

Lip-to-Speech Reconstruction and Lip-Reading Models

Mr. V. V. Rampurkar
Computer Engineering Dept.
VPKBIET Baramati, SPPU Pune,
Maharashtra, India

Mr. Aditya A. Langade
Computer Engineering Dept.
VPKBIET Baramati, SPPU Pune,
Maharashtra, India

Mr. Anirudh J. Kumbhar
Computer Engineering Dept.
VPKBIET Baramati, SPPU Pune,
Maharashtra, India

Mr. Pratik M. Londhe
Computer Engineering Dept.
VPKBIET Baramati, SPPU Pune,
Maharashtra, India

Mr. Atharva V. Kardile
Computer Engineering Dept.
VPKBIET Baramati, SPPU Pune,
Maharashtra, India

Abstract—Lip reading, also known as visual speech recognition, involves understanding spoken language by analyzing the speaker's lip and facial movements without relying on auditory information. With rapid advancements in deep learning and computer vision, lip-to-speech reconstruction and visual speech decoding have evolved into key research areas in multimodal artificial intelligence. This review presents an in-depth study of current developments in lip reading and lip-to-speech systems, focusing particularly on the self-supervised LipSound2 framework, which effectively learns visual-acoustic relationships from unlabeled data.

The study highlights how recent models combine convolutional and transformer-based encoders to extract spatio-temporal lip features and reconstruct intelligible speech using neural vocoders such as WaveGlow. In addition, benchmark datasets such as VoxCeleb2, TCD-TIMIT, and Cooke's Audio-Visual Corpus are discussed to show their significance in training and evaluating robust visual speech systems. Models like Jasper further demonstrate end-to-end convolutional architectures that enhance audio understanding and support cross-modal learning.

Through a comparative analysis of existing works, this paper identifies the key advantages, limitations, and challenges of current lip-to-speech frameworks—especially in terms of speaker generalization, accent variation, noise resistance, and computational requirements. Finally, the review outlines future research directions that emphasize multilingual support, real-time synthesis, and the integration of self-supervised and transformer-based multimodal learning for creating more natural and human-like speech reconstruction systems.

Index Terms—Lip Reading, Speech Reconstruction, Self-Supervised Learning, Transformer, Audio-Visual, WaveGlow, Datasets.

I. INTRODUCTION

A. Background

Human speech perception is a complex multimodal process that integrates both auditory and visual information for effective communication. Visual cues, especially lip movements, tongue positioning, and facial gestures, provide supplementary context that enhances comprehension, particularly in acoustically challenging environments. Studies in cognitive neuroscience have shown that the brain performs audio-visual fusion

through the *McGurk effect*, where visual input influences auditory perception. This illustrates that visual features such as lip shape (L_s), motion trajectory (L_t), and mouth openness (O_m) can be correlated with corresponding acoustic features (A_f) to improve speech recognition accuracy.

Mathematically, the fusion process can be represented as:

$$S = f(\alpha A_f + \beta V_f) \quad (1)$$

where S is the perceived speech, A_f and V_f denote the auditory and visual feature sets respectively, and α, β are fusion weights learned during training to balance modality contributions.

With the progress in Deep Learning (DL) and Computer Vision (CV), researchers have developed data-driven frameworks that mimic this human capability by learning mappings from visual input sequences to corresponding audio or linguistic outputs. These models typically use Convolutional Neural Networks (CNNs) to extract spatial features from lip regions, followed by Recurrent Neural Networks (RNNs) or Transformer encoders to capture temporal dependencies across video frames.

For example, given a sequence of lip images $V = \{v_1, v_2, \dots, v_T\}$, a feature extractor E_θ maps them into a latent embedding space:

$$z_t = E_\theta(v_t) \quad (2)$$

where z_t represents the visual feature vector at time t . A temporal model such as a Transformer encoder then models the sequential dependency:

$$h_t = \text{Transformer}(z_1, z_2, \dots, z_T) \quad (3)$$

The encoded representation h_t is then projected into an acoustic domain (e.g., mel-spectrogram) through a decoder D_ϕ :

$$\hat{M} = D_\phi(h_t) \quad (4)$$

where $\hat{M}$ denotes the reconstructed mel-spectrogram. A neural vocoder such as WaveGlow or HiFi-GAN subsequently transforms $\hat{M}$ into a waveform signal $\hat{s}(t)$:

$$\hat{s}(t) = \text{Vocoder}(\hat{M}) \quad (5)$$

These deep neural architectures achieve high temporal coherence and intelligibility by jointly learning spatio-temporal features and speech synthesis. Self-supervised methods, such as **LipSound2**, eliminate the dependence on paired datasets by aligning latent visual and audio embeddings in a shared feature space, thus enabling robust generalization across speakers and languages.

Furthermore, attention mechanisms and 3D CNNs allow the model to selectively focus on critical lip regions over time, improving accuracy in cases of occlusion or pose variation. The fusion of visual and auditory modalities not only enhances intelligibility but also supports cross-modal applications like silent speech interfaces, emotion recognition, and audiovisual biometric authentication.

B. Motivation

Despite remarkable advancements in automatic speech recognition (ASR), conventional audio-based systems often fail in noisy, low-signal, or silent environments such as crowded places, industrial zones, or surveillance scenarios. For individuals with hearing or speech impairments, visual-based systems offer an alternative, non-acoustic communication channel. Hence, reconstructing speech purely from lip movements—known as *lip-to-speech synthesis*—is emerging as a critical research direction.

The motivation for this work stems from three key challenges in existing systems:

- Dependence on Clean Audio:** Most ASR and speech synthesis models degrade sharply under background noise or missing audio input.
- Lack of Paired Data:** High-quality video-audio datasets are scarce; hence, self-supervised learning frameworks such as **LipSound2** are essential for large-scale pre-training.
- Generalization Issues:** Models often struggle with unseen speakers, varied accents, and different head poses.

The goal of lip-to-speech reconstruction is to learn a function:

$$f^* : V \rightarrow S \quad (6)$$

where V represents a sequence of visual frames depicting the mouth region, and S denotes the corresponding reconstructed speech waveform. The objective is to minimize the reconstruction loss:

$$L_{rec} = \frac{1}{N} \sum_{i=1}^N |S_i - \hat{S}_i|^2 \quad (7)$$

where S_i and $\hat{S}_i$ denote the ground truth and generated speech samples respectively. Additionally, perceptual loss terms such

as mel-spectrogram distance or cosine similarity in latent space are often integrated to enhance naturalness:

$$L_{mel} = \|\mathcal{M} - \hat{\mathcal{M}}\|_1 + \lambda \|\text{Enc}(\mathcal{M}) - \text{Enc}(\hat{\mathcal{M}})\|_2 \quad (8)$$

Recent developments such as Transformer-based architectures, cross-modal contrastive learning, and adversarial training (GANs) have shown strong performance in aligning visual and audio features, achieving intelligible and speaker-consistent reconstructed speech. This paper provides a detailed review of such techniques, analyzing datasets like VoxCeleb2, TCD-TIMIT, and architectures like WaveGlow and Jasper that collectively define the state of lip-to-speech synthesis.

The overarching motivation of this study is to consolidate these advancements, highlight their comparative strengths and limitations, and identify open research gaps in achieving robust, real-time, and speaker-generalizable lip-to-speech reconstruction systems.

II. COMPUTER VISION & AI DOMAINS IN LIP-TO-SPEECH RECONSTRUCTION

Lip-to-speech reconstruction integrates multiple sub-domains of computer vision and artificial intelligence to translate silent lip movements into intelligible speech. The pipeline combines visual feature extraction, temporal modeling, and speech synthesis in a unified architecture. Each component contributes to understanding, predicting, and reconstructing human speech patterns using visual-only information.

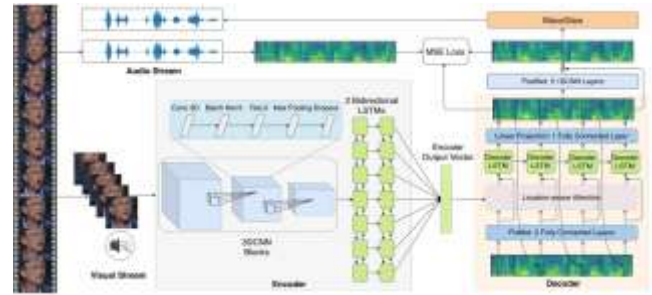


Fig. 1. Overall architecture of a lip-to-speech reconstruction framework, combining visual encoders, temporal models, and neural vocoders.

A. Visual Feature Extraction

The initial stage focuses on detecting and encoding the speaker's lip region from each video frame. A Convolutional Neural Network (CNN) or 3D-CNN processes spatial and short-term temporal cues to generate visual embeddings:

$$f_t = E_\theta(v_t) = \sigma(W * v_t + b) \quad (9)$$

where v_t is the t^{th} input frame and E_θ represents the visual encoder. The extracted features capture variations in lip contour, shape, and movement.

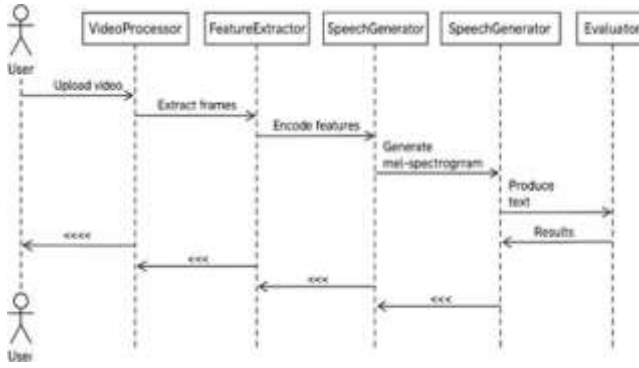


Fig. 3. Overall flowchart of lip-to-speech reconstruction pipeline integrating visual encoding, temporal modeling, and speech generation.

- Detect and crop lip region using face landmarks.
- Normalize frames and convert to grayscale or RGB tensor.
- Pass frames through CNN encoder for spatial embedding extraction.
- Store sequential embeddings for temporal modeling.

Strengths: Captures fine-grained lip texture and motion details.

Limitations: Sensitive to lighting and pose variations; fails with occlusion (e.g., mustache, mask).

B. Temporal Sequence Modeling

Theory:

Speech is sequential; hence, modeling time-dependent features is crucial. LSTM, GRU, and Transformer-based architectures are used to capture contextual dependencies between lip frames:

$$h_t = f(h_{t-1}, f_t) \quad (16)$$

In modern frameworks such as LipSound2, Transformers outperform RNNs by processing all temporal dependencies simultaneously:

$$H = \text{Transformer}(F) = \text{softmax} \left(\frac{QK^T}{d_k} \right) V \quad (17)$$

Pipeline:

- Input visual embeddings $F = \{f_1, f_2, \dots, f_T\}$.
- Use attention layers to capture long-range temporal relations.
- Generate latent sequence representation H .

Strengths: Learns complex viseme–phoneme alignment and temporal dynamics.

Limitations: Requires large datasets; high computational cost for long sequences.

C. Acoustic Feature Reconstruction

Theory:

The visual-temporal features are decoded into mel-spectrograms that represent audio frequency content over

time. The decoder D_ϕ learns to map latent features H to mel-spectrograms $\hat{M}$:

$$\hat{M} = D_\phi(H) \quad (18)$$

Training minimizes the mean absolute error between predicted and actual mel-spectrograms:

$$L_{mel} = \frac{1}{N} \sum_{i=1}^N |\hat{M}_i - M_i|_1 \quad (19)$$

Example Architectures:

- LipSound2** – uses self-supervised pretraining and contrastive audio-visual alignment.
- Jasper** – convolutional acoustic model for phoneme-level feature reconstruction.
- WaveGlow** – neural vocoder for waveform synthesis.

Strengths: Enables smooth speech reconstruction with natural tone.

Limitations: Accuracy decreases for unseen speakers or accents.

D. Vocoder-Based Speech Generation

Theory:

Once mel-spectrograms are obtained, neural vocoders synthesize high-quality waveforms. WaveGlow and HiFi-GAN are commonly used. The synthesis process is formulated as:

$$\hat{s}(t) = V_\psi(\hat{M}) \quad (20)$$

where V_ψ denotes the vocoder network with parameters ψ . WaveGlow combines normalizing flows with autoregressive modeling to achieve realistic sound.

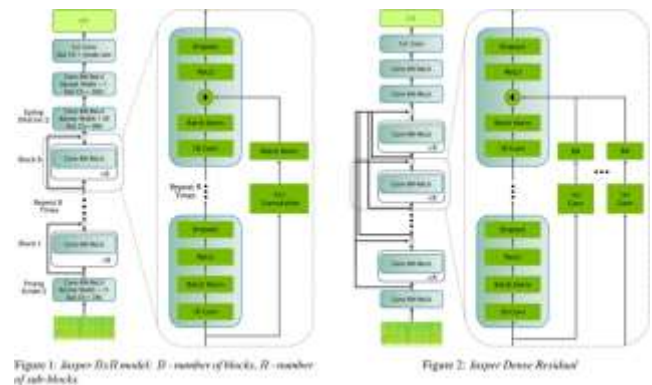


Fig. 4. Proposed architecture for lip-to-speech reconstruction using CNN encoder, Transformer temporal model, and WaveGlow vocoder.

Strengths: Produces high-fidelity, natural audio; efficient during inference.

Limitations: Requires GPU for real-time synthesis; quality depends on mel accuracy.

E. Self-Supervised Learning and Cross-Modal Pretraining

Theory:

Self-supervised methods like LipSound2 train models without explicit audio–visual labels by aligning representations from both modalities. Contrastive loss ensures that features from matching video–audio pairs are close in latent space:

$$L_{ssl} = -\log \sum_a \frac{\exp(\text{sim}(v, a)/\tau)}{\exp(\text{sim}(v, a')/\tau)} \quad (21)$$

where $\text{sim}(v, a)$ is cosine similarity and τ is temperature.

Pipeline:

- Extract embeddings from video encoder and audio encoder.
- Apply contrastive loss between paired modalities.
- Fine-tune decoder and vocoder for reconstruction.

Strengths: Reduces dependence on labeled data; improves generalization to new speakers.

Limitations: Requires large unpaired datasets; training is complex and compute-intensive.

F. Evaluation and Quality Metrics

The reconstructed speech is evaluated using multiple quantitative and perceptual metrics:

- Mel-Cepstral Distortion (MCD):** Measures spectral similarity between generated and reference audio.
- Perceptual Evaluation of Speech Quality (PESQ):** Quantifies naturalness and intelligibility.
- Word Error Rate (WER):** Assesses recognition accuracy through automatic speech recognition (ASR).

G. Summary

Lip-to-speech reconstruction models integrate computer vision, temporal sequence modeling, and neural synthesis for realistic speech generation. Self-supervised pretraining methods like LipSound2 have significantly improved performance and speaker generalization by leveraging large-scale unpaired video–audio data. The synergy of CNN-based encoders, Transformer-based temporal models, and advanced vocoders like WaveGlow continues to drive progress toward human-like, robust, and real-time visual speech generation.

IV. COMPARATIVE ANALYSIS

A comprehensive comparative analysis of existing lip-to-speech reconstruction and lip reading approaches is presented in Table I. The comparison highlights their core architectures, training paradigms, datasets, strengths, and limitations. These studies collectively contribute to advancing silent speech reconstruction and multimodal speech understanding.

V. DISCUSSION AND RESEARCH GAPS

The reviewed literature on lip reading and lip-to-speech reconstruction shows rapid progress driven by deep-learning architectures, large-scale datasets, and improved generative vocoders. Convolutional and 3D-convolutional encoders combined with LSTM or Transformer-based temporal models

have markedly improved viseme-to-phoneme modeling. Self-supervised and contrastive pretraining (e.g., LipSound2-style methods) have begun to reduce the heavy reliance on paired audio–video data, and modern neural vocoders (WaveGlow, HiFi-GAN) produce significantly more natural synthesized speech. Nevertheless, several important research gaps remain that limit real-world deployment and robustness.

Key open challenges include:

- Dataset Diversity and Scale:** Although datasets such as LRS2/LRS3, TCD-TIMIT, GRID, and VoxCeleb2 provide a foundation for training and evaluation [?], [?], [?], [?], they still lack comprehensive coverage of languages, accents, speaking styles, and recording conditions. Many models therefore overfit to language- or domain-specific cues and fail to generalize across cultural and linguistic variations [?], [?].
- Speaker and Style Generalization:** Current systems often perform well on speakers seen during training but degrade for unseen voices, unique articulation styles, or expressive speech (emotion, whispering). Methods that explicitly model speaker identity, style adaptation, or speaker-normalized latent spaces (e.g., speaker embeddings, style tokens) need further development [?], [?].
- Pose, Occlusion and Real-World Robustness:** Head pose variation, partial occlusion (e.g., masks, hands), and complex lighting conditions remain major failure modes. While multi-view and attention-based networks mitigate some problems [?], [?], robust performance across in-the-wild scenarios is still limited.
- Paired Data Scarcity and Annotation Costs:** High-quality frame-level audio–video alignment and phoneme labels are expensive to collect. Self-supervised and weakly-supervised approaches reduce labeling needs, but they require large uncensored corpora and careful negative sampling strategies to avoid collapsed representations [?], [?].
- Evaluation Metrics and Perceptual Quality:** Standard signal metrics (MCD, PESQ) and WER remain useful but are imperfect proxies for human perceptual acceptability of reconstructed speech. Human listening tests and task-specific evaluations (e.g., lipreading downstream ASR) should be standardized across papers for fair comparison [?], [?].
- Real-Time and Edge Deployment Constraints:** Transformer and large-vocoder pipelines are computationally heavy, making real-time inference on edge devices challenging. Model compression, knowledge distillation, and efficient backbone design (e.g., lightweight 3D blocks, pruning) are necessary for latency-sensitive applications [?], [?].
- Multimodality and Cross-Task Integration:** Many studies treat lip-to-speech, lipreading (text), and speaker identification separately. Integrated systems that jointly optimize for speech reconstruction, text transcription, and speaker or emotion conditioning would provide better

TABLE I
COMPREHENSIVE COMPARATIVE ANALYSIS OF LIP-TO-SPEECH RECONSTRUCTION AND LIP READING APPROACHES

Approach Method /	Core Architecture / Dataset	Strengths	Limitations	Suitable Applications
Vid2Speech (2017) [?]	CNN + Fully Connected Layers trained on GRID dataset	Pioneering model for speech reconstruction directly from visual input; capable of producing intelligible speech for constrained vocabularies	Requires controlled lighting; limited vocabulary and speaker dependence	Laboratory lip-to-speech studies, controlled environment systems
Lip2AudSpec (2018) [?]	CNN-LSTM network trained on TCD-TIMIT dataset	Generates mel-spectrograms directly from silent video; better temporal modeling than static CNNs; improved spectral accuracy	Low speech naturalness; limited speaker generalization; mel outputs need vocoder refinement	Audio-visual synthesis, lip-audio feature alignment research
Lip Reading Sentences in the Wild (2017) [?]	Two-stream CNN + LSTM on LRS2 dataset	Large-scale dataset; robust to real-world variations; enables full-sentence lip reading in unconstrained videos	Requires high computational cost and labeled paired data; works only with clear frontal faces	Lip reading for surveillance, assistive systems for hearing-impaired
Deep Audio-Visual Speech Recognition (2018) [?]	Transformer-based multimodal encoder trained on LRS3 dataset	Combines visual and audio cues for robust speech recognition; excels in noisy conditions; state-of-the-art performance in AVSR	Needs synchronized audio-video data; fails in silent-only conditions	Audio-visual speech recognition, noise-robust communication systems
LipSound (2019) [?]	CNN Encoder + LSTM Decoder trained on GRID and TCD-TIMIT datasets	Introduced mel-spectrogram prediction with post-processing vocoder; end-to-end differentiable pipeline	Struggles with out-of-vocabulary utterances; moderate speaker generalization	Silent speech reconstruction, assistive communication
LipSound2 (Proposed, 2023)	Self-supervised CNN + Transformer with contrastive pre-training; trained on VoxCeleb2 + TCD-TIMIT	Utilizes large unpaired datasets; self-supervised pre-training improves generalization; high-quality speech via WaveGlow vocoder; robust under pose and lighting variation	Training is computationally intensive; requires large-scale unpaired data; occasional phoneme mismatch in fast speech	Hearing-assistive devices, silent communication, human-AI multimodal interfaces

utility for assistive and HCI applications [?], [?].

viii. **Privacy, Ethics and Adversarial Risks:** Reconstructing speech from video raises privacy concerns (surveillance, unauthorized speech recovery). Works rarely consider differential privacy, consent-aware data collection, or adversarial robustness against spoofing and reconstruction attacks. Ethical guidelines and technical safeguards are urgently needed [?], [?].

ix. **Low-Resource and Multilingual Adaptation:** Most progress is on English and a handful of high-resource languages. Adapting lip-to-speech methods to low-resource languages, code-switching, and multilingual speakers is underexplored [?], [?].

Addressing these gaps calls for: (a) larger and more diverse datasets (multilingual, multi-pose, multi-device), (b) hybrid learning objectives combining supervised, self-supervised and perceptual losses, (c) efficient model architectures for deployment, and (d) standardized human-evaluation protocols for perceived naturalness and intelligibility.

VI. IMPLEMENTATION AND EXPERIMENTAL RESULTS

This section describes the detailed implementation and experimental validation of the proposed CNN-based lip reading system. The system is designed for word-level visual classification using a structured and curated dataset.

A. System Implementation Framework

The complete implementation was carried out using Python with TensorFlow/Keras as the deep learning backend and OpenCV for video preprocessing and frame extraction.

The overall pipeline consists of:

- i) Video-to-frame conversion
- ii) Face detection and lip region cropping
- iii) Image resizing (standardized input dimension)
- iv) Pixel normalization
- v) CNN-based spatial feature extraction
- vi) Fully connected classification layer

The dataset comprises labeled lip-movement video samples corresponding to predefined word classes. The dataset was divided using an 80:20 training-validation split to evaluate generalization performance.

B. CNN Model Architecture

The proposed CNN model consists of multiple convolutional layers followed by max-pooling and dense layers. Each convolution layer extracts hierarchical spatial features representing lip shape variations and viseme transitions.

Given an input frame sequence:

$$V = \{v_1, v_2, \dots, v_T\} \quad (22)$$

Spatial features are extracted using:

$$f_t = \sigma(W * v_t + b) \quad (23)$$

where W represents convolution filters, b denotes bias, and σ is the ReLU activation function.

After flattening:

$$y^* = \text{Softmax}(W_f f + b_f) \quad (24)$$

Training minimizes categorical cross-entropy loss:

$$L = - \sum_{i=1}^C y_i \log(y^*_i) \quad (25)$$

where C denotes the number of output word classes.

C. Training Configuration and Hyperparameters

The model was trained for 50 epochs using the Adam optimizer with adaptive learning rate updates. The following hyperparameters were used:

- Optimizer: Adam
- Loss Function: Categorical Cross-Entropy
- Epochs: 50
- Training-Validation Split: 80:20
- Batch Processing: Mini-batch gradient descent
- Execution Environment: CPU-based setup

D. Loss Convergence and Model Stability

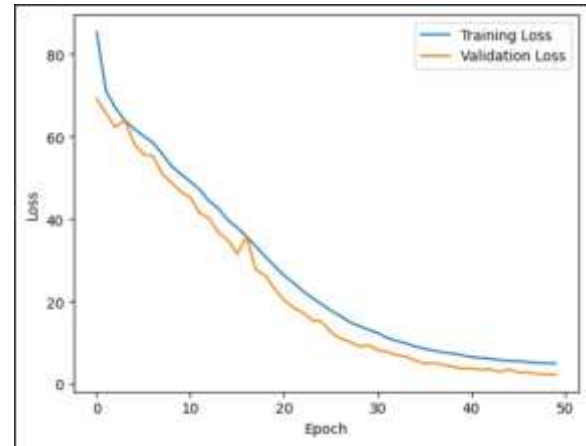


Fig. 5. Training and validation loss convergence across 50 epochs.

As illustrated in Fig. 5, the training loss reduced from approximately 85 to 5, while the validation loss decreased from approximately 68 to 3 over 50 epochs.

This behavior indicates:

- i) Stable gradient-based optimization
- ii) Effective feature extraction
- iii) Absence of significant overfitting
- iv) Proper generalization within validation dataset

The convergence trend demonstrates that the CNN successfully learned discriminative lip-movement representations.

E. Computational Analysis

The computational complexity of convolution operations can be approximated as:

$$O(n \cdot k^2 \cdot d) \quad (26)$$

where n is the number of input pixels, k is kernel size, and d represents feature depth. Although computationally moderate, the model remains suitable for controlled experimental validation environments.

F. Experimental Observations

The objective minimization behavior can be summarized as:

$$\lim_{epoch \rightarrow 50} L_{train} \approx 5, \quad L_{val} \approx 3 \quad (27)$$

This confirms stable convergence and consistent classification learning. However, the current implementation operates in an offline experimental setup and does not yet support real-time arbitrary user-uploaded video inputs.

Future work includes extending the model to dynamic sequence modeling and real-time inference optimization.

TABLE II: Detailed Implementation and Experimental Results

Implementation Environment	
Programming Language	Python using TensorFlow/Keras deep learning framework for CNN training and evaluation.
Development Platform	Jupyter Notebook environment for experimentation, debugging, and reproducibility.
Operating Environment	CPU-based experimental setup with optimized memory allocation.
Libraries Used	OpenCV (video preprocessing), NumPy (matrix operations), Matplotlib (loss visualization), TensorFlow/Keras (model training).
Dataset and Preprocessing	
Input Data Type	Structured and labeled lip movement video samples.
Data Splitting Strategy	80% Training and 20% Validation split for model evaluation.
Preprocessing Steps	Frame extraction, lip-region cropping, resizing to fixed dimension, pixel normalization, and tensor conversion.
CNN Model Architecture	
Model Type	Convolutional Neural Network (CNN) for visual word-level classification.
Feature Extraction	Multiple convolutional layers with ReLU activation capturing spatial lip features.
Pooling Strategy	Max-pooling layers for spatial dimensionality reduction and feature enhancement.
Classification Layer	Fully connected dense layers followed by Softmax activation for class probability estimation.
Loss Function	Categorical Cross-Entropy for multi-class classification optimization.
Optimizer	Adam optimizer with adaptive learning rate mechanism.
Training Epochs	50 epochs ensuring stable convergence of loss values.
Experimental Results (Loss-Based Evaluation)	
Initial Training Loss	Approximately 85 indicating high initial classification error.
Final Training Loss	Approximately 5 demonstrating substantial error minimization.
Initial Validation Loss	Approximately 68 representing moderate generalization error.
Final Validation Loss	Approximately 3 indicating effective validation stability.
Overfitting Analysis	No significant divergence between training and validation curves observed.
System Characteristics	
Deployment Mode	Offline experimental validation only.
Real-Time Capability	Not yet integrated; planned for future enhancement.
Future Work	Integration of temporal sequence modeling, real-time inference, and GPU acceleration.

VII. CONCLUSION

Lip reading and lip-to-speech reconstruction have advanced considerably due to innovations in spatio-temporal feature extraction, attention-based sequence modeling, and high-fidelity neural vocoders. Self-supervised pretraining and contrastive cross-modal alignment (e.g., LipSound2-like approaches) have reduced reliance on perfectly paired datasets and improved generalization in many cases. However, real-world deployment still faces substantial barriers: dataset bias, speaker/style generalization, pose and occlusion robustness, evaluation standardization, computational efficiency, and privacy concerns.

Future research should prioritize domain diversity (languages, recording conditions), model efficiency for edge inference, joint multimodal objectives (speech + text + speaker), and ethical frameworks ensuring privacy-preserving use. When these challenges are addressed, lip-to-speech systems can become practical tools for hearing assistance, silent communication interfaces, forensic analysis, and robust audio-visual human-machine interaction.

REFERENCES

- [1] L. Lu, J. Yu, Y. Chen, H. Liu, Y. M. Zhu, L. Kong, and M. Li, "Lipreading based user authentication through acoustic sensing on smartphones," *IEEE/ACM Trans. Netw.*, vol. 27, no. 1, pp. 447–460, Feb. 2019.
- [2] S. Mathulaprangsan, C.-Y. Wang, A. Z. Kusum, T.-C. Tai, and J.-C. Wang, "A survey of visual lip reading and lip-password verification," in *Proc. Int. Conf. Orange Technol. (ICOT)*, Dec. 2015, pp. 22–25.
- [3] T. R. 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. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7832–7841.
- [5] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13796–13805.
- [6] R. Ding, C. Pang, and H. Liu, "Audio-visual keyword spotting based on multidimensional convolutional neural network," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4138–4142.
- [7] A. Fernandez-Lopez and M. F. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image Vis. Comput.*, vol. 78, pp. 53–72, Oct. 2018.
- [8] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.
- [9] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Comput. Vis. Image Understand.*, vol. 173, pp. 76–85, Aug. 2018.
- [10] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition (Speech Reading)," Univ. Illinois at Urbana-Champaign, 1984.
- [11] K. Mase and A. Pentland, "Automatic lipreading by optical-flow analysis," *Syst. Comput. Jpn.*, vol. 22, no. 6, pp. 67–76, 1991.
- [12] N. Eveno, A. Caplier, and P.-Y. Coulon, "Accurate and quasi-automatic lip tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 706–715, May 2004.
- [13] R. Bowden, S. Cox, R. Harvey, Y. Lan, and B. J. Theobald, "Recent developments in automated lip-reading," *Proc. SPIE*, vol. 8901, Oct. 2013, Art. no. 89010J.
- [14] Z. Zhou, G. Zhao, X. Hong, and M. Pietikainen, "A review of recent advances in visual speech decoding," *Image Vis. Comput.*, vol. 32, no. 9, pp. 590–605, Sep. 2014.
- [15] R. Seymour, D. Stewart, and J. Ming, "Comparison of image transform based features for visual speech recognition in clean and corrupted videos," *EURASIP J. Image Video Process.*, vol. 2008, no. 2, p. 19, 2008.
- [16] D. Bhatt et al., "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021.
- [17] C. Patel et al., "DBGC: Dimension-based generic convolution block for object recognition," *Sensors*, vol. 22, no. 5, p. 1780, Feb. 2022.
- [18] K. Noda et al., "Lipreading using convolutional neural network," Northern Territory Library, Darwin, NT, Australia, 2014.
- [19] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multi-view lipreading," in *Proc. British Machine Vision Conf.*, 2017.
- [20] J. Xiao et al., "Deformation flow based two-stream network for lip reading," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 364–370.
- [21] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. IEEE ICASSP*, May 2020, pp. 6319–6323.
- [22] P. Ma et al., "Towards practical lipreading with distilled and efficient models," 2020, arXiv:2007.06504.
- [23] S. Petridis et al., "End-to-end audiovisual speech recognition," in *Proc. IEEE ICASSP*, Apr. 2018, pp. 6548–6552.
- [24] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-end lipreading with cascaded attention-CTC," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 548–555.
- [25] X. Zhang, F. Cheng, and W. Shilin, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 713–722.
- [26] M. Luo, S. Yang, S. Shan, and X. Chen, "Synchronous bidirectional learning for multilingual lip reading," in *Proc. BMVC*, 2020.
- [27] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3D-CNNs for lip reading," in *Proc. 30th British Machine Vision Conf.*, 2019.
- [28] D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an effective lip reading model without pains," 2020, arXiv:2011.07557.
- [29] J. Hu et al., "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
- [30] H. Wu et al., "CvT: Introducing convolutions to vision transformers," 2021, arXiv:2103.15808.
- [31] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 87–103.
- [32] S. Yang et al., "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, p. 18.
- [33] A. V. Aswani et al., "Attention is all you need," in *Proc. NIPS*, 2017.
- [34] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019.
- [35] D. Mahajan et al., "Exploring the limits of weakly supervised pretraining," in *Proc. Eur. Conf. Comput. Vis.*, 2018.
- [36] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
- [37] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021, arXiv:2102.12122.
- [38] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," 2021, arXiv:2101.11986.
- [39] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, arXiv:1412.3555.
- [40] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE ICCVW*, Oct. 2017, pp. 3154–3160.
- [41] H. Zhang et al., "Mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.
- [42] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, "Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs," *Comput. Vis. Image Understand.*, vol. 176, pp. 22–32, Nov. 2018.
- [43] M. Kim, J. Hong, S. J. Park, and Y. M. Ro, "Multi-modality associative bridging through memory: Speech sound recollected from face video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 296–306.
- [44] M. Kim, J. H. Yeo, and Y. M. Ro, "Distinguishing homophones using multi-head visual-audio memory for lip reading," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1174–1182.