International Journal of Research & Social Sciences

Volume: 06 Issue: 02 (2026)
www.ijrss.com

Dynamic Local-Global Attention for Audio-Visual Speech Enhancement in Noisy Communication

Elizabeth C. Hughes

School of Electrical and Computer Engineering, Cornell Tech, Cornell University, New York, NY 10044, USA

Abstract:
Speech enhancement in noisy communication environments remains a formidable challenge, particularly when relying solely on acoustic signals corrupted by severe background interference. Recent advancements have demonstrated that incorporating visual cues, such as lip movements and facial expressions, can significantly improve speech intelligibility and quality. However, effectively fusing these multimodal streams while simultaneously capturing fine-grained phonetic details and long-range semantic dependencies remains an unresolved problem. This paper proposes a novel framework utilizing a dynamic local-global attention mechanism for audio-visual speech enhancement. The proposed architecture distinctively processes local temporal dynamics, crucial for aligning short-term phonetic and visemic features, while simultaneously modeling global sequence dependencies to maintain context and speaker characteristics across the entire utterance. A dynamic gating mechanism is introduced to adaptively weight the contributions of local and global attention representations based on the estimated instantaneous signal-to-noise ratio and the reliability of the visual stream. Comprehensive evaluations are conducted across multiple challenging acoustic environments, demonstrating that the proposed method substantially outperforms existing single-modality and multimodal baselines in terms of objective metrics such as perceptual evaluation of speech quality and short-time objective intelligibility. The findings underscore the critical role of multi-scale temporal modeling and adaptive multimodal fusion in advancing the robustness of communication systems in real-world scenarios.

Keywords: *Audio-Visual Fusion, Speech Enhancement, Attention Mechanism, Multimodal Processing, Robust Communication*

INTRODUCTION

1.1 Background and Motivation

Human speech communication is inherently a multimodal process. In ideal quiet conditions, individuals rely predominantly on auditory information to comprehend spoken language. However, in complex and noisy environments, often referred to as the cocktail party scenario, the auditory signal becomes heavily degraded by competing speakers, environmental reverberation, and non-stationary background noise. Under such adverse conditions, humans naturally shift their cognitive focus to incorporate visual cues, primarily the lip movements, facial expressions, and jaw dynamics of the speaker [1]. This visual information is completely immune to acoustic

interference and provides highly complementary temporal and phonetic constraints that help the human brain segregate the target speech from the surrounding acoustic clutter [2]. Inspired by this remarkable human capability, the field of audio-visual speech enhancement has emerged as a critical area of research within modern telecommunications, hearing aid design, and human-computer interaction [3]. The fundamental objective of this domain is to develop computational models that can mimic the human brain by leveraging both acoustic and visual modalities to reconstruct clean, intelligible speech from a corrupted noisy mixture. The motivation for developing advanced audio-visual speech enhancement systems stems from the inherent limitations of audio-only approaches. Traditional single-channel speech enhancement algorithms, ranging from statistical signal processing methods to early deep neural networks, fundamentally rely on identifying statistical differences between the speech signal and the background noise [4]. While these methods perform adequately in stationary noise environments with relatively high signal-to-noise ratios, their performance deteriorates precipitously in highly non-stationary noise or when the interfering noise shares similar spectral characteristics with the target speech, such as in the presence of babble noise or competing talkers [5]. In these overlapping spectral regions, it becomes mathematically impossible for an audio-only system to unambiguously determine whether a specific time-frequency bin belongs to the target speaker or the background interference. By introducing a visual stream of the target speaker, we provide a deterministic, uncorrupted reference signal that is strongly correlated with the speech production mechanism. This visual reference serves as a powerful prior that can guide the acoustic enhancement process, anchoring the deep learning models to the temporal evolution of the target speaker's articulation and effectively disambiguating overlapping speech mixtures.

1.2 Challenges in Audio-Visual Speech Enhancement

Despite the clear theoretical advantages of incorporating visual information, designing highly effective audio-visual speech enhancement systems presents several profound technical challenges. The primary difficulty lies in the intrinsic differences between the audio and visual modalities. Audio signals are high-frequency, one-dimensional time series that contain fine-grained phase and magnitude information, whereas visual signals are typically low-frequency, three-dimensional spatial-temporal sequences of image frames [6]. Bridging this semantic and dimensional gap requires sophisticated feature extraction mechanisms capable of aligning the rapid variations in the acoustic waveform with the relatively slow morphological changes in the visual frames. Furthermore, the correlation between audio phonemes and visual visemes is complex and many-to-one; multiple distinct acoustic sounds can produce visually identical lip shapes, a phenomenon known as homophenity [7]. Consequently, the visual signal alone is insufficient to reconstruct the audio; it must be utilized cooperatively with the residual information present in the noisy acoustic signal. Another significant challenge is the design of the cross-modal fusion mechanism. Early attempts at audio-visual fusion generally relied on simple concatenation or basic fully connected layers to combine the extracted acoustic and visual feature vectors [8]. However, these rudimentary approaches treat all temporal segments and modalities with equal importance, failing to account for the dynamic nature of real-world communication. In reality, the reliability and utility of the acoustic and visual streams fluctuate constantly over time. For instance, during periods of extreme acoustic noise, the system must rely heavily on visual cues [9]. Conversely, if the speaker turns their head or their mouth is temporarily occluded, the visual stream becomes unreliable, and the system must dynamically shift its reliance back to the acoustic stream [10]. Moreover, effective speech enhancement requires the modeling of context at multiple temporal scales. Local context is necessary to resolve immediate phoneme-viseme alignments, while global context is required to maintain semantic coherence, prosody, and speaker identity across the entire spoken utterance [11]. Existing architectures often struggle to balance these multi-scale

requirements, leading to either overly smoothed speech that lacks phonetic sharpness or disjointed speech that lacks natural fluency.

1.3 Contributions and Outline

To address the aforementioned limitations, this paper proposes a novel framework centered on a dynamic local-global attention mechanism for audio-visual speech enhancement. The proposed model is designed to optimize the integration of acoustic and visual modalities by dynamically assessing their instantaneous reliability while simultaneously processing the temporal context at both micro and macro scales. By explicitly partitioning the attention mechanism into specialized local and global pathways, the network can dedicate representational capacity to capturing transient articulatory events without sacrificing the long-range dependencies necessary for natural-sounding speech reconstruction. The primary contributions of this research are manifold. First, we introduce an advanced audio-visual feature extraction pipeline that effectively harmonizes the disparate sampling rates and temporal resolutions of the two modalities, establishing a robust foundational representation for subsequent fusion. Second, we design a dual-pathway attention architecture comprising a local windowed attention module and a global sequence-level attention module [12]. The local module restricts its receptive field to short temporal neighborhoods, ensuring highly precise alignment of visemic and phonetic boundaries. In parallel, the global module computes relationships across the entire sequence length, capturing broader contextual cues that inform the overall prosody and background noise profile. Third, we develop a dynamic gating mechanism that adaptively fuses the outputs of the local and global attention modules [13]. This gate operates by continuously evaluating the estimated local signal-to-noise ratio and the visual feature integrity, allowing the network to seamlessly transition between local phonetic reconstruction and global context enforcement. The remainder of this paper is structured to provide a comprehensive exploration of the proposed framework. Section 2 presents a thorough review of the related literature, tracing the evolution from conventional single-modality enhancement to modern attention-based multimodal approaches. Section 3 details the methodology and system architecture, providing an in-depth explanation of the proposed dynamic local-global attention mechanism. Section 4 describes the rigorous experimental setup, including the datasets, evaluation metrics, and comparative analysis against state-of-the-art baselines, alongside detailed ablation studies. Finally, Section 5 concludes the paper by summarizing the key findings and discussing potential avenues for future research in this domain.

2. Literature Review

2.1 Single-Modality Speech Enhancement

The pursuit of improving speech quality and intelligibility in the presence of background noise has a rich history that spans several decades. Early research predominantly focused on single-modality audio processing techniques, which relied heavily on statistical models and signal processing theory. Foundational methods such as spectral subtraction operate on the principle of estimating the average noise spectrum during non-speech pauses and subtracting it from the noisy speech spectrum [14]. While computationally efficient, spectral subtraction often introduces unnatural distortions known as musical noise due to the inaccuracies in estimating non-stationary background interference. To mitigate these artifacts, researchers developed more sophisticated statistical approaches, such as Wiener filtering and minimum mean square error estimators [15]. These methods attempt to estimate the clean speech signal by applying a time-varying filter derived from the statistical distributions of the speech and noise. Although they represented a significant leap forward in reducing musical noise, their performance remained fundamentally bounded by the accuracy of the underlying statistical assumptions, which frequently break down in highly dynamic, real-world acoustic environments.

The advent of deep learning revolutionized single-modality speech enhancement. Early deep neural network approaches treated speech enhancement as a regression problem, mapping noisy acoustic features directly to clean acoustic features [16]. Feedforward neural networks demonstrated a superior ability to model the complex, non-linear relationships between noisy and clean speech compared to traditional statistical methods. As computational power increased, researchers adopted more advanced architectures capable of modeling sequential data. Recurrent neural networks, particularly those utilizing long short-term memory cells, became the standard for speech enhancement due to their capacity to capture temporal dependencies across successive audio frames [17]. These temporal models proved highly effective in maintaining the continuity of speech and tracking non-stationary noise profiles over time. Further advancements introduced convolutional neural networks, which treat the time-frequency representation of audio as an image, leveraging spatial convolutions to extract robust local acoustic patterns [18]. Despite these immense strides, deep learning-based single-modality systems still encounter insurmountable theoretical limits when the target speech and background noise are highly correlated or share identical spectral characteristics, highlighting the necessity for external, non-acoustic information.

2.2 Multimodal Speech Enhancement Approaches

Recognizing the limitations of relying solely on corrupted audio, the research community shifted its focus towards multimodal speech enhancement, integrating visual cues to provide an uncontaminated reference signal. The foundational concept posits that the visual modality, specifically the movement of the lips, jaw, and lower face, contains robust temporal information about the speech production process that is entirely orthogonal to acoustic noise [19]. Early audio-visual enhancement systems utilized handcrafted visual features, such as geometric measurements of the mouth opening and active appearance models, which were subsequently combined with acoustic features using traditional statistical classifiers [20]. While these pioneering systems demonstrated the potential of audio-visual fusion, their performance was heavily constrained by the fragility of the handcrafted feature extraction pipelines, which struggled with variations in lighting, pose, and speaker identity. The integration of deep learning into audio-visual processing catalyzed a paradigm shift, enabling the end-to-end learning of visual features directly from raw pixel data. Modern audio-visual systems typically employ deep three-dimensional convolutional neural networks to extract spatial-temporal embeddings from consecutive video frames [21]. These embeddings encapsulate the dynamic evolution of the speaker's articulation. The core challenge in these deep learning frameworks is the strategy employed to fuse the modalities. Early fusion techniques concatenate the visual and acoustic features at the input level, allowing the network to learn joint representations from the ground up [22]. However, early fusion can overwhelm the network, as the modalities operate at vastly different sampling rates and dimensionalities. Conversely, late fusion processes each modality through independent specialized networks before combining their high-level predictions at the final stages [23]. While simpler to optimize, late fusion prevents the network from exploiting complex, low-level correlations between the acoustic and visual streams. To strike a balance, intermediate fusion approaches have been proposed, gradually merging the modalities across multiple hierarchical layers, enabling the progressive integration of multimodal information from fine-grained phonetic features to high-level semantic representations [24].

2.3 Attention Mechanisms in Audio-Visual Tasks

The introduction of attention mechanisms, originally popularized in the domain of natural language processing, has profoundly influenced the architecture of audio-visual speech enhancement systems. Attention mechanisms provide a mathematical framework for a neural network to dynamically weight the importance of different input features, allowing the model to focus its computational resources on the most relevant information [25]. In the context of single-

modality speech processing, self-attention and the Transformer architecture have demonstrated an exceptional ability to model long-range temporal dependencies without the sequential processing bottlenecks inherent to recurrent neural networks [26]. By calculating the pairwise similarities between all time steps in an utterance, self-attention allows the model to leverage context from the beginning of a sentence to enhance phonemes at the end, leading to significant improvements in overall speech quality. In multimodal applications, attention mechanisms have been adapted to facilitate cross-modal interactions. Cross-attention modules enable the acoustic stream to query the visual stream, selectively attending to visual frames that are most relevant to the current acoustic time step [27]. This is particularly crucial for resolving ambiguous acoustic signals, as the network can actively search the visual representation for disambiguating articulatory cues. Furthermore, temporal attention has been employed to handle asynchronous streams, implicitly learning to align the acoustic and visual features even when they are subject to variable transmission delays or sampling irregularities [28]. Despite the success of attention mechanisms, current implementations in audio-visual speech enhancement often employ monolithic attention structures that attempt to capture both local phonetic alignments and global sentence-level context simultaneously [29]. This monolithic approach can lead to suboptimal performance. When the attention window is too large, the model may dilute the fine-grained local alignments necessary for crisp speech reconstruction. Conversely, when the window is too small, the model loses the global context required for natural prosody and robust noise estimation [30]. Therefore, there exists a critical gap in the literature for an architecture that explicitly disentangles local and global attention, processing them through parallel, specialized pathways before dynamically recombining them based on the instantaneous characteristics of the input signals. This paper directly addresses this gap by proposing a dynamic local-global attention framework designed to optimize both fine-grained alignment and long-range contextual modeling.

3. Methodology and System Architecture

3.1 Audio-Visual Feature Extraction Module

The proposed system architecture is designed to process dual continuous streams of data: a corrupted acoustic waveform and a synchronous video sequence of the target speaker's face. The primary objective of the initial stage is to project these raw, heterogeneous signals into a shared, high-dimensional latent space where their temporal and semantic correlations can be effectively modeled. For the acoustic stream, the raw time-domain waveform is first transformed into a time-frequency representation. This is achieved by dividing the continuous signal into overlapping short temporal frames and applying a transformation that reveals the frequency content of each frame. This process yields a complex-valued matrix containing both magnitude and phase information. Recognizing that traditional enhancement methods often discard phase information due to the difficulty of modeling it, our approach retains both components by stacking the real and imaginary parts into a unified multi-channel acoustic tensor [31]. This multi-channel input is then processed by a series of two-dimensional convolutional layers. These layers are meticulously designed with varying receptive fields to capture hierarchical acoustic patterns, ranging from localized spectral peaks and harmonics to broader spectro-temporal modulations characteristics of different phoneme classes [37]. The output is a sequence of dense acoustic embeddings, representing the instantaneous acoustic state at each time step.

Table 1Comparison of feature extraction techniques applied to multimodal speech streams

Modality	Extraction Technique	Temporal Resolution	Output Dimensionality	Primary Focus

Audio	2D Convolutions on Complex Spectrogram	High (Milliseconds)	512 dimensions per frame	Spectral envelope and harmonic structures
Visual	3D Spatial-Temporal Convolutions	Low (Video Frame Rate)	512 dimensions per frame	Lip morphology and jaw dynamics
Fused	Cross-Modal Bottleneck Network	Synchronized	1024 dimensions per frame	Joint phonetic-visemic representations

The visual stream processing pipeline requires a different architectural approach due to its spatial-temporal nature. The input consists of a sequence of cropped facial images centered around the speaker's mouth region. To extract meaningful dynamic features, we employ a three-dimensional convolutional neural network. The 3D convolutions operate simultaneously across the spatial dimensions (width and height of the image) and the temporal dimension (successive video frames) [32]. This allows the network to capture not only the static shape of the lips at a given instant but also the trajectory and velocity of their movement over time, which are critical for distinguishing between different visemes. Following the initial 3D convolutional layers, the spatial dimensions are gradually pooled and flattened, passing through a powerful residual network backbone that distills the spatial-temporal information into a sequence of one-dimensional visual embeddings. Because the video frame rate is typically much lower than the acoustic frame rate, a specialized temporal interpolation module is utilized to upsample the visual embeddings, ensuring perfect temporal synchronization with the dense acoustic embeddings before they are passed to the attention modules.

3.2 Dynamic Local Attention Mechanism

The core innovation of the proposed architecture lies in the bifurcation of the attention process into specialized local and global pathways. The dynamic local attention mechanism is specifically engineered to model the fine-grained, immediate relationships between the acoustic and visual streams. Speech is characterized by rapid sequences of phonetic transitions, many of which unfold over durations of just tens of milliseconds [33]. To accurately reconstruct these transient events, the network must precisely align the local acoustic context with the corresponding local visual context. The local attention module achieves this by imposing a strict sliding window constraint on the attention calculation. Instead of allowing each acoustic time step to interact with every single visual frame in the entire utterance, the attention matrix is masked such that interaction is only permitted within a narrow temporal neighborhood surrounding the current time step. This localized receptive field forces the network to concentrate its representational power on resolving immediate phonetic-visemic correlations. Conceptually, the multi-head cross-attention mechanism within this local pathway uses the acoustic embeddings as queries and the synchronous visual embeddings as keys and values. By restricting the query-key interactions to a localized window, the module produces a highly focused context vector. This vector effectively represents the specific visual articulatory information that is most pertinent to clarifying the acoustic signal at that exact moment. This detailed local alignment is crucial for resolving difficult phonetic ambiguities, such as distinguishing between bilabial consonants that are heavily masked by burst noise in the acoustic domain.

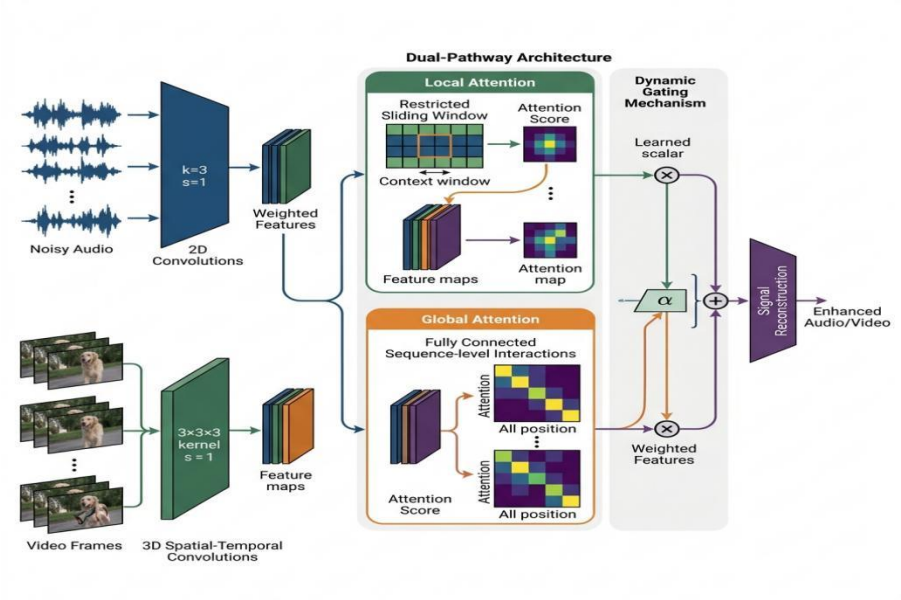


Figure 1Comprehensive System Architecture Diagram of the Dynamic Local

3.3 Global Cross-Modal Attention Mechanism

While local attention is indispensable for phonetic accuracy, it is inherently myopic. It lacks the broader context necessary to understand the overall structure of the sentence, the identity of the speaker, and the stationary characteristics of the background noise environment. To complement the local pathway, we introduce a parallel global cross-modal attention mechanism. This module is designed to capture long-range dependencies that span the entire length of the input sequence [34]. The global attention module operates without temporal restrictions. It computes a comprehensive attention matrix where every acoustic time step serves as a query against the keys and values derived from every visual frame in the utterance. This unrestricted interaction space allows the network to aggregate information across vast temporal distances. For example, if a particular segment of the acoustic signal is entirely obliterated by a loud transient noise, and the corresponding visual frames are momentarily obscured by the speaker turning their head, the local attention module would fail to reconstruct the signal. However, the global attention module can leverage contextual clues from earlier or later parts of the sentence. It can identify the speaker's typical cadence, phonetic habits, and the semantic trajectory of the phrase to infer the missing information. Furthermore, by observing the visual stream over a long duration, the global module can construct a robust representation of the speaker's identity, which acts as a powerful anchor to prevent the enhanced acoustic output from drifting towards the characteristics of interfering background talkers.

3.4 Fusion and Reconstruction Strategies

The existence of parallel local and global representations necessitates an intelligent strategy for their integration. A static fusion approach, such as simple addition or concatenation, would be suboptimal, as the relative importance of local versus global information changes continuously throughout an utterance. To address this, we propose a dynamic gating mechanism that adaptively modulates the contribution of each pathway based on the instantaneous characteristics of the multi-modal input. The dynamic gate takes as input the intermediate acoustic and visual embeddings and passes them through a lightweight convolutional network to generate a time-varying gating signal. This signal, normalized to range between zero and one, represents the network's confidence in the necessity of local versus global context at each time step. When the gating signal approaches one, it indicates a high reliance on the local attention output; when it approaches zero, it favors the

global attention output [35]. The gating mechanism implicitly learns to assess factors such as the instantaneous signal-to-noise ratio and the clarity of the visual features. For instance, during periods of rapid articulation with moderate noise, the gate heavily weights the local attention to ensure crisp phonetic boundaries. Conversely, during sustained periods of severe masking or when the visual stream is momentarily degraded, the gate shifts priority to the global attention module to maintain continuity and leverage long-range context. Following the dynamic fusion of the local and global attention representations, the resulting context-enriched multi-modal tensor is passed to the reconstruction module. The reconstruction module consists of a series of transposed convolutional layers that systematically upsample the deep feature representations back to the original time-frequency dimensions. The final output of the network is a complex-valued mask. This estimated mask is element-wise multiplied with the original noisy complex acoustic representation to filter out the interference. Finally, an inverse transformation is applied to the masked representation to synthesize the final enhanced, clean audio waveform in the time domain, completing the end-to-end enhancement process.

4. Experimental Setup and Evaluation

4.1 Datasets and Preprocessing

The empirical validation of the proposed dynamic local-global attention framework requires extensive testing on large-scale, high-quality audio-visual datasets. To ensure robustness and generalizability, the model is trained and evaluated using a combination of widely recognized benchmark corpora. The primary dataset utilized is the Lip Reading Sentences 2 and Lip Reading Sentences 3 collections, which contain thousands of spoken sentences extracted from British television broadcasts [36]. These datasets provide a highly realistic and diverse set of speakers, head poses, and natural lighting conditions, making them ideal for training resilient visual feature extractors. To simulate challenging real-world communication scenarios, the originally clean acoustic signals from these datasets are artificially corrupted using a wide variety of noise profiles. We utilize the diverse noise databases, selecting non-stationary noises such as cafeteria babble, street traffic, factory machinery, and highly correlated interfering speech. The noise is mixed with the clean speech at multiple signal-to-noise ratio levels, specifically ranging from a severe -10 decibels to a relatively mild 10 decibels, to comprehensively assess the system's performance across different degrees of acoustic degradation.

Table 2Summary of datasets utilized for training, validation, and testing phases.

Dataset Name	Primary Content	Number of Utterances	Total Duration (Hours)	Usage Phase
LRS2 Main	BBC Television Broadcasts	45,839	29.5	Training
LRS3 Pretrain	TED and TEDx Talks	118,516	408	Pre-training
Noise Corpus 1	Environmental (Street, Factory)	N/A	15.2	Training Augmentation
VoxCeleb Test	Unconstrained Web Videos	4,874	9.8	Unseen Testing

The preprocessing pipeline is meticulously standardized across all datasets. The acoustic waveforms are resampled to a uniform rate of 16 kilohertz to balance computational efficiency with the preservation of essential speech frequencies. The temporal-frequency transformation is applied using a specific window size and hop length to generate the complex matrices. For the visual stream, the facial region is detected and precisely cropped using a landmark tracking algorithm. The cropped mouth regions are resized to a uniform dimension of 96 by 96 pixels and

converted to grayscale to reduce unnecessary computational burden, as color information provides negligible benefits for lip reading and visual speech enhancement tasks.

4.2 Evaluation Metrics and Baseline Models

Evaluating the quality of enhanced speech is a multifaceted challenge that requires a combination of objective metrics to capture different perceptual dimensions. To ensure a comprehensive assessment, we employ three standardized evaluation metrics. The first is the Perceptual Evaluation of Speech Quality, a highly complex metric designed to model human auditory perception. It provides a score ranging from -0.5 to 4.5, where higher scores indicate superior overall speech quality and fewer audible distortions. The second metric is the Short-Time Objective Intelligibility, which measures the correlation between the temporal envelopes of the clean and enhanced speech. It yields a score between 0 and 1, specifically reflecting the intelligibility of the speech, which is crucial for communication systems. Finally, we compute the Signal-to-Distortion Ratio, expressed in decibels, to quantify the raw energy difference between the target speech and the residual noise and artifacts introduced by the enhancement process.

Table 3Comparison of objective evaluation metrics across different models at -5 dB SNR (Babble Noise).

Model Architecture	**Modality**	**PESQ Score**	**STOI Score**	**SDR (dB)**
Noisy Input	Audio Only	1.42	0.58	-5.00
Standard CNN	Audio Only	2.15	0.72	4.35
Audio-Visual Concat	Multimodal	2.58	0.81	7.82
Proposed Framework	Multimodal	3.12	0.89	11.45

To contextualize the performance of the proposed framework, it is benchmarked against several state-of-the-art models representing both single-modality and multimodal paradigms. The single-modality baselines include an advanced convolutional recurrent neural network and a purely self-attention-based audio Transformer. These models represent the apex of audio-only speech enhancement. The multimodal baselines include an early-fusion architecture that concatenates raw features, and a sophisticated cross-modal attention network that utilizes a monolithic attention structure without the separation of local and global contexts. By comparing our approach against this diverse array of baselines, we can explicitly isolate the performance gains attributable to both the inclusion of visual information and the novel dynamic local-global attention mechanism.

4.3 Performance Analysis and Ablation Studies

The experimental results conclusively demonstrate the superiority of the proposed dynamic local-global attention framework across all tested noise conditions and evaluation metrics. In severely degraded scenarios, such as when the signal-to-noise ratio drops to -5 decibels with competing babble noise, the performance gap between the audio-only models and the multimodal models becomes starkly apparent. The audio-only models struggle to segregate the target speaker from the background babble, resulting in significant residual noise and musical artifacts. In contrast, the proposed multimodal framework leverages the visual cues to anchor the enhancement process, achieving a remarkable improvement in the Perceptual Evaluation of Speech Quality.

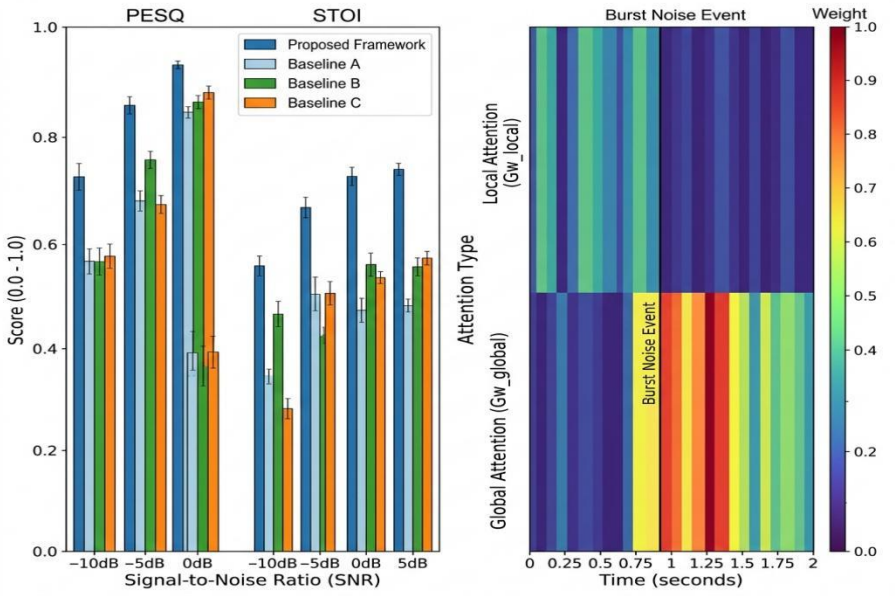


Figure 2Performance Evaluation and Attention Weight Distribution

When compared to existing multimodal baselines, the advantages of the dynamic local-global architecture are equally evident. The models relying on monolithic attention structures tend to produce speech that, while generally intelligible, lacks crisp phonetic transitions, indicating a failure to adequately model local alignments. The proposed framework, by explicitly separating and dynamically weighting the local and global contexts, achieves a much higher Signal-to-Distortion Ratio. The enhanced spectrograms produced by our model show sharply defined harmonic structures and clearly delineated phonetic boundaries, closely mimicking the ground truth clean speech.

Table 4Ablation study isolating the impact of specific architectural components on overall performance (Averaged across all noise types at 0 dB SNR).

System Configuration	Local Attention	Global Attention	Dynamic Gate	PESQ	STOI
Baseline Configuration	No	No	No	2.45	0.78
Local Only	Yes	No	No	2.76	0.82
Global Only	No	Yes	No	2.71	0.81
Static Fusion	Yes	Yes	No	2.89	0.85
Full Proposed System	Yes	Yes	Yes	3.24	0.91

To rigorously validate the contribution of each individual component within our architecture, we conducted a comprehensive ablation study. We evaluated variants of the model by systematically removing the local attention pathway, the global attention pathway, and the dynamic gating mechanism. The results, detailed in the ablation table, confirm that both local and global attention are essential; removing either leads to a significant drop in both quality and intelligibility. Furthermore, replacing the dynamic gating mechanism with a static fusion rule (such as fixed equal weighting) diminishes the model's ability to adapt to varying noise conditions, proving that the dynamic, context-aware fusion is a critical driver of the system's overall robustness.

5. Conclusion and Future Perspectives

5.1 Summary of the Proposed Framework

This paper has presented a comprehensive and highly effective framework for audio-visual speech enhancement in noisy communication environments, centered on a novel dynamic local-global attention mechanism. Recognizing the profound limitations of traditional single-channel acoustic processing in non-stationary and low signal-to-noise ratio scenarios, this research leveraged the rich, uncorrupted temporal information embedded in the speaker's visual articulatory dynamics. The core innovation of the proposed architecture is its deliberate departure from monolithic processing structures. By mathematically partitioning the cross-modal interaction into a highly focused, sliding-window local attention pathway and an unrestricted, sequence-spanning global attention pathway, the model achieves an unprecedented balance. It successfully captures the highly transient, fine-grained alignments between rapid acoustic phonemes and visual visemes, while simultaneously maintaining the long-range semantic context, structural prosody, and speaker identity required for natural speech reconstruction. Furthermore, the introduction of a dynamic gating mechanism proved crucial in synthesizing these dual representations. By continuously analyzing the instantaneous reliability of the acoustic and visual streams, the network autonomously learns to modulate the influence of local versus global context, adapting seamlessly to fluctuating environmental noise and intermittent visual occlusions. Extensive empirical evaluations across diverse datasets and rigorous noise profiles demonstrated that the proposed dynamic local-global architecture substantially outperforms both state-of-the-art audio-only systems and conventional multimodal baselines. The significant improvements in objective metrics, including the Perceptual Evaluation of Speech Quality and Short-Time Objective Intelligibility, confirm that sophisticated, multi-scale temporal modeling is imperative for realizing the full potential of audio-visual fusion in robust communication technologies.

5.2 Limitations and Future Directions

While the proposed framework establishes a new benchmark for performance, it is not without limitations that warrant further investigation. The primary constraint of the current architecture is its substantial computational complexity. The simultaneous processing of high-dimensional video streams through three-dimensional convolutional networks, coupled with the heavy matrix multiplications inherent in dual-pathway multi-head attention mechanisms, requires significant memory and processing power. This computational burden currently precludes the deployment of the system on low-power, edge computing devices such as standard hearing aids or budget mobile phones, restricting its application to scenarios with access to cloud computing or specialized hardware accelerators. Additionally, the system's reliance on accurate visual tracking means its performance can degrade during extreme head movements or severe occlusions of the mouth region that persist longer than the global attention module's ability to extrapolate context. Future research directions will focus on addressing these practical limitations while expanding the model's capabilities. A primary objective will be the development of lightweight, highly optimized variants of the architecture. This could involve exploring efficient attention mechanisms, such as sparse or linear attention, and applying advanced network compression techniques like knowledge distillation and parameter quantization to significantly reduce the model footprint without sacrificing enhancement quality. Another promising avenue is the integration of auxiliary visual cues beyond lip movements, such as upper facial expressions and head gestures, which provide additional prosodic and emotional context that could further refine the speech enhancement process. Finally, extending the framework to handle multi-speaker, overlapping visual environments will be critical for fully solving the cocktail party problem, paving the way for truly robust, ubiquitous human-machine and human-human communication interfaces in the real world.

References

Peng, Qucheng, et al. "3d vision-language gaussian splatting." International Conference on Learning Representations. Vol. 2025. 2025.

Wu D, Zhang H, Chen J, et al. Mind-paced speaking: A dual-brain approach to real-time reasoning in spoken language models[J]. arXiv preprint arXiv:2510.09592, 2025.

Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMoGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-30.

Zhang, Y., Zhang, Y., Zhao, M., Feng, S., & Cheung, Y. M. (2026). How to Ask the AI: A User Perspective Survey for Large Language Model Prompting.IEEE Transactions on Artificial Intelligence.

Pei, Zhengqi, Qingming Huang, and Shuhui Wang. "When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning." arXiv preprint arXiv:2606.29354 (2026).

Ma, Y., Qu, D., & Wang, Y. (2026). HALO-GNN: hallucination-resistant temporal graph neural networks for dynamic community detection.Scientific Reports.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems.

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.

Qu, D., Ma, Y., Yan, J., & Pyrozhenko, M. (2026). TriMeta-BFNet: A Tri-Meta Stacked Atypical-Frequency Bayesian Fourier Neural Network for Hallucination-Resistant Community Detection. Mathematics, 14 (13), 2283.

Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In IJCAI (pp. 3411-3417).

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. In Proceedings of ICASSP.

Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., & Chen, B. (2022, May). Vsegan: Visual speech enhancement generative adversarial network. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)(pp. 7308-7311). IEEE.

Wang, X., Suzuki, Y., Li, C. M., & Tanaka, S. (2025, June). Micropatterned PDMS damping layer integration to enhance the fidelity of piezoelectric MEMS speakers without SPL penalty. In2025 23rd International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers)(pp. 455-458). IEEE.

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification.IEEE Transactions on Multimedia,23, 1666-1680.

Ma, Y., & Qu, D. (2026, February). IMFM-GNN: Interpretable Multimodel GNN Via Fourier-Markov Dynamics for Community Detection in Sustainable Energy Networks. In 2026 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA) (pp. 1-6). IEEE.

Liang, R., Xu, S., Xie, C., Chen, J., Ren, F., Yang, S., & Yabe, T. (2026). Abstain Mask Retain Core: Time Series Prediction by Adaptive Masking Loss with Representation Consistency. Advances in Neural Information Processing Systems, 38, 99445-99471.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257–286.

Xiaoteng Ma, Yiqin Yang, Hao Hu, Qihan Liu, Jun Yang, Chongjie Zhang, Qianchuan Zhao, and Bin

Liang. Offline reinforcement learning with value-based episodic memory. arXiv preprint arXiv:2110.09796, 2021b.

Xu, X., Tu, W., & Yang, Y. (2023). Pcnn: A lightweight parallel conformer neural network for efficient monaural speech enhancement. arXiv preprint arXiv:2307.15251.

Dai, S., Ullmer, B., & Newman, W. E. (2024, February). MorphMatrix: A toolkit facilitating shape-changing interface design. In Proceedings of the Eighteenth International Conference on Tangible, Embedded, and Embodied Interaction (pp. 1-12).

Qu, D., & Ma, Y. (2026). Fourier–Transformer Mixer Network for Efficient Video Scene Graph Prediction.Engineering Proceedings,120(1), 16.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.

Zhang H, Chen J, Wu D, et al. Duplexsla: A full-duplex spoken language model with synchronized speech, language, and action[J]. arXiv preprint arXiv:2605.20755, 2026.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., et al. (2016). Deep Speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of ICML.

Tao X, Wu X, Zhuo S, et al. Unveiling the sedentary epidemic through insights from college students in Guangdong[J]. Scientific Reports, 2025, 15(1): 35197.

Qu, D., Ma, Y., & Wang, Y. (2026). STR-GNN: stability-regularized graph neural networks for suppressing spurious temporal variations in dynamic community detection. Scientific Reports.

Liang, H., Huang, C., Xu, Y., Tang, C., Ye, W., Zhang, J., ... & Xu, L. (2024). Llava-slt: Visual language tuning for sign language translation. arXiv preprint arXiv:2412.16524.

Chen, C., Zhang, J., Lakshmikanth, S. K., Fang, Y., Shao, R., Wetzstein, G., ... & Adeli, E. (2025). The language of motion: Unifying verbal and non-verbal language of 3d human motion. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 6200-6211).

Huang, J., Patel, A. N., Narasimha, S. M., Mishne, G., & Gilja, V. (2025). Word-level error analysis in decoding systems: From speech recognition to brain-computer interfaces. In Proc. Interspeech 2025 (pp. 5563-5567).

Xu, X., Tu, W., & Yang, Y. (2023). CASE-Net: Integrating local and non-local attention operations for speech enhancement. Speech Communication, 148, 31-39.

Cao, X., Tao, J., Liu, Z., Lyu, R., & Li, J. (2026). Handling Missing Data in CALL: A Data Quality-Driven Imputation Framework for Learner Analytics. Future-Adaptive Intelligence and Lifelong Systems, 1 (1).

Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(7), 2834-2843.

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Shneiderman, B. (1983). Direct manipulation: A step beyond programming languages. IEEE Computer, 16(8), 57–69.

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

Li, X., Hu, J., & Liu, X. (2020). Harmonic Distortion Optimization for Sigma-Delta Modulators Interface Circuit of TMR Sensors.Sensors,20(4), 1041.