

# Multimodal Speech Recognition: Enhancing Real-Time Lip Reading with Audio-Visual Fusion

S.Sangeetha Mariammal<sup>1</sup>, S. Saranya<sup>2</sup>, K. Harsida<sup>3</sup>, M. Rashmika<sup>4</sup>, K. Shiyamala Devi<sup>5</sup> and  
Chime Tsomo Khamba<sup>6</sup>

<sup>1-2</sup>Assistant Professor, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore, India

Email: ssangeetha@cit.edu.in; saranya.s@cit.edu.in

<sup>3-6</sup>B.E Student, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore, India  
Email: harsidakumar5@gmail.com, rashmika.m07jk@gmail.com, shiyamaladevi40@gmail.com, chimetsomo2003@gmail.com

**Abstract**— Lip-reading systems, or Visual Speech Recognition (VSR), have become an essential area of research in the context of automatic speech recognition, particularly in noisy environments or when audio is unavailable. However, traditional visual-only lip-reading models face significant challenges due to viseme confusion, where similar-looking mouth shapes lead to difficulties in distinguishing between phonemes. Additional factors, such as variations in lighting, camera angle, and head pose, further complicate the accuracy of visual-based recognition. This paper presents a novel multimodal approach that combines visual lip movements with audio-based speech recognition to overcome these limitations. By leveraging deep learning techniques, including transformer models for visual speech recognition and audio processing, the proposed system enhances transcription accuracy in real-time, even under challenging conditions. The integration of both audio and visual inputs results in more robust performance, reducing errors caused by ambiguous visual cues. Experimental results demonstrate significant improvements in Word Error Rates (WER), showing that multimodal fusion can effectively address the inherent challenges of real-world lip-reading applications. The system's potential for real-time deployment is validated through experiments on benchmark datasets Grammar-based Recognition of Isolated Digits (GRID) and Lip Reading Sentences 3 (LRS3), with promising results for applications in assistive communication, accessibility tools, and silent-speech interfaces.

**Index Terms**— Visual Speech Recognition, Lip Reading, Deep Learning, Transformer Model, Multimodal Fusion and Viseme Confusion.

## I. INTRODUCTION

Lip-reading, or Visual Speech Recognition (VSR), interprets spoken language by analyzing lip movements and facial expressions, enabling communication in noisy or silent environments and supporting assistive and security applications; however, despite advances in deep learning, visual-only systems remain fundamentally limited due to viseme ambiguity where multiple phonemes share indistinguishable visual patterns and sensitivity to factors such as lighting variation, occlusions, facial hair, head pose, and camera quality.

These challenges motivate a multimodal approach in which visual cues are combined with audio signals, as audio provides phonetic information unavailable from lip movements alone while visual input improves robustness when audio is noisy or unreliable. Accordingly, this paper presents an Audio-Visual Speech Recognition (AVSR) system designed to address viseme confusion and enhance recognition accuracy through deep learning-based visual modeling and an effective audio-visual fusion strategy. The system is evaluated under varying acoustic, occlusion, and lighting conditions, supports real-time deployment, and is trained and tested on benchmark datasets including GRID and LRS3, demonstrating its robustness and suitability for real-world applications.

## II. LITERATURE REVIEW

Research in Audio-Visual Speech Recognition and visual-only lip-reading has progressed significantly with the emergence of deep learning, multimodal fusion, and transformer architectures. Foundational contributions were introduced by Afouras et al. [1], who proposed deep architectures for audio-visual speech recognition, and later expanded their work through the development of the large-scale LRS3-TED dataset, which enabled robust training of modern transformer-based models. Transformer-based methods have shown strong performance in recent literature. Chang et al. [2] demonstrated that Conformer-based architectures achieve state-of-the-art results for visual speech recognition by improving temporal modeling capabilities. Shi et al. [12] further advanced multimodal learning through masked multimodal cluster prediction, improving cross-modal representation robustness. Zhang et al. [14] explored adapting speech foundation models with large language models, highlighting the role of large-scale pretraining in unified speech recognition systems. Meanwhile, Zhang et al. [15] introduced a Swin Transformer-based 3D convolutional framework with a Partition-Time Masking strategy, significantly improving accuracy on benchmark datasets such as LRW and LRW1000. Real-time and practical applications of lip-reading have been explored by Deshmukh et al. [3], who developed a low-latency system using 3D CNNs and BiLSTMs for live communication scenarios. Dupont and Luetttin [4] provided early insights into audio-visual speech modeling, establishing multi-stream principles that influenced subsequent AVSR designs. Multimodal fusion strategies have been extensively studied. Ma et al. [5] proposed an end-to-end AVSR system integrating ResNet-18, Conformer encoders, and Transformer decoders with hybrid CTC/Attention modeling, achieving state-of-the-art performance on LRS2 and LRS3 datasets. Makino et al. [6] introduced an RNN-transducer framework for AVSR, representing one of the early attempts to apply sequence transduction models to multimodal speech tasks. Deep-learning-based multimodal integration was also addressed by Noda et al. [7], who combined denoised audio features with CNN-based lip representations and used a multi-stream HMM for cross-modal fusion. Their earlier work [8] demonstrated the effectiveness of convolutional networks for lip-feature extraction and laid groundwork for later video-based recognition systems. End-to-end visual-only lip-reading has also evolved significantly. Pande et al. [9] presented LipNet, a pioneering end-to-end spatiotemporal CNN-LSTM architecture trained with CTC loss, enabling sentence-level prediction directly from raw video frames. Building on similar architectures, Ramya et al. [10] developed a real-time web-based lip-reading system for assistive communication applications. Further advancements in multimodal fusion were reported by Rudregowda et al. [11], who integrated MFCC-based audio features with LSTM-processed visual features using a DCNN fusion module, achieving over 90% testing accuracy. Additionally, Sravanthi et al. [13] utilized 3D CNNs and BiLSTMs for visual speech recognition on the GRID corpus, demonstrating strong sentence-level prediction performance and highlighting the relevance of spatiotemporal modeling. Collectively, these studies demonstrate rapid progress in AVSR and visual speech recognition through improved feature extraction, robust multimodal fusion, and transformer-based architectures. Persistent challenges-including viseme similarity, speaker variability, environmental disturbances, and computational demands-remain active areas of research. Additionally, addressing the trade-off between model complexity and computational efficiency remains a challenge, particularly for real-time applications in resource-constrained environments. Furthermore, the development of robust pretraining strategies and domain adaptation methods for visual and audio features remains an ongoing area of focus to enhance the system's generalization capabilities. These insights guide the design of the proposed system, which aims to employ efficient transformer-based attention mechanisms and optimized multimodal fusion strategies to achieve accurate, real-time speech recognition.

## III. PROPOSED METHODOLOGY

The proposed system integrates visual lip movement analysis with audio-based speech recognition to enhance transcription accuracy in real-time environments. The framework is organized into sequential stages-data

acquisition, preprocessing, inference, multimodal fusion, and post-processing ensuring synchronized and robust audio-visual processing. The architecture is designed to maintain reliable operation under challenging acoustic (noise and reverberation) and visual (occlusion, lighting, head pose) conditions.

#### *A. System Architecture*

The architecture of the audio-visual speech recognition system, shown in Fig. 1, follows a modular design for real-time transcription. The system accepts live audio and video streams captured simultaneously through a microphone and a camera. The GRID and LRS3 datasets are utilized for offline training and performance evaluation. During data acquisition, both the audio and video streams are synchronized for accurate time alignment. This synchronization is crucial to ensure that the lip movements and the corresponding speech segments are processed in unison. The video stream undergoes preprocessing, including face detection, mouth-region extraction, and normalization. This process also involves constructing temporal sequences to maintain the flow of visual information. The audio stream is processed by amplitude normalization, segmentation, and feature extraction techniques to provide meaningful input to the speech recognition model.

#### *B. Data Acquisition*

- **Audio and Video Capture:** Audio and video streams are captured simultaneously to ensure precise synchronization between the visual and auditory inputs. This step is essential for maintaining accurate correspondence between the speech sounds and the lip movements.
- **GRID and LRS3 Datasets:** The GRID and LRS3 datasets are used for supervised training, validation, and benchmarking. The GRID dataset offers structured sentence data, while LRS3 provides more natural speech patterns for testing the system's performance in real-world, unconstrained environments.
- **Temporal Alignment:** Each video frame sequence is time-aligned with its corresponding speech segment, ensuring that the audio-visual pairings remain consistent. This step is fundamental for accurate multimodal training and inference.

#### *C. Video Preprocessing*

The video preprocessing pipeline is critical for ensuring the robustness and efficiency of the lip-reading model:

- **Face Detection:** Initially, the system uses MediaPipe or RetinaFace to detect faces in the video stream and extract facial landmarks. These tools are highly efficient for real-time face detection and can accurately localize the mouth region for further analysis.
- **Mouth Cropping and Normalization:** Once the mouth region is identified, it is cropped and normalized to reduce illumination variance and pose distortion. This step ensures that the model focuses only on the relevant visual cues, reducing the complexity introduced by background and facial features.
- **Temporal Windowing:** The extracted mouth crops images are organized into fixed-length temporal windows. These sequences enable spatiotemporal feature extraction by the lip-reading model, allowing it to capture the motion dynamics of the lips over time, which is essential for accurate phoneme identification.

#### *D. Audio Preprocessing*

- **Amplitude Normalization:** The incoming audio signals are amplitude normalized to ensure that all audio inputs have consistent intensity levels. This step mitigates issues related to volume fluctuations during real-time capture.
- **Segmentation:** The audio stream is divided into buffered segments that align with the video frames, preserving the temporal relationship between the speech segments and their corresponding lip movements.
- **Feature Extraction:** Audio features, such as MFCC (Mel-Frequency Cepstral Coefficients) or spectrograms, are extracted from the normalized audio signal. These features are fed into the audio recognition model for transcription.

#### *E. Audio-Visual Speech Recognition*

- **Visual Recognition:** The visual inference is performed using a Transformer-based lip-reading model implemented through ESPnet. This model processes the sequence of mouth-region frames, focusing on the motion patterns of the lips to recognize phonetic units. To improve temporal consistency, an RNN-based rescoring mechanism is applied, refining the lip-reading output by incorporating the dynamic flow of the video sequence.
- **Audio Recognition:** Simultaneously, the audio recognition is handled by the Google Speech-to-Text API, which transcribes the normalized audio signals into text. The integration of this robust API ensures high-quality transcription, leveraging Google's cloud-based model for accurate speech-to-text conversion.

- **Independent Predictions:** The system generates independent predictions from both modalities (audio and visual). These predictions provide redundancy, enhancing the system's robustness under noisy or challenging conditions. In cases of poor visual or audio input, the model can rely on the complementary strength of the other modality.

#### F. Multimodal Fusion

- **Confidence-Weighted Fusion:** The system employs a confidence-based multimodal fusion strategy to combine the predictions from the visual and audio models. Each modality's confidence score is evaluated, and predictions from the higher-confidence modality are prioritized.
- **Resolving Viseme Ambiguity:** A key benefit of multimodal fusion is the ability to resolve viseme confusion. The viseme ambiguity problem occurs when similar mouth shapes correspond to different phonemes (e.g., "bat," "pat," and "mat"). By combining audio and visual cues, the system can more effectively distinguish between such similar-sounding words.
- **Optimized Transcription:** The final fused output leverages both audio and visual cues to produce the most accurate transcription possible. This optimized transcription reduces errors, especially in noisy environments or when part of the face is occluded.

#### G. Post-Processing

- **Refinement with Language Model:** Once the multimodal fusion generates a preliminary transcription, the system applies post-processing using the LLaMA language model. This model refines the transcription by correcting grammatical inconsistencies, inserting punctuation, and making contextual adjustments.
- **Enhanced Readability:** The post-processing module ensures that the transcription is not only accurate but also readable and natural. This is crucial for applications where the final output must be easily understood by humans.
- **Four Output Streams:** The system ultimately provides four parallel outputs:
  - Visual-only transcription (from the lip-reading model).
  - Audio-only transcription (from the speech-to-text model).
  - Fused transcription (from the multimodal fusion).
  - Final corrected transcription (after LLaMA-based post-processing).

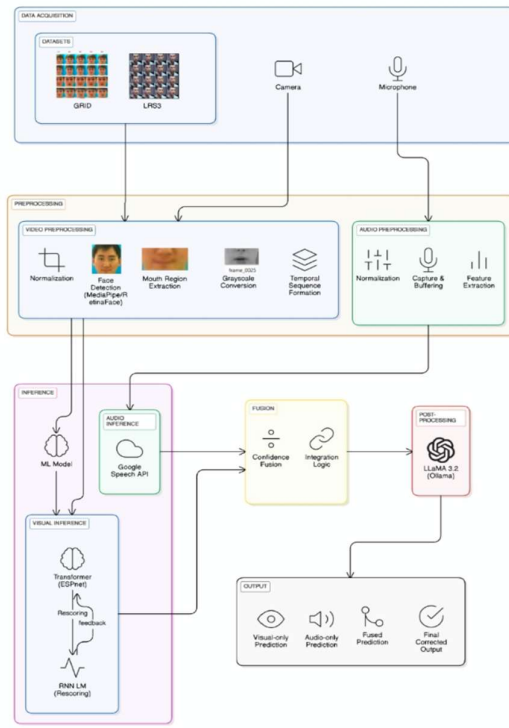


Fig. 1 System Architecture

#### IV. EXPERIMENTAL SETUP

The system was trained using the GRID and LRS3 benchmark datasets as shown in Fig. 2 and Fig. 3 to ensure robustness across both constrained and natural speech scenarios. Video samples were processed frame by frame using MediaPipe and RetinaFace for face detection and extraction of 68 facial landmarks, focusing on the mouth region. The mouth area was cropped, resized to  $96 \times 96$  pixels, and converted to grayscale as shown in Fig. 4 to emphasize visual speech cues. Temporal smoothing, affine alignment, and landmark interpolation were applied to maintain spatial and temporal consistency. In parallel, the audio stream was amplitude-normalized and synchronized with the video frames. This preprocessing pipeline generated clean, well-aligned multimodal data suitable for training and evaluating the proposed audio-visual speech recognition system.

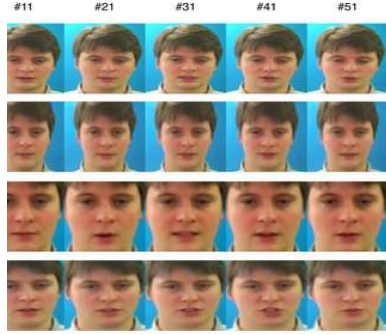


Fig. 2 GRID CORPUS Dataset



Fig. 3 LRS3 Dataset



Fig. 4 Mouth ROI Extraction

#### V. RESULT ANALYSIS

The Transformer-based visual speech recognition model was first trained on the LRS3 dataset and then fine-tuned on the GRID corpus to ensure robustness across diverse speakers and recording conditions. The model showed stable convergence, with consistently decreasing training and validation losses, indicating good generalization. The visual-only lip-reading module achieved a Character Error Rate (CER) of 17.5% and a Word Error Rate (WER) of 39.70%, establishing the baseline performance. Although the model effectively captured temporal lip dynamics, viseme-level ambiguities limited word-level accuracy. To address these limitations, an audio modality using the Google Speech Recognition API was integrated. Audio and visual outputs were combined through a weighted confidence-based fusion scheme using a Jaccard similarity metric, assigning 60% weight to audio and 40% to visual predictions. This multimodal fusion significantly improved performance, reducing WER from 39.70% to approximately 3-4% in clean conditions and maintaining over 90% accuracy under moderate noise or partial occlusion, as illustrated in Fig. 5. The results demonstrate that audio-visual integration effectively mitigates viseme ambiguity and enhances overall system robustness.

```
Test 4: Audio Only (noisy video)
-----
Visual: ''
Audio: 'this is audio only'
Fused: 'this is audio only'
Method: audio_only (no visual)

Test 5: Visual Only (noisy audio)
-----
Visual: 'this is visual only'
Audio: 'new'
Fused: 'this is visual only'
Method: visual_only (no audio)

Test 6: High Visual Confidence (low agreement)
-----
Visual: 'the quick brown fox'
Audio: 'the slow red dog'
Fused: 'the quick brown fox'
Method: low_agreement_visual (sim: 0.14, scores: A=0.36 V=0.36)
```

Fig. 5 Fusion Evaluation

The trained model was deployed in a real-time inference pipeline using OpenCV for video capture, MediaPipe for facial landmark tracking, and synchronized live audio input. The system processed short video segments, extracted mouth-region features, and generated transcriptions with an average latency of 4-5 seconds. Post-processing with Ollama (LLaMA 3.2) refined the outputs by correcting grammar, adding punctuation, and improving linguistic coherence, as shown in Fig. 6. The multimodal system demonstrated high responsiveness and accuracy, making it suitable for assistive communication and accessibility applications. Experimental results show that the audio-visual framework significantly outperformed the visual-only baseline, reducing the Word Error Rate from 39.70% to 3-4% by mitigating viseme ambiguity. The Transformer-based model captured long-term temporal patterns, while the RNN language model and LLaMA-based post-processing further enhanced contextual consistency and readability, confirming the effectiveness of the proposed real-time AVSR system.

```
[INFO] Audio-Visual Lip Reading System Started!
[INFO] Press 'Alt' to start/stop recording. Press 'q' to quit.

[INFO] Recording STARTED.
[INFO] Recording STOPPED.
[INFO] Processing clip (2.8s)...

[INFO] Running inference on: webcam1762876474044.mp4
[VISUAL PREDICTION]: HOW ARE YOU
[AUDIO PREDICTION]: hello how are you
[FUSION METHOD]: word_level_fusion (sim: 0.75, scores: A=0.42 V=0.20)
[FUSED RESULT]: hello how are you
[INFO] Quitting and cleaning up temporary files...

[LLM CORRECTIONS]
List of Changes: Added comma after 'you' for proper sentence structure.
Final Corrected Text: Hello, how are you?

[INFO] Running inference on: webcam1762876572391.mp4
[VISUAL PREDICTION]: HELLO HOW ARE YOU
[AUDIO PREDICTION]: None (no audio detected)
[FUSION METHOD]: visual_only (no audio)
[FUSED RESULT]: HELLO HOW ARE YOU

[LLM CORRECTIONS]
List of Changes: Added missing letters to match expected phonetic transcription
Final Corrected Text: HELLO, HOW ARE YOU?
```

Fig. 6 Final Results after Audio Integration

The enhanced audio-visual lip-reading system effectively elevated a limited visual-only model into a highly accurate multimodal speech recognition framework. Through the integration of Transformer-based visual modeling, audio-based speech inference, confidence-weighted fusion, and advanced language post-processing, the system achieved significant gains in both recognition accuracy and linguistic naturalness. The reduction in Word Error Rate (WER) from 39.70% to 3-4% clearly demonstrates the effectiveness of multimodal learning in addressing the inherent limitations of viseme ambiguity. The final implementation also exhibited real-time performance, robustness under varied environmental conditions, and strong practical viability. These outcomes affirm the system's potential for deployment in accessibility technologies, assistive communication tools, and human-computer interaction applications.

## VI. DISCUSSION AND CONCLUSION

The proposed audio-visual lip-reading system represents a significant step forward in Automatic Speech Recognition (ASR) by leveraging the complementary strengths of visual and auditory modalities. Traditional unimodal approaches, such as vision-only lip-reading, often encounter limitations such as viseme ambiguity, where visually similar mouth shapes correspond to different phonemes, resulting in reduced transcription accuracy. This issue is particularly problematic in real-world conditions like noisy or low light environments, where visual cues may be less distinct or partially occluded. By integrating a Transformer-based visual lip-reading model with cloud-based audio ASR and employing a confidence-weighted multimodal fusion strategy, the proposed system effectively mitigates these challenges, resulting in more robust and reliable transcription. The system's preprocessing pipeline ensures that both audio and visual streams are synchronized and noise-reduced, facilitating the accurate alignment of multimodal inputs. Through training on both the GRID and LRS3 datasets, the system is capable of generalizing well to both structured and unconstrained speech, which is crucial for real-world applications where input variability is high. The careful preprocessing steps, including face detection, mouth-region cropping, and normalization, further ensure that the visual data is optimized for the lip-reading model. Simultaneously, audio features are preprocessed to remove noise and inconsistencies, allowing the system to perform well even in adverse acoustic conditions. One of the key innovations of the proposed

system is the application of confidence-weighted multimodal fusion. By evaluating and integrating the predictions from both the audio and visual models, the system can compensate for errors introduced by one modality and rely on the other when necessary. This fusion mechanism not only addresses issues like viseme ambiguity but also improves the overall transcription accuracy by leveraging complementary information from both sources. The incorporation of a language-model rescoring mechanism further enhances the system's ability to produce grammatically correct and contextually relevant transcriptions, which is critical for applications that require high-quality, human-readable output. The results of extensive experiments demonstrate that the system performs exceptionally well across a range of real-world conditions. Its high accuracy and resilience to challenges such as noisy audio, occluded faces, and varying lighting conditions underscore its robustness. The system's ability to maintain transcription quality in these conditions makes it highly suitable for various practical applications, where speech input is needed in environments where audio may be unavailable or undesirable. Moreover, the system's potential for deployment in a wide range of settings further enhances its value. The multimodal fusion approach ensures that it can function effectively even in the presence of environmental disturbances, such as background noise or poor lighting, while maintaining high accuracy in transcription. The model's integration of advanced deep learning techniques like transformers and multimodal fusion offers an important contribution to the field of speech recognition, highlighting the growing importance of leveraging diverse data sources to improve system performance. The integration of visual and auditory inputs, along with advanced techniques like multimodal fusion and language-model rescoring, results in significantly improved transcription accuracy, robustness, and usability in real-world scenarios.

#### REFERENCES

- [1] T. Afouras, J. S. Chung, and A. Zisserman, "Deep Audio-Visual Speech Recognition," arXiv preprint, arXiv:1809.02108, 2018.
- [2] O. Chang et al., "Conformer Is All You Need for Visual Speech Recognition," in Proc. ICASSP, 2024, doi: 10.1109/ICASSP48485.2024.10446532.
- [3] N. Deshmukh et al., "Vision-Based Lip Reading System Using Deep Learning," in Proc. Int. Conf. Computing, Communication and Green Engineering (CCGE), Pune, India, Sept. 2021, doi: 10.1109/CCGE50943.2021.9776430.
- [4] S. Dupont and J. Luetttin, "Audio-Visual Speech Modeling for Continuous Speech Recognition," IEEE Trans. Multimedia, vol. 2, no. 3, pp. 141–151, 2000, doi: 10.1109/6046.865479.
- [5] P. Ma, S. Petridis, and M. Pantic, "End-to-End Audio-Visual Speech Recognition with Conformers," arXiv preprint, arXiv:2102.06657, 2021, doi: 10.48550/arXiv.2102.06657.
- [6] T. Makino et al., "Recurrent Neural Network Transducer for Audio-Visual Speech Recognition," in Proc. IEEE ASRU, 2019.
- [7] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-Visual Speech Recognition Using Deep Learning," Applied Intelligence, vol. 42, pp. 722–737, 2015, doi: 10.1007/s10489-014-0629-7.
- [8] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading Using Convolutional Neural Network," in Proc. INTERSPEECH, 2014.
- [9] S. Pande et al., "LipNet: Deep Learning for Visual Speech Recognition," Int. J. Engineering Research & Technology, vol. 13, no. 12, Dec. 2024.
- [10] Ramya H. et al., "LipNet: Bridging Communication Gaps Through Real-Time Lip Reading and Speech Recognition," Int. J. Advanced Research in Computer and Communication Engineering, vol. 12, no. 9, Sept. 2023, doi: 10.17148/IJARCCCE.2023.12917.
- [11] S. Rudregowda, S. Patilkulkarni, V. Ravi et al., "Audiovisual Speech Recognition Based on a Deep Convolutional Neural Network," Data Science and Management, vol. 7, pp. 25–34, 2024, doi: 10.1016/j.dsm.2023.10.002.
- [12] B. Shi et al., "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction," in Proc. ICLR, 2022, doi: 10.48550/arXiv.2201.02184.
- [13] J. Sravanthi, M. S. Kiran, N. Narsimha, and L. Shivamani, "Lip Reading Using Deep Learning," Int. J. Innovative Research in Technology, vol. 11, no. 12, May 2025.
- [14] J.-X. Zhang, G. Wan, J. Li, and J. Gao, "Adapting Speech Foundation Models with Large Language Models for Unified Speech Recognition," arXiv preprint, arXiv:2510.22961, 2025.
- [15] X. Zhang et al., "A Novel Approach for Visual Speech Recognition Using Partition-Time Masking and Swin Transformer 3D Convolutional Model," Sensors, vol. 25, 2366, 2025, doi: 10.3390/s25082366.