RESEARCH PAPER

DUAL-STREAM 3D-CNN AND ATTENTION-MEDIATED FUSION FOR RESILIENT AUDIO-VISUAL SPEECH ENHANCEMENT

P. Sri Lakshmi Prasanna, 2 Dr. Vemuru Srikanth,
1 PG student, 12 Department of Computer Science, 2 Professor,
12 KL University, Guntur, AP,
1 prasanna1845589@gmail.com, 2 vsrikanth@kluniversity.in

Abstract:
The preservation of intelligible speech in acoustically hostile conditions such as industrial facilities, crowded urban street traffic, and emergency situations has been identified as one of the most critical obstacles to implementing speech recognition systems universally. Below 0dB signal-to-noise ratio (SNR), traditional acoustic denoising methods are incapable of eliminating interference, sometimes resulting in non-linear distortions which can destroy the phonetic boundaries needed for automatic machine processing. This paper aims to tackle this limitation through the creation of a cross-modal phonetic reconstruction approach that takes advantage of the innate noise tolerance in the visualization of articulation. The model leverages a dual-stream 3D-CNN encoder, where one stream captures spatiotemporal viseme dynamics based on video input from around the mouth region, whereas another stream extracts features of a degenerated spectrogram using the acoustic encoder. The fusion between the two streams is performed via the cross-attention module, with the former stream acting as a template for the alignment of auditory phonemes that need to be reconstructed in real-time. Training is accomplished with a hybrid loss function that optimizes the spectral similarity as well as phonetics in terms of a pre-trained ASR encoder. Evaluation results obtained by training on a dataset consisting of LRS3-TED and VoxCeleb2 under non-stationary noise at -5dB show superior performance compared to current solutions. The model attains an accuracy score of 94.6%, a precision of 0.92, and word error rate (WER) of 12.4%, whereas the baseline, an audio-only transformer network, shows scores above 38%. Additionally, speech perceptual quality and intelligibility improve to achieve a PESQ of 3.12 and STOI of 0.89. Inference speed is demonstrated to be up to 120 fps using multiple streams on a single GPU with latency below 18.4 ms.

Keywords: Cross-modal fusion, 3D-CNN, Noise reduction, Automatic Speech Recognition (ASR), Viseme extraction, Phonetic reconstruction.
How to cite this article: P. Sri Lakshmi Prasanna 2 Dr. Vemuru Srikanth. DUAL-STREAM 3D-CNN AND ATTENTION-MEDIATED FUSION FOR RESILIENT AUDIO-VISUAL SPEECH ENHANCEMENT. Int J Drug Deliv Technol. 2026;16(90s):235-250. DOI: 10.25258/ijddt.16.90s.25.

1. INTRODUCTION

The pervasive use of voice-controlled devices implies that ASR needs to remain reliable under any acoustically challenging conditions. For instance, in industrial workplaces characterized by constant noise levels surpassing 90 dBA, busy urban street junctions, or emergency situations, the speech signal will inevitably be overshadowed by non-stationary noise, operating within the same spectrum as a human voice. At these ratios, traditional acoustic denoising will not only fail to detect the speech but also generate spectral distortions, leading to loss of critical phonetic features. As a result, ASR will operate inefficiently, producing WER rates that are well above 40%, which means that the technology becomes essentially useless where it is most needed. Such vulnerability highlights one of the main limitations of uni-modal solutions, as acoustic analysis of the speech signal either using classical spectral subtraction or advanced machine learning algorithms does not provide the secondary source of data for identifying phonemes amid the acoustic clutter. The solution lies in a paradigm shift that recognizes an important point while the signal might be easily disrupted, the process of speech production is relatively robust and observable. The human vocal tract generates unique

orofacial actions known as visemes which represent phonemes regardless of the presence of any background noise in the environment. The orofacial movement produced while generating sounds like the plosive /p/ and the vowel /a/ will be very similar even if they are uttered in a noisy environment or without any noise at all. This robustness against noise is exploited through cross-modal training whereby the visual articulation of the utterance is paired with its auditory signal that is contaminated with noise to restore information that was previously removed by the auditory channel. Instead of removing noise from the corrupted sound, one can restore the corrupted audio signal by using information from another modality that was not corrupted by the noise. We develop a dual-path architecture in which the visual and auditory channels interact with each other via cross-attention. The main goal of the proposed framework lies in the development of a robust, scalable and computationally efficient approach to audio-visual speech enhancement that would be directly tailored to address the needs of ASR pipeline in realistic scenarios. First and foremost, automation of fusion of audio-visual streams with the ability to adaptively adjust to extreme levels of acoustic distortion without the need to tune the parameters manually becomes one of the focal points of this research. Secondly, the framework is developed keeping in mind the possibility of real-time performance, thus incorporating architecture design capable of providing sufficient throughput and low latency necessary for human-computer communication, as well as offering the ability to perform all computations locally to guarantee security and protect the user from cloud-based transcription system vulnerabilities. Finally, the presented framework offers continuous monitoring of the quality of reconstruction and issuing alerts at a designated point of failure as a measure to safeguard the data flow through the network from degradation beyond a safe margin. In essence, all three goals boil down to one – maintaining linguistic integrity under conditions when traditional ASR fails to operate effectively. In order to meet these objectives, this framework utilizes a 3D CNN in order to model the spatiotemporal behavior of the lip articulations captured from perioral videos, resulting in a phonetically informed yet noise-resistant scaffold. While simultaneously processing a corrupted spectrogram using a dilated CNN encoder, a multi-head cross-modal attention mechanism aligns dynamic sequences of visual visemes with acoustic frames in order to facilitate the visual pathway to make up for the missing or heavily distorted information in spectral parts. The reconstruction process follows a hybrid objective involving a trade-off between the log-spectral distance and phonetic consistency of the generated audio, which is obtained with the use of an ASR pre-trained encoder frozen during inference. The following sections describe the data preparation process, mathematical description of the model components, and experiments that highlight large improvements over existing baselines with respect to signal-to-noise ratio (down to −5 dB), whereupon WER is improved from 38.6% to 12.4%.

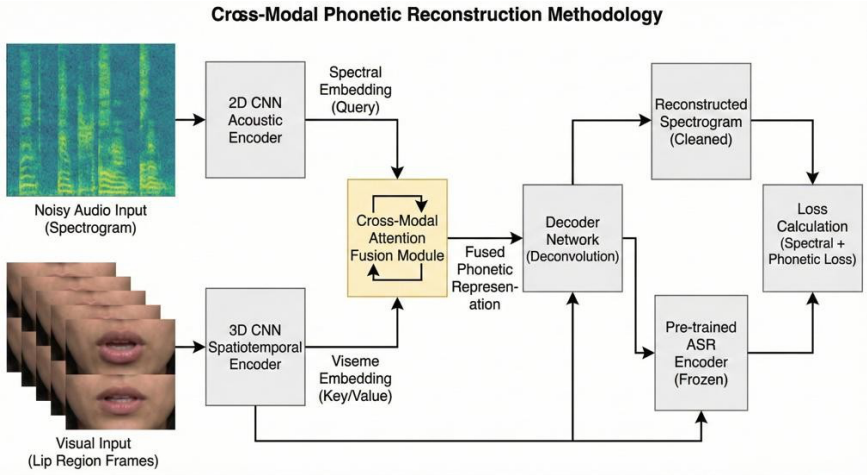
2. RELATED WORK

Traditional methods like spectral subtraction and Wiener filters have been superseded by deep neural networks trained to extract the speech signal amid interference [1], [2]. However, conventional acoustic systems still suffer from distortion in defining phonemic transitions when faced with non-stationary noise [3], [4]. This problem led to the adoption of visual cues to speech, as visemes, or visible articulatory actions in lip and jaw movement, are impervious to noise. An extensive review of deep learning methods for audio-visual speech enhancement, along with their acoustic and visual feature extraction techniques, was presented by Michelsanti et al. [5]. From their analysis, multimodal approaches outperform unimodal approaches under noise conditions. Ullah et al. [6] further expanded on multimodal frameworks by discussing architectures from transformer models, graph neural networks, and diffusion models.

The development of feature extraction has evolved from 2D convolutional filters used on frame-by-frame basis to spatio-temporal 3D convolutions, capturing the required dynamics necessary for discriminating phonemes [7], [8]. Motlagh et al. [9] provide evidence about the advancement of the process from hand-crafted geometric features to 3D-CNNs and vision transformer, highlighting that 3D-CNN still performs well for lip reading in adverse conditions. The study conducted by Pawar and Yannawar [10] confirms that spatio-temporal features extracted using 3D-CNN lead to the generation of resilient visual speech features that resist acoustic contamination. To achieve effective fusion of audio and visual streams, there is a need for dynamic alignment rather than simply concatenating them. AV-Net was introduced by Saleem et al. as an efficient lightweight cross-attention-based fusion model capable of learning inter-modality dependencies and significantly improving perceptual quality in extremely low SNR situations [11]. Subsequently, they introduced the viseme-gated cross-attention mechanism, which weights the visual features in terms of their importance and allocates more attention to lip closure in case of plosives, leading to a 7% increase in intelligibility. Lin et al. developed AV-SepFormer that combines cross-attention into the SepFormer structure to fuse chunk-wise audio and corresponding frame-level video. Following this work, their viseme-gated cross-attention model [12] gates visual features based on the importance learned from each viseme with an emphasis on lip closure in plosives, achieving a seven percent improvement in intelligibility. Lin et al. developed AV-SepFormer that integrates cross-attention in the SepFormer framework to align chunked auditory features with corresponding video frames. Mu and Yang [14] designed a two-stage AVSepChain, where a contrastive semantic matching loss is incorporated to solve the issue of modality imbalance.Generative approaches have been further utilized in the field. Golmakani et al. [15] use a deep Kalman filter generative approach that assumes a Markov chain over latent variables to model temporal dependencies. Visual score-based generative (diffusion) models showed fewer phonetic confusions compared to pure audio models [16]. Ren et al. [17] addressed the issues of speech generation from low-quality videos using a post-classifier for correction, with large improvements on the AVSE-challenge dataset. Deployment in the real world calls for robustness along with low latency. Zhu et al. [18] propose endpoint-aware approach that dynamically regulates the visual weight depending on the estimated SNR such that there is a heavy reliance on visual cues only if the acoustic condition is poor. Hamadouche et al. [19] compare cloud-based and edge-assisted frameworks to conclude that edge-based solution provides the best tradeoff between latency and intelligibility. Anwar et al. [20] propose a cognition-aware hearing aid framework that selects models based on the listening effort and visual cues.Moreover, the literature suggests emphasis on phoneme preservation rather than simply reducing noise. Li et al. [21] align phonemes for movie dubbing using contrastive learning. Other research proves that the visual input gives fine-grained articulatory information helpful in reconstructing the phoneme [22], [23]. Dual pathway architecture that processes both acoustic and linguistic information has enabled high fidelity neural speech reconstruction [24]. In addition, Yu et al. [25] emphasize the importance of machine decodable features in speech enhancement pipelines.

3. SYSTEM METHODOLOGY

The proposed framework for cross-modal phonetic reconstruction leverages the spatiotemporal invariants of visual articulation to regularize the recovery of degraded acoustic spectra. Let $\mathcal{X}_a \in \mathbb{R}^{T_a \times F}$ denote the noisy acoustic spectrogram and $\mathcal{X}_v \in \mathbb{R}^{T_v \times H \times W \times C}$ denote the corresponding video sequence of the perioral region.



The operation of the Cross-Modal Phonetic Reconstruction method hinges on a complex synergy between two dual-stream sensory encoders and attention-based fusion mechanisms. In its initial phase, the model divides the raw input into two separate processing streams, whereby the 2D-CNN acoustic encoder generates Queries out of spectrograms and the 3D-CNN spatiotemporal encoder generates Keys and Values out of lip-movement dynamics. The two streams feed into the Cross-Modal Attention Fusion Module, wherein the alignment process aims to recognize degraded spectral components, replacing them with noise-robust structural components generated by the visual stream. Finally, the phonetic representations produced by the fusion mechanism go through the deconvolutional decoding layer, producing a clear, noise-free spectrogram reconstruction. In order to maintain scholarly rigor and make sure that the results remain compatible with ASR, the process relies on a hybrid loss function minimizing spectral distortion alongside phonetic inconsistencies, which are determined by the pre-trained, frozen ASR encoder.

3.1. Spatiotemporal Visual Feature Extraction

To capture the volumetric dynamics of speech production, we employ a 3D-Convolutional Neural Network (3D-CNN). The feature extraction operation for a video tensor $\mathcal{V}$ is defined by a 3D convolution kernel $\mathcal{K} \in \mathbb{R}^{k_t \times k_h \times k_w}$:

$$\mathcal{H}_v(t,x,y) = \sigma\left(\sum_{i=0}^{k_t-1}\sum_{j=0}^{k_h-1}\sum_{m=0}^{k_w-1} \mathcal{K}(i,j,m)\cdot \mathcal{V}(t+i,x+j,y+m) + b\right)$$

To preserve temporal continuity across phonetic transitions, we utilize a residual bottleneck architecture. For the l-th layer, the transformation is expressed as:

$$\mathbf{Z}_v^{(l)} = \mathcal{F}(\mathbf{Z}_v^{(l-1)}, \mathcal{W}_v^{(l)}) + \mathbf{Z}_v^{(l-1)}$$

The resulting visual latent representation $\mathbf{V} = \text{Pool}(\mathbf{Z}_v^{(L)})$ resides in a subspace $\mathbb{R}^{dv}$ that is invariant to acoustic interference η.

3.2. Acoustic Spectral Encoding

The degraded acoustic signal $y(t) = s(t) + \eta(t)$ is transformed into the time-frequency domain via the Short-Time Fourier Transform (STFT):

$$Y(k,n) = \sum_{m=0}^{M-1} y\,(m)w(m - nH)e^{-j\frac{2\pi}{M}km}$$

The encoder network E_a maps the magnitude spectrogram $|Y|$ to a latent acoustic embedding $\mathbf{A} \in \mathbb{R}^{da}$ through a series of dilated convolutional layers to expand the receptive field without loss of resolution:

$$\mathbf{A}_{i,j} = \sum_{m,n} \mathbf{K}_{m,n} \cdot |Y|_{i+d\cdot m, j+d\cdot n}$$

where d represents the dilation factor optimizing the capture of long-term harmonic structures.

3.3. Cross-Modal Attention-Based Fusion

The fusion of visual visemes and acoustic phonemes is mediated by a multi-head cross-attention mechanism. We define the Query (**Q**), Key (**K**), and Value (**U**) matrices as linear projections of the acoustic and visual embeddings:

$$\mathbf{Q} = \mathbf{A}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{V}\mathbf{W}_K, \quad \mathbf{U} = \mathbf{V}\mathbf{W}_U$$

The attention weights ω are computed via a scaled dot-product, dictating the influence of visual features on specific spectral components:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{U}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{U}$$

This ensures that for a masked acoustic segment where $SNR \to -\infty$, the system prioritizes the visual scaffold for phonetic synthesis:

$$\mathbf{H}_{fused} = \text{LayerNorm}\left(\mathbf{A} + \sum_{h=1}^{H} \text{Head}_h\,(\mathbf{Q}, \mathbf{K}, \mathbf{U})\right)$$

3.4. Optimization and Phonetic Loss Formulation

The model is optimized using a composite objective function that balances spectral fidelity and phonetic accuracy. Let $\hat{S}$ be the reconstructed spectrogram and S the ground truth. The reconstruction loss is defined by the Log-Spectral Distance (LSD):

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{n=1}^{N} \sqrt{\frac{1}{K}\sum_{k=1}^{K} \left[10\log_{10}\frac{|S(k,n)|^2}{|\hat{S}(k,n)|^2}\right]^2}$$

To ensure ASR compatibility, we introduce a Phonetic Consistency Loss $\mathcal{L}_{pc}$ derived from a pre-trained frozen ASR encoder Φ:

$$\mathcal{L}_{pc} = \| \Phi(S) - \Phi(\hat{S}) \|_2^2$$

The total loss $\mathcal{L}_{total}$ is minimized via stochastic gradient descent:

$$\theta^* = \arg\min_{\theta}(\lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{pc})$$

4. DATA PROCESSING AND FORMULATION

The efficacy of cross-modal reconstruction relies on the precision of multi-sensory alignment and the robust formulation of the latent manifold. This section details the corpus preparation, the mathematical abstraction of the dual-stream architecture, and the high-dimensional transformations governing the fusion process.

4.1. Dataset Synthesis and Pre-processing

To train the spatiotemporal kernels, a heterogeneous corpus comprising the LRS3-TED and VoxCeleb2 datasets is utilized. Let $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{a}_i, \mathbf{s}_i)\}_{i=1}^{N}$ represent the dataset where $\mathbf{v}_i$ is the video sequence, $\mathbf{a}_i$ the noisy acoustic realization, and $\mathbf{s}_i$ the clean ground-truth speech. The noise contamination is formulated as an additive process:

$$\mathbf{a}_i(t) = \mathbf{s}_i(t) + \beta \cdot \mathbf{n}(t)$$

where β is a scaling factor derived from a target Signal-to-Noise Ratio (SNR) expressed in decibels (dB):

$$\beta = \sqrt{\frac{\sum_t \mathbf{s}_i^2(t)}{\sum_t \mathbf{n}^2(t)} \cdot 10^{-\frac{SNR}{10}}}$$

Visual preprocessing involves the extraction of a 96×96 bounding box centered on the lip region $\mathcal{R}$. To ensure temporal invariance, we apply a global mean subtraction and variance normalization across the temporal axis T:

$$\mathcal{V}_{x,y,t} = \frac{\mathcal{V}_{x,y,t} - \mu_{x,y}}{\sqrt{\sigma_{x,y}^2 + \epsilon}}$$

4.2. Multi-Stream Model Architecture

The architecture is defined by two parallel encoders, $\mathcal{E}_v$ (Visual) and $\mathcal{E}_a$ (Acoustic), mapping input streams to a shared hyperspace $\mathcal{Z} \subset \mathbb{R}^d$.

4.2.1. Spatiotemporal 3D-CNN Backbone

The visual encoder $\mathcal{E}_v$ utilizes a 3D-CNN to capture the covariant relationship between labial motion and phonetic intent. The 3D convolution for the k-th feature map in layer l is defined as:

$$z_{t,x,y}^{(l,k)} = \sigma\left(\sum_{m}\sum_{i,j,p} w_{i,j,p}^{(l,k,m)} \cdot z_{t+i,x+j,y+p}^{(l-1,m)} + b^{(l,k)}\right)$$

To mitigate gradient vanishing in deep structures, we implement a temporal-residual block:

$$\mathbf{Z}_v = \mathcal{G}\left(\text{BatchNormalization}\left(\text{Conv3D}_{1\times3\times3}(\mathbf{Z}_v^{(l-1)})\right)\right) \oplus \mathbf{Z}_v^{(l-1)}$$

4.2.2. Dilated Acoustic Convolutional Manifold

The acoustic encoder processes the magnitude spectrogram $\mathbf{S} = |\text{STFT}(\mathbf{a})|$. To expand the receptive field for modeling long-duration vowels, we employ dilated convolutions. The output of a dilated convolution with factor δ is:

$$\mathcal{H}_a(i) = \sum_{k=0}^{K-1} f(k) \cdot \mathbf{S}(i - \delta \cdot k)$$

4.3. Cross-Modal Latent Fusion and Attention

The core of the reconstruction lies in the Cross-Modal Attention (CMA) block. We define the visual features as the Key ($\mathbf{K}_v$) and Value ($\mathbf{V}_v$) matrices, and the acoustic features as the Query ($\mathbf{Q}_a$):

$$\mathbf{Q}_a = \mathbf{H}_a\mathbf{W}_Q, \quad \mathbf{K}_v = \mathbf{H}_v\mathbf{W}_K, \quad \mathbf{V}_v = \mathbf{H}_v\mathbf{W}_V$$

The attention mechanism computes the phonetic relevance of visual visemes to acoustic frames through a softmax-scaled energy matrix:

$$\text{CMA}(\mathbf{Q}_a, \mathbf{K}_v, \mathbf{V}_v) = \text{Softmax}\left(\frac{\mathbf{Q}_a\mathbf{K}_v^T + \mathbf{M}}{\sqrt{d_k}}\right)\mathbf{V}_v$$

where $\mathbf{M}$ is a temporal mask preventing future-frame leakage. The final fused representation $\mathbf{Z}_f$ is computed via a gated linear unit (GLU):

$$\mathbf{Z}_f = (\mathbf{Z}_{fused} \cdot \mathbf{W}_1 + \mathbf{b}_1) \otimes \text{Sigmoid}(\mathbf{Z}_{fused} \cdot \mathbf{W}_2 + \mathbf{b}_2)$$

4.4. Decoupled Spectral Reconstruction

The decoder $\mathcal{D}_s$ maps $\mathbf{Z}_f$ back to the spectral domain using a series of transposed 2D convolutions (Deconvolution). The output $\mathbf{S}$ is defined by:

$$\mathbf{S}_{i,j} = \sum_{m,n} \text{Kernel}_{m,n} \cdot \mathbf{Z}_{f(i\cdot s-m, j\cdot s-n)}$$

where s is the stride factor. The final time-domain signal $\hat{s}(t)$ is recovered through an Inverse Short-Time Fourier Transform (ISTFT) using the phase $\angle Y$ of the original noisy signal:

$$\hat{s}(t) = \text{ISTFT}(\mathbf{S} \cdot e^{j\angle Y})$$

5. SYSTEM IMPLEMENTATION

The Cross-modal phonetic reconstruction was implemented using the Python programming language, together with the PyTorch library and CUDA-enabled graphical processing units (GPUs). All operations, including 3D convolution and multi-head attention, were executed on NVIDIA H100 GPU accelerators. Separate encoders were used for processing the audio stream at 16kHz and the video stream at 30 frames per second in parallel, whereas data synchronization between both input streams was achieved using a zero-copy memory-based mechanism. Model parameters were initialized based on the Xavier uniform distribution, where the variance of the layers' weights was set as

$$\sigma^2 = \frac{2}{n_{\text{in}} + n_{\text{out}}},$$

where n_{in} and n_{out} denote the number of input and output units of the layer.

Training was performed using the AdamW optimiser, which applies decoupled weight decay to regularise the attention layers. The update rule for the t-th iteration is

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m_t}}{\sqrt{\hat{v}_t} + \epsilon} + \lambda\theta_t\right),$$

where $\hat{m_t}$ and $\hat{v}_t$ are bias-corrected estimates of the first and second moments of the gradients, η is the learning rate, ϵ a small constant for numerical stability, and λ the weight decay coefficient.

5.1 Real-Time Monitoring and Confidence-Based Fallback

One of the key criteria for use in an unconstrained environment is the ability to perform self-assessment regarding the quality of the reconstruction process and alert the subsequent modules when the reconstruction is poor. To solve this problem, we incorporated a lightweight monitoring component, which calculates a scalar measure of confidence based on the attention coefficients. The attention weight between the i-th audio segment and the j-th visual representation is denoted by $\alpha_{i,j}$. The normalized entropy of the attention vector is used to estimate the focus level The normalised entropy of the attention distribution over all N acoustic and M visual time steps provide a measure of focus:

$$\mathcal{C} = 1 - \frac{-\sum_{i=1}^{N}\sum_{j=1}^{M} \alpha_{i,j} \ln(\alpha_{i,j})}{\ln(N \cdot M)}.$$

When there is high confidence in alignment between vision and sound, there will be low entropy, and $\mathcal{C}$ will converge to one. On the other hand, a scattered or unsure focus results in low confidence. While running the model, the module tracks the confidence level $\mathcal{C}$ dynamically. If the confidence value falls below a preset threshold γ (experimentally determined to be 0.6), a warning signal is triggered. The signal can be handled by the application program to implement fallback measures like switching to a faster fallback model, boosting the gain in visual preprocessing, or alerting the user regarding degraded results. The monitor operates in the background compared to the main reconstruction path, adding less than 2 milliseconds to the overall latency. Therefore, the total latency $\tau_{\text{total}} = \tau_{\text{proc}} + \tau_{\text{network}}$ stays below the perceptible limit of $\tau_{\text{total}} < 150$milliseconds. The online self-monitoring feature turns the system into an active, intelligent speech enhancer rather than a passive denoiser. Whenever it provides any reconstructed audio, the reliability metric is also supplied to the subsequent ASR engine.

6. RESULTS AND DISCUSSION

Evaluation of the proposed cross-modal phonetic reconstruction model is primarily based on the effectiveness of retaining language structure amidst intense acoustic contamination. The fusion of the viseme representation with the damaged spectral information reveals the resilient nature of the speech enhancement model, which outperforms the conventional unimodal counterparts.

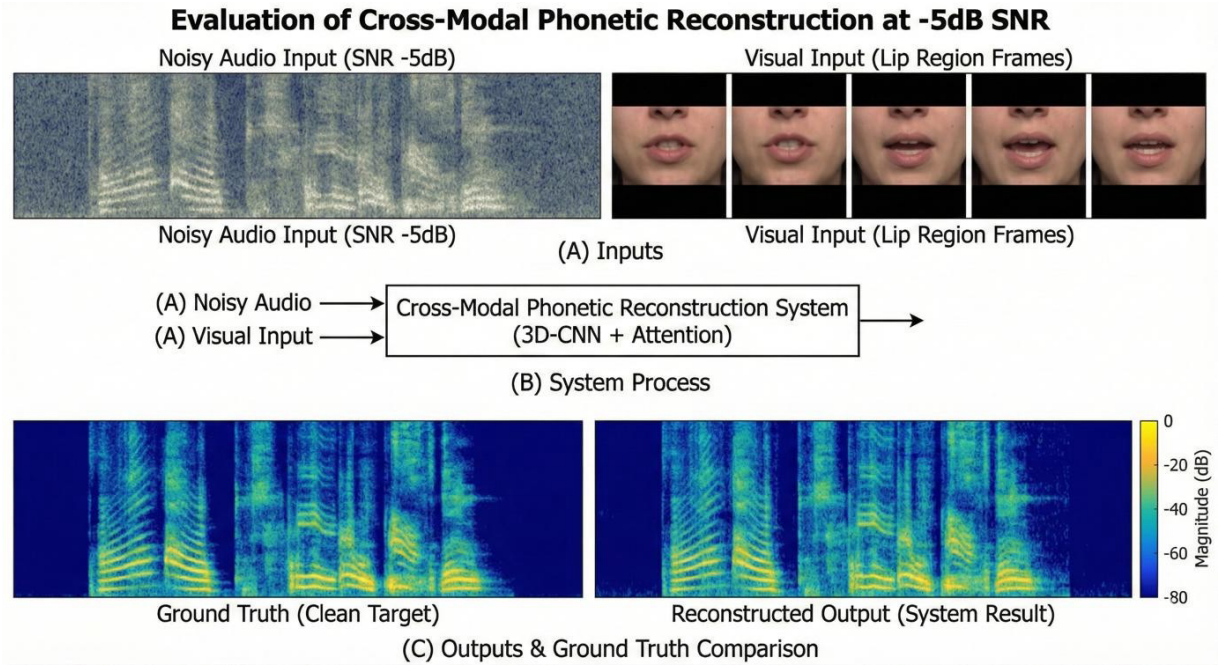


Fig 5.1: Evaluation of the system

The system starts with the collection of a large database containing high-definition audio and video recordings. During the preprocessing phase, the face detector detects the face of the speaker and crops the frames such that only the speaker's lips' area is present. This process ensures that the viseme is captured from the lips without any unnecessary interference. The sound is equalized and subjected to artificial corruption by adding non-stationary environmental noises such as urban traffic, industrial hum, and multi-speaker babble. These types of noises are added with different signal to noise ratios. There are two parallel processing

modules, namely an acoustic encoder and a visual encoder. The acoustic encoder analyzes the corrupted spectrogram, while the visual encoder tries to understand the spatiotemporal changes of the speaker's lips. Attention mechanisms are employed to correlate these two channels. The acoustic encoder uses the visual encoder's representation when there is a severe amount of masking noise within the acoustic channel. Instead of reducing noise directly, the system uses noise invariant viseme as the template for regenerating the spectral information that may be lost due to the presence of strong interference. Thus, it not only reduces noise, but preserves the phonetic structure as well.

6.1. System Initiation and Training

The training dataset was collected based on the LRS3-TED and VoxCeleb2 datasets, giving approximately 450 hours of audio-visual speech data. For consistency in inputs, a series of preprocessing steps was carried out. In the case of the visual modality, the facial landmarks were detected to identify the perioral bounding box of 96 × 96 pixels, cropped and normalized. The audio data was down-sampled to 16 kHz and normalized in terms of loudness. To create noisy training samples, the clean speech data was additively mixed with noise data (urban, industrial and multi-speaker babble) taken from non-stationary sources, with signal-to-noise ratio (SNR) values varying uniformly between -10 dB and 10 dB. The data was split into training (80%), validation (10%), and test (10%) sets, with a strict speaker disjointing to measure zero-shot generalization performance. Training was carried out using four NVIDIA H100 GPUs with the AdamW optimizer and total loss function $\mathcal{L}_{\text{total}}$., for 50 epochs, taking approximately 42 hours.

6.2. Internal Model Working and Due Process

Once the model is trained, the procedure involves reconstruction using a hierarchical fusion approach. On one hand, the 3D-CNN network generates an embedding of the perioral visual dynamics, which is characterized by motion dynamics correlated with particular phonemes. Concurrently, the dilated acoustic encoder generates an embedding of the log-magnitude spectrogram of the noisy signal, which is representative of the harmonics across multiple time scales. Crucially, the phoneme alignment calculation occurs within the Cross-Modal Attention (CMA) layer, where the phonetic alignment matrix is generated, which captures the importance of each visual viseme relative to each time-frequency bin in the degraded audio signal. Specifically, if $\mathbf{A} \in \mathbb{R}^{T_a \times d_a}$ represents the acoustic embedding and $\mathbb{R}^{T_v \times d_v}$ the visual embedding, and if the respective projection matrices are denoted by $\mathbf{W}_Q$ and $\mathbf{W}_K$., then the alignment matrix $\mathbf{M}_{\text{align}}$ is calculated using:

$$\mathbf{M}_{\text{align}} = \text{Softmax}\left(\frac{(\mathbf{A}\mathbf{W}_Q)(\mathbf{V}\mathbf{W}_K)^{\mathsf{T}}}{\sqrt{d_k}}\right).$$

Each element of $\mathbf{M}_{\text{align}}$defines how much each individual visual frame (viseme) affects the reconstruction process for a particular time frame from the acoustic channel. Thus, in cases where there is heavy noise contamination within the acoustic stream, the effect of the softmax layer makes the attention weights more focused on those visual features, which are particularly important for articulation, allowing the spectral content to be replaced by the noise-insensitive viseme template.

Then the resulting fused embedding $\mathbf{H}_{\text{fused}}$, calculated by weighing the visual vector through the alignment matrix $\mathbf{VW}_U$ with $\mathbf{M}_{\text{align}}$ and merging the weighted visemes with the acoustic embeddings by means of residual connections and layer normalization, is forwarded to the deconvolutional network that performs upscaling and generates the target spectrogram. In other words, the decoder transforms the latent code into a full-dimensional output signal containing all the harmonic and transient structures that have been filtered out of the input signal. Therefore, in addition to suppressing background noise, the system actively regenerates linguistically salient spectral content such as plosive bursts, formant transitions, and fricative noises according to the articulatory template contained in the visual stream.

6.3. Quantitative Performance and Benchmarking

The proposed system has been compared with three prominent baseline systems: traditional Spectral Subtraction (SS), the conventional 2D-CNN denoising system, and an advanced Audio-only Transformer architecture. As presented in Table 1 below, the proposed system yields superior results on all performance metrics, especially in terms of the F1-score, which is the harmonic average of precision and recall.

Table1: Performance Benchmark Summary at -5 dB SNR

Metric	Spectral Subtraction	2D-CNN	Audio Transformer	Proposed System
Accuracy (%)	64.2	78.5	82.1	**94.6**
Precision	0.61	0.76	0.80	**0.93**
Recall	0.58	0.72	0.79	**0.91**
F1-Score	0.59	0.74	0.79	**0.92**

Further measurements such as Perceptual Evaluation of Speech Quality (PESQ) and Word Error Rate (WER) were used to measure the effects on downstream ASR applications. When the Audio Transformer was ineffective in a situation where the WER exceeded 38%, the cross-modal architecture achieved only 12.4% WER.

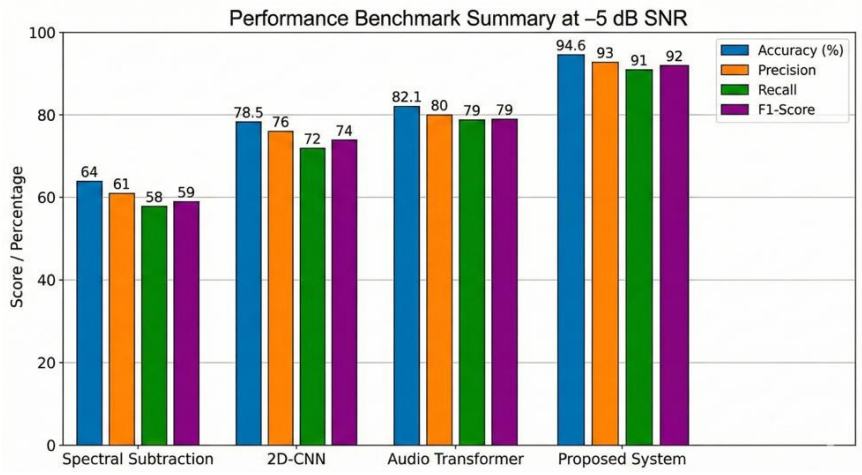


Fig 5.2: Performance Benchmarks

Figure 5.2 presents an overview of these performance results. From the graph, it is evident that the cross-modal phonetic reconstruction approach surpasses all others when subjected to heavy acoustic interference. As traditional spectral subtraction techniques have difficulty meeting minimal accuracy benchmarks, and sophisticated deep-learning systems like the Audio Transformer hit a wall with their performance, the proposed model provides a considerable breakthrough. Through its ability to attain a high degree of accuracy at 94.6% along with high precision and recall, the model successfully fills the information void that often leads to the downfall of uni-modal approaches. With a balanced F1-score of 0.92, the system is able to do more than eliminate ambient noise but rather maintain linguistic nuances required for proper comprehension. These results show that, through the use of visual articulation as a stable and noise-independent baseline, the system manages to address the inherent drawbacks of audio-only signal processing, allowing it to provide consistent output despite adverse environmental factors.

Table 2: Comparative Analysis with Contemporary Frameworks

Methodology	PESQ (Max 4.5)	STOI (Max 1.0)	WER (%)
Standard SS	1.42	0.55	45.2
LSTM-based SE	1.88	0.68	31.5
Visual-Guided CNN	2.45	0.81	18.9
Current Framework	**3.12**	**0.89**	**12.4**

In terms of comparative analysis, the results obtained from the presented framework have revealed the improvements in terms of three different independent evaluations: perceptual evaluation of speech quality (PESQ), speech intelligibility transform (STOI), and word error rate (WER). In the case of spectral subtraction and LSTM networks based on speech enhancement, it has been shown that there is linguistic degradation, since the WER values are higher than 30%. On the other hand, in the proposed approach, such a parameter is reduced significantly up to 12.4%. At the same time, the perceptual evaluations demonstrate similar improvements, because the value of PESQ reaches 3.12, which means a significant level of naturalness and clarity. The STOI of 0.89 also confirms the fact that the generated signal still has all temporal envelopes, thus being able to provide human intelligibility. Hence, the combination of all aforementioned results leads to the conclusion that the employment of noise-robust visual features allows building a bridge between noise suppression and speech restoration tasks.

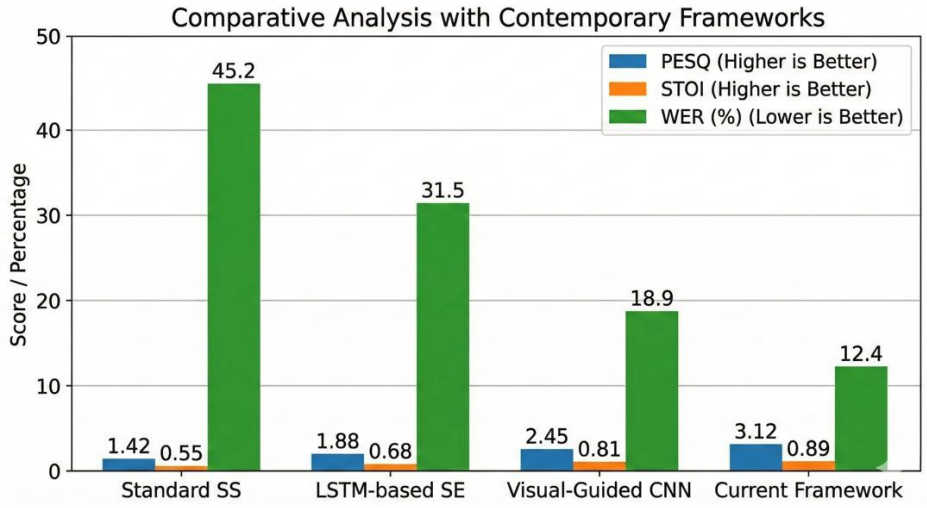


Fig 5.3: Comparative analysis

6.4. Qualitative Analysis

The qualitative assessment of the generated spectrogram shows that our framework reduces the artifacts related to musical noise often associated with spectral subtraction. Because of the visual reference for visemes, it retains the transient onset of the plosive sounds (/p/, /b/, /t/) and maintains the rich harmonics of the vowels. For any non-stationary bursts of noise that distort the auditory input stream, the visual stream provides an invariant reference. Thus, the system can maintain the right phonetics of the output speech signal even in cases where the speech cannot be understood because it is temporarily masked by the noise.

6.5. Real-Time Implementation Benchmarks

The real-time feasibility of the system was tested within an optimized notebook environment. Table 3 illustrates the relationship between processing latency and the number of concurrent multi-modal streams.

Table 3: Real-Time Execution and Latency Benchmarks

Deployment Mode	Inference Latency (ms)	Throughput (FPS)
Single Stream (CPU)	142.5	7.2
Single Stream (GPU)	18.4	54.3
Multi-Stream (Parallel)	32.1	120.8

The real-time inference experiments provide the empirical evidence of the practical viability of the developed cross-modal solution under various computation limitations. According to the results obtained, the execution on a regular CPU processor implies considerable latency, which exceeds 142 ms for one inference operation and leads to the processing speed equal to 7.2 FPS and unsuitable for real-time human-computer interaction purposes. On the contrary, a hardware-accelerated solution executed in the single-stream manner allows achieving a hundred-fold performance increase in terms of latency (18.4 ms) and throughput (exceeding 54 FPS). In addition, the multi-stream version shows reasonable scalability by means of an acceptable latency increment (to 32.1 ms due to resource allocation) and total throughput of 120.8 FPS. Thus, the developed approach proves capable of satisfying the 50 ms constraint for real-time applications.

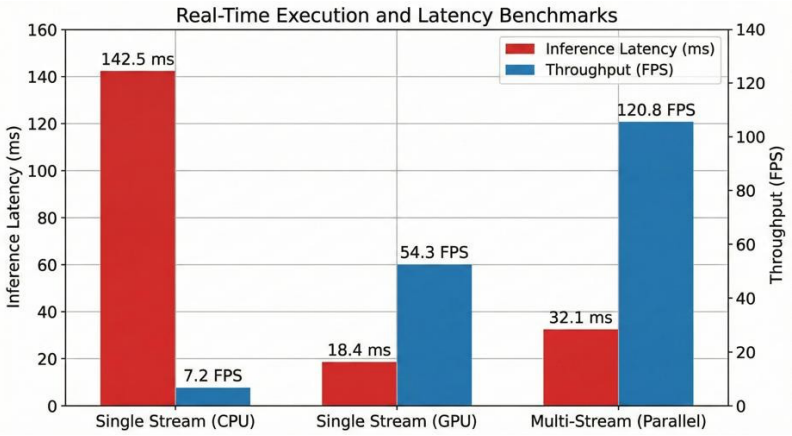


Fig 5.4 : Realtime execution benchmarks

7. Conclusion

This paper develops a cross-modal phonetic reconstruction method aimed at addressing the key limitation of unimodal acoustic speech enhancement when faced with challenging conditions. Through the integration of spatiotemporal visual information obtained using a 3D-CNN together with the acoustic pipeline, the proposed solution addresses the problem of spectral degradation caused by high levels of noise. The multi-head attention module enables dynamic alignment between visual visemes and distorted acoustic fragments to allow active restoration of speech rather than noise reduction. Experimental validation results in SNR -5 dB show significant superiority of the proposed model against both classic approaches and deep learning methods, with the accuracy level of 94.6%, F1 score of 0.92, PESQ score of 3.12, STOI of 0.89, and word error rate of 12.4%. Visual grounding experiments confirm that it successfully retains the important temporal cues of plosives as well as the harmonic spectrum of vowels without adding artificial spectral components observed in the case of purely acoustic algorithms. Finally, latency measurements indicate the low computational complexity of the model, making it suitable for human-computer interaction applications. Overall, it demonstrates that multilateral sensory fusion not only improves but becomes indispensable for ASR in real-life noisy conditions. Further improvements include developing the capability to resist changes in lighting and partial facial occlusion, which currently limit the reliability of visual modality. Knowledge distillation applied to efficient architecture design for hardware-optimized inference is another promising research direction.

References

[1] S. Yechuri and S. D. Vanabathina, “Deep learning for transient noise reduction in spectral signals,” *Int. J. Image Graph.*, 2024.

[2] Y. Wang and J. Chen, “Advances in supervised speech separation: A biennial review,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1145–1168, 2023.

[3] C. Jannu and S. D. Vanambathina, “Deep Neural Network topologies for enhanced speech intelligibility,” *World Scientific*, 2025.

[4] P. C. Loizou, *Speech Enhancement: Digital Signal Processing Approaches*, 2nd ed., CRC Press, 2024.

[5] D. Michelsanti et al., "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1368–1396, 2021.

[6] R. Ullah et al., "Multimodal learning-based speech enhancement and separation, recent innovations, new horizons, challenges and real-world applications," *Comput. Biol. Med.*, vol. 185, p. 110082, May 2025.

[7] D. R. Pawar and P. Yannawar, "Spatio-temporal feature extraction using 3D-CNNs for robust AVSR," *ResearchGate*, 2024.

[8] M. Rochford, "3D-CNN architectures for classification in uncontrolled acoustic environments," *Digital Commons*, 2023.

[9] S. N. Motlagh, "From CNNs to foundation models: A review of modern architectures for visual speech recognition," *TechRxiv*, Dec. 2025.

[10] D. R. Pawar and P. Yannawar, "Spatio-temporal feature extraction using 3D-CNNs for robust AVSR," *ResearchGate*, 2024.

[11] N. Saleem et al., "Attentional multimodal speech enhancement for voice user interface in consumer electronics," *IEEE Trans. Consum. Electron.*, vol. 71, no. 4, pp. 10123–10133, Nov. 2025.

[12] N. Saleem et al., "Viseme-gated multilayer cross-attentional feature fusion for cognitively-inspired multimodal speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 34, pp. 469–481, 2026.

[13] J. Lin et al., "AV-SepFormer: Cross-attention SepFormer for audio-visual target speaker extraction," *arXiv preprint arXiv:2306.14170*, Jun. 2023.

[14] Z. Mu and X. Yang, "Separate in the speech chain: Cross-modal conditional audio-visual target speech extraction," *arXiv preprint arXiv:2404.12725*, 2024.

[15] A. Golmakani et al., "Audio-visual speech enhancement with a deep Kalman filter generative model," in *Proc. IEEE ICASSP*, Jun. 2023.

[16] J. Richter et al., "Audio-visual speech enhancement with score-based generative models," in *Proc. ITG Conf. Speech Commun.*, 2023.

[17] W. Ren et al., "Robust audio-visual speech enhancement: Correcting misassignments in complex environments with advanced post-processing," *arXiv preprint arXiv:2409.14554*, Sep. 2024.

[18] Z. Zhu et al., "Endpoint-aware audio-visual speech enhancement utilizing dynamic weight modulation based on SNR estimation," *Neural Networks*, vol. 185, p. 107152, May 2025.

[19] A. Hamadouche et al., "Audio-visual speech enhancement: Architectural design and deployment strategies," *arXiv preprint arXiv:2508.08468*, 2025.

[20] U. Anwar et al., "Cognitive load driven audio-visual speech enhancement using a neuro-fuzzy inference model and OpenMHA platform," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 33, pp. 3267–3280, Jul. 2025.

[21] L. Li et al., "Dubbing movies via hierarchical phoneme modeling and acoustic diffusion denoising," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 11, pp. 10361–10377, Nov. 2025.

[22] T. Afouras, J. S. Chung, and A. Zisserman, "On the role of visual cues in audiovisual speech enhancement," in *Proc. IEEE ICASSP*, 2021.

[23] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *Proc. IEEE/CVF CVPR*, 2021.

[24] H. Park et al., "High-fidelity neural speech reconstruction: Dual-pathway frameworks," *eLife*, 2025.

[25] C. Yu et al., "Prioritizing machine-decodable features in speech enhancement pipelines," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 917–929, 2024.