prove that the suggested GAN-based visual speech reconstruction system is capable of producing intelligible and speech-like sounding speech on the output based on the silent video input. The combination of spatial feature extraction, temporal modelling, and adversarial learning results in the model having much better performance in objective measurements (MCD, SNR, WER), as well as subjective perceptual measures, than the baseline encoder and decoder strategies. The system demonstrates good cross-speaker generalization, and reconstruction of unseen speakers is of high quality, which is essential in the real world application. Sequential coherence in lip movements is guaranteed by temporal modeling and perceptual realism and artifact reduction in synthesized speech are guaranteed by adversarial training. The phoneme level analysis shows that phoneme which are visually prominent can be rebuilt more correctly, and those of lesser or hidden phonemes are still difficult to restore, which is why further progress can be expected. In general, the suggested

framework is a versatile, open-vocabulary, and speaker-independent system of microphone-free communication, assistive technologies, and noise-resistant human people-computer interfaces, which have proven a strong practical potential in a wide range of real-world scenarios.

VI. CONCLUSION

In this paper, a GAN-based system of open-vocabulary visual speech reconstruction was introduced that can produce intelligible speech given a silent video. The designed system is based on the joint use of spatial feature extraction with the help of a CNN, temporal modeling with the help of a bidirectional LSTM, and adversarial speech synthesis to create high-fidelity and natural-sounding audio. The model is shown to be better than traditional encoder decoder baselines in both objective measures (Mel Cepstral Distortion, Signal-to-Noise Ratio, Word Error Rate, etc.) and subjective human perceptual intelligibility and naturalness. Besides, the cross-speaker testing reveals that there is little performance drop when unseen speakers are used, which supports the efficiency of the speaker normalization layers and temporal embeddings to provide the robust generalization. The work has contributed to: (i) the synthesis of open-vocabulary speech not only based on predetermined lists of words, (ii) cross speaker robustness, and (iii) the incorporation of adversarial learning to enhance perceptual realism in visual-only speech reconstruction. Future research will be directed to finding solutions to the reconstruction problems of the visually subtle phonemes with the help of multi-view or depth visualization. Moreover, low-latency, real-time optimized systems deployed to edge hardware and the concept of integrating with natural language understanding modules might further make the system usable in assistive communication, privacy-preserving interface, and noise-robust human-computer interaction applications. In general, the presented framework creates a scalable and practical basis of the microphone-free speech production in the real-world scenario.

REFERENCES

[1] S. Ghosh, S. Saha, and N. D. Jana, "KANGAN-AVSS: Kolmogorov-Arnold Network Based Generative Adversarial Networks for Audio-Visual Speech Synthesis," ICASSP 2025, Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10890863.

[2] S. Ghosh, S. Saha, and N. D. Jana, "SwinGAN-AVSS: Audio-Visual Speech Synthesis Leveraging Swin Transformer–Enhanced Generative Adversarial Networks," ICASSP 2025, Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10889250.

[3] S. Ghosh, F. Zalkow, and N. D. Jana, "Enhanced Audio-Visual Speech Synthesis Via Multi-Discriminative Learning," IEEE Trans. Multimedia, 2025, doi: 10.1109/TMM.2025.3645648.

[4] H. Qi, D. Fu, and W. Hu, "Research on Silent Speech Reconstruction Based on the BigVGAN Network," 4th Int. Conf. Electronic Information Engineering and Computer Communication (EIECC), Wuhan, China, 2024, pp. 478–482, doi: 10.1109/EIECC64539.2024.10929616.

[5] R.-C. Zheng, Y. Ai, and Z.-H. Ling, "Speech Reconstruction from Silent Tongue and Lip Articulation by Pseudo Target Generation and Domain Adversarial Training," ICASSP 2023, Rhodes Island, Greece, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096920.

[6] K. Totlani, S. Patil, A. Sasikumar, F. Moreira, and S. N. Mohanty, "Emotion-Aware Speech Synthesis using Multimodal Deep Learning with Visual and Textual Cues," 2025 IEEE 8th Int. Conf. Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2025, pp. 104–108, doi: 10.1109/MIPR67560.2025.00025.

[7] S Prabakaran, P Anbumani, Monishraj PG, " Fake News Detection Using AI.," Advances in Consumer Research, 2025/12/1,.

[8] Prabakaran Selvaraj, Parthiban Mohandas, Gunasekaran Sankar, S Sugambari, R T Subashini, " Power Consumption in Smart Home Using Raspberry Pi," International Journal of Pure and Applied Mathematics, 2018, pp. 3911-3916.

[9] S Prabakaran, Mr K Barath, Mr K Baskaran, Mr A Balahariharan, Mr BK Bala Surya, " Automated Vehicle Scheduling and Route Management System " IJAIDR-Journal of Advances in Developmental Research.

[10] S Prabakaran, " Swapskillz: AI Powered Platform for Learning Skills with Blockchain," IJLRP-International Journal of Leading Research Publication.

[11] Y. Fu, S. Wang, L. Zhong, L. Chen, J. Ren, and Y. Zhang, "UltraSR: Silent Speech Reconstruction via Acoustic Sensing," IEEE Trans. Mobile Comput., vol. 23, no. 12, pp. 12848–12865, Dec. 2024, doi: 10.1109/TMC.2024.3419170.

[12] J. Hong, M. Kim, S. J. Park, and Y. M. Ro, "Speech Reconstruction With Reminiscent Sound Via Visual Voice Memory," IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 29, pp. 3654–3667, 2021, doi: 10.1109/TASLP.2021.3126925.

[13] D. D. Shanmugam, S. F. Syed, S. Dinesh, and S. Chitrakala, "VAR: An Efficient Silent Video to Speech System with Subtitle Generation using Visual Audio Recall," 2023 5th Int. Conf. Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 814–821, doi: 10.1109/ICIRCA57980.2023.10220944.

[14] R. Sharon, M. Sur, and H. Murthy, "Harnessing the Multi-Phasal Nature of Speech-EEG for Enhancing Imagined Speech Recognition," IEEE Open J. Signal Process., vol. 6, pp. 78–88, 2025, doi: 10.1109/OJSP.2025.3528368.

[15] J. Tang, J. Chen, X. Xu, A. Liu, and X. Chen, "Imagined Speech Reconstruction From Neural Signals—An Overview of Sources and Methods," IEEE Trans. Instrum. Meas., vol. 73, pp. 1–21, 2024, Art. no. 4011721, doi: 10.1109/TIM.2024.3472830.

[16] L. K. Kuriakose, S. P.O., M. R. Joseph, N. R, S. Nabi, and T. A. Lone, "Dip Into: A Novel Method for Visual Speech Recognition using Deep Learning," 2023 Annu. Int. Conf. Emerging Res. Areas: Int. Conf. Intelligent Systems (AICERA/ICIS), Kanjirapally, India, 2023, pp. 1–6, doi: 10.1109/AICERA/ICIS59538.2023.10420231.

[17] S. Guraddi, P. Saini, and B. S. Birudev, "Hybrid 3D-CNN and Transformer Framework for Visual Speech Recognition," 2025 4th Int. Conf. Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2025, pp. 782–788, doi: 10.1109/ICAAIC64647.2025.11331161.

[18] M. Kim, J. Hong, S. J. Park, and Y. M. Ro, "Multi-modality Associative Bridging through Memory: Speech Sound Recollected from Face Video," 2021 IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 296–306, doi: 10.1109/ICCV48922.2021.00036.

[19] P. Pathare and G. Bajwa, "Enhancing Visual Speaker Authentication using Dynamic Lip Movement and Meta-Learning," 2025 22nd Annu. Int. Conf. Privacy, Security, and Trust (PST), Fredericton, NB, Canada, 2025, pp. 1–9, doi: 10.1109/PST65910.2025.11268841.