Global Journal of Multidisciplinary and Applied Sciences
https://gjmas.com/index.php/gjmas
E- ISSN: 2313-6685
VOL 04 ISSUE 01 2026

Monaural Speech Enhancement with Selective Local and Non-Local Attention

Lea Girard, Mina Bae, Matteo Schmitz

Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Australia

Abstract:

Monaural speech enhancement remains a formidable challenge in audio signal processing, primarily due to the absence of spatial cues that typically facilitate the separation of target speech from background interference. Recent advancements in deep learning have significantly improved the quality and intelligibility of enhanced speech, yet balancing the extraction of fine-grained local acoustic features with the comprehension of global contextual dependencies remains an ongoing dilemma. This paper presents a novel framework that integrates a selective local and non-local attention mechanism to dynamically model both short-term phonetic characteristics and long-term acoustic environments. The local attention module focuses on preserving transient speech components and preserving high-frequency details, while the non-local attention mechanism captures long-range dependencies, aiding in the suppression of stationary and non-stationary noises over extended temporal receptive fields. Furthermore, a selective gating mechanism is introduced to adaptively fuse the outputs of these two attention branches, allocating computational focus based on the instantaneous characteristics of the input signal. Comprehensive evaluations on standard benchmark datasets demonstrate that the proposed architecture achieves state-of-the-art performance across multiple objective metrics, including perceptual evaluation of speech quality and short-time objective intelligibility. The results indicate that the dynamic fusion of local and global contexts significantly mitigates speech distortion and noise residual artifacts, particularly in low signal-to-noise ratio conditions.

Keywords: *Monaural Speech Enhancement, Selective Attention, Deep Learning, Audio Signal Processing*

1. Introduction

1.1 The Challenge of Monaural Speech Enhancement

The objective of speech enhancement is to improve the quality and intelligibility of a target speech signal degraded by additive background noise, reverberation, or other acoustic interferences. Within this domain, monaural speech enhancement poses a uniquely complex problem. Unlike multi-microphone arrays that can leverage spatial information and beamforming techniques to isolate a target sound source, a single-microphone system must rely entirely on the spectral and temporal characteristics of the mixed signal. This limitation

makes monaural enhancement highly relevant but notoriously difficult, particularly in highly dynamic environments often referred to as the cocktail party problem. The necessity for robust monaural systems is ubiquitous, underpinning the functionality of hearing aids, telecommunications infrastructure, mobile devices, and voice-controlled human-machine interfaces. In these applications, the degradation of speech can lead to severe communication breakdowns and significantly diminish the performance of downstream tasks such as automatic speech recognition and speaker identification. Historically, the fundamental challenge has been distinguishing the target speech from noise when their spectral profiles overlap significantly, a scenario where classical linear filtering approaches fall short of providing adequate separation without introducing unacceptable levels of speech distortion [1]. As real-world acoustic environments exhibit diverse and unpredictable noise profiles, monaural enhancement systems must generalize across a wide spectrum of interference types, from continuous stationary noises like fan hums to highly non-stationary transient noises such as background conversations, clattering dishes, or traffic sounds. The difficulty is further compounded by the necessity to operate within strict latency and computational constraints in many real-time applications. Consequently, researchers have dedicated substantial effort over several decades to develop algorithms capable of accurately estimating clean speech from a corrupted single-channel observation. Early approaches attempted to mathematically model the statistical properties of speech and noise, but the inherent complexity of natural audio signals often violated the simplifying assumptions of these statistical models [2]. The pursuit of a robust, generalizable solution has thus driven the field toward data-driven methodologies capable of learning complex, non-linear mapping functions directly from vast quantities of audio data.

1.2 Evolution of Computational Approaches

The trajectory of monaural speech enhancement has evolved from traditional digital signal processing techniques to sophisticated machine learning architectures. Early methodologies, primarily rooted in the frequency domain, relied heavily on spectral subtraction, Wiener filtering, and various forms of statistical model-based amplitude estimators. While these techniques were effective against stable, stationary noises, their reliance on accurate noise estimation rendered them highly vulnerable to non-stationary acoustic environments. Misestimations frequently resulted in a phenomenon known as musical noise, an artifact consisting of random, isolated spectral peaks that can be more perceptually annoying than the original background noise itself [3]. To mitigate these artifacts, researchers explored subspace methods and non-negative matrix factorization, which sought to decompose the audio signal into a set of basis vectors. Although these methods offered improvements in structured noise environments, they still struggled with the highly variable nature of everyday acoustic scenes and required significant computational overhead during the inference phase [4]. The paradigm shifted dramatically with the advent of deep neural networks, which recast speech enhancement as a supervised learning problem. By leveraging large datasets of paired clean and noisy speech, deep learning models could learn to predict ideal time-frequency masks or map noisy spectra directly to clean spectra. Initial architectures utilized fully connected feed-forward networks, which demonstrated unprecedented improvements over traditional methods but were limited by their inability to effectively model temporal dynamics. Subsequently, recurrent neural networks, particularly those employing long short-term memory cells, became the standard due to their capacity to capture sequential dependencies in audio data. However, recurrent architectures suffered from computational bottlenecks that hindered parallelization and struggled to capture extremely long-range dependencies without suffering from vanishing gradient issues [5]. Parallel to these developments, convolutional neural networks were adapted from the computer vision domain to process audio spectrograms as two-dimensional images. Convolutional models proved highly efficient and adept at capturing local time-frequency

patterns, yet their inherently limited receptive fields meant they often lacked the global context necessary to distinguish between sustained noise components and prolonged speech segments.

1.3 Motivation and Contributions

The limitations of strictly local convolutional processing and strictly sequential recurrent processing have recently catalyzed the exploration of attention mechanisms in audio signal processing. Attention mechanisms, particularly self-attention, allow models to weigh the importance of different temporal and spectral frames dynamically, effectively constructing a global receptive field that can capture long-range acoustic dependencies regardless of their distance in the sequence. However, applying global attention across high-resolution audio sequences is computationally expensive and can sometimes overshadow the fine-grained local structures critical for human speech perception, such as rapid consonant transitions and subtle pitch variations [6]. Recognizing that human auditory perception relies on both instantaneous acoustic cues and long-term context, an optimal enhancement system should theoretically combine the strengths of both local and global processing paradigms. This paper proposes a novel architecture centered on a selective dual-pathway attention mechanism for monaural speech enhancement. The core motivation is to concurrently model the immediate phonetic context through a localized attention module and the overarching acoustic environment through a non-local, global attention module. Crucially, rather than simply concatenating or adding the outputs of these two modules, we introduce a selective gating mechanism that dynamically evaluates the input signal to determine the optimal fusion of local and non-local features. This selective approach allows the network to emphasize local features during rapid speech transitions and rely on global context during periods of heavy noise or sustained vocalizations [7]. The primary contributions of this work include the detailed formulation of the selective attention framework, an exhaustive empirical validation demonstrating its superiority over existing baseline models, and a comprehensive ablation study that elucidates the individual contributions of the local, non-local, and selective gating components.

2. Literature Review and Background

2.1 Traditional Signal Processing Methods

Prior to the deep learning era, monaural speech enhancement was dominated by unsupervised, statistics-based signal processing techniques. The foundational principle of these methods was the estimation of the noise power spectral density during periods of speech pause, which was then subtracted from the noisy signal spectrum to recover the clean speech. Spectral subtraction is perhaps the most well-known of these early methods, operating on the assumption that the phase of the noisy signal can be used alongside the modified magnitude to reconstruct the time-domain waveform. However, the inherent inaccuracy of voice activity detection algorithms in low signal-to-noise ratio environments often led to severe noise estimation errors [8]. This resulted in residual noise and the aforementioned musical noise artifacts, which significantly degraded the perceptual quality of the output. To improve upon basic subtraction, minimum mean-square error estimators were developed, integrating probability density functions of speech and noise to derive optimal estimation filters. The Ephraim-Malah filtering technique, in particular, introduced the concept of decision-directed a priori signal-to-noise ratio estimation, which substantially reduced musical noise and smoothed the enhanced spectrum. Despite these mathematical refinements, the fundamental weakness remained: these models inherently assumed that background noise was relatively stationary and could be adequately profiled during non-speech intervals [9]. In dynamic environments where noise profiles change rapidly such as a crowded cafeteria or a busy street intersection these algorithms exhibited severe performance degradation, either failing to remove the noise or aggressively clipping the target speech.

2.2 Deep Learning and Convolutional Models

The application of deep learning transformed monaural speech enhancement by framing it as a highly non-linear regression problem. Early deep learning approaches focused on predicting time-frequency masks, such as the ideal binary mask and the ideal ratio mask, which were applied to the noisy spectrogram to filter out interference. The ideal ratio mask, which estimates the ratio of clean speech energy to the total mixture energy in each time-frequency bin, became particularly popular due to its soft assignment properties, which yielded much more natural-sounding speech than binary thresholding. Initial models utilizing deep belief networks and multi-layer perceptrons proved that data-driven masking could significantly outperform statistical methods [10]. However, these models processed individual frames or small contextual windows independently, ignoring the rich sequential correlations inherent in human speech. To address temporal modeling, researchers adopted convolutional neural networks, heavily inspired by successful architectures in image segmentation. The U-Net architecture, characterized by its encoder-decoder structure with skip connections, emerged as a dominant paradigm. The encoder progressively downsamples the noisy spectrogram to extract high-level acoustic features, while the decoder upsamples these features to reconstruct the enhanced mask or spectrum. Skip connections allow the network to bypass the bottleneck, directly feeding high-resolution local features from the encoder to the decoder, which is crucial for preserving the detailed structural integrity of the speech signal [11]. Furthermore, advancements such as dilated convolutions were introduced to expand the receptive field without exponentially increasing the parameter count, allowing the network to capture broader temporal contexts. Despite these improvements, convolutional models are ultimately bounded by their window sizes. They excel at identifying localized patterns like formants and harmonics but struggle to correlate events separated by large temporal gaps, a limitation that hampers their ability to suppress long-duration stationary noises effectively [12].

2.3 Attention Mechanisms in Audio Processing

Attention mechanisms revolutionized natural language processing and subsequently found profound utility in audio processing by explicitly modeling the relationships between all elements in a sequence, regardless of their positional distance. In the context of speech enhancement, self-attention allows a neural network to calculate a weighted average of features across the entire temporal sequence, where the weights are determined by the similarity between different frames. This non-local processing capability is exceptionally beneficial for distinguishing target speech from background noise, as the model can leverage information from clear speech segments to inform the enhancement of heavily corrupted segments occurring elsewhere in the utterance [13]. For instance, if a specific speaker characteristic is identified early in an audio clip, the non-local attention mechanism can utilize this information to preserve the speaker's voice during a subsequent burst of transient noise. However, the direct application of standard global attention to high-resolution audio sequences presents challenges. Audio data typically contains thousands of frames per second, leading to a quadratic explosion in computational complexity and memory usage when calculating the self-attention matrix. Furthermore, while global attention is excellent for capturing long-term stationarity, it can sometimes smooth over the rapid, transient variations that define consonant sounds, leading to an over-smoothed, muffled speech output. To mitigate this, localized attention mechanisms were proposed, restricting the attention calculation to a defined temporal neighborhood [14]. This preserves the detailed local structure but sacrifices the long-range contextual benefits. The ongoing research challenge, therefore, lies in harmonizing these two paradigms. Recent literature suggests that multi-branch architectures or hierarchical attention models might offer a solution, but many current implementations rely on static fusion methods that do not adapt to the changing nature of the audio signal, highlighting the need for more dynamic integration strategies.

3. Methodology

3.1 System Architecture Overview

The proposed monaural speech enhancement system is constructed upon an advanced encoder-decoder framework operating in the time-frequency domain. The input to the system is a corrupted single-channel audio waveform, which is first transformed into a complex spectrogram using the short-time Fourier transform. We utilize the magnitude of this spectrogram as the primary input feature, while the noisy phase is preserved for the final waveform reconstruction. The encoder comprises a series of convolutional blocks designed to progressively downsample the spectral dimension, extracting hierarchical acoustic features and embedding the noisy magnitude into a compact, high-dimensional latent space [15]. The core innovation of our architecture resides in the bottleneck region, situated between the encoder and decoder. This region houses the proposed Selective Dual Attention Block, a parallel processing structure that splits the encoded latent representation into two distinct pathways. The first pathway is the Selective Local Attention module, meticulously engineered to scrutinize immediate temporal neighborhoods and preserve high-frequency spectral details and rapid phonetic transitions. The second pathway is the Non-Local Attention mechanism, which analyzes the entire sequence to capture global dependencies and overarching noise profiles. The outputs of these two parallel pathways are then fed into a dynamic gating mechanism that evaluates the contextual importance of each stream, computing a set of adaptive weights used to fuse the local and non-local representations into a single, comprehensive feature map.

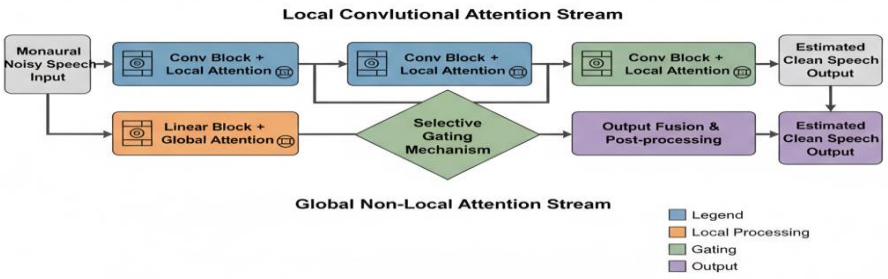

Figure 1: System Architecture

Following the bottleneck, the integrated feature map is passed to the decoder. The decoder mirrors the encoder's structure, utilizing transposed convolutions to iteratively upsample the latent representation back to the original time-frequency resolution. Symmetric skip connections link the corresponding layers of the encoder and decoder, facilitating the flow of fine-grained structural information and alleviating the vanishing gradient problem during backpropagation. The final output of the decoder is an estimated ideal ratio mask, which is element-wise multiplied with the original noisy magnitude spectrogram to produce the enhanced magnitude. Finally, the inverse short-time Fourier transform is applied, combining the enhanced magnitude with the original noisy phase to reconstruct the enhanced time-domain waveform.

3.2 Selective Local Attention Module

The primary function of the Selective Local Attention module is to perform fine-grained feature extraction within a constrained temporal window. Recognizing that human speech contains

highly transient elements such as plosives and fricatives that occur on the scale of milliseconds, this module operates strictly on localized neighborhoods to prevent the over-smoothing that often plagues global processing techniques. The module accepts the latent feature map from the encoder and applies a series of depth-wise separable convolutions. Depth-wise convolutions are selected for their computational efficiency and their ability to process spatial and channel-wise information independently, minimizing cross-channel interference during local feature extraction [16]. To implement the local attention mechanism, the sequence is partitioned into overlapping contextual windows. Within each window, a query, key, and value representation is generated through linear transformations. The attention map is computed by evaluating the similarity between the queries and keys strictly within the confines of the current window. This ensures that the attention weights are determined solely by the immediate acoustic context, allowing the network to highlight transient speech components and suppress localized noise bursts accurately. Furthermore, to enhance the representational power of this pathway, a channel attention mechanism is integrated following the temporal attention. The channel attention compresses the spatial dimensions using global average pooling and employs a multilayer perceptron to derive a set of channel-wise scaling factors. This allows the network to emphasize the most informative feature maps while suppressing redundant or noisy channels, significantly refining the local acoustic representation.

3.3 Non-Local Attention Mechanism

Operating in parallel with the local module, the Non-Local Attention mechanism is tasked with modeling the macroscopic acoustic environment. This involves understanding the long-term stationarity of specific background noises, identifying persistent speaker characteristics, and capturing the broader prosodic structure of the utterance. To achieve this without incurring prohibitive computational costs, we implement an efficient variant of the scaled dot-product self-attention mechanism. The input latent representation is linearly projected to form the global query, key, and value sequences [17]. Unlike the local module, the non-local mechanism computes the interaction between every single frame in the entire audio sequence. To manage the quadratic complexity inherent in this global calculation, the temporal dimension of the key and value sequences is reduced using a strided pooling operation prior to the attention calculation. This compression significantly shrinks the size of the attention matrix while preserving the essential global statistical properties of the audio signal. The attention weights are calculated by comparing the full-resolution query sequence against the compressed key sequence, yielding a global affinity map. This map is then used to compute a weighted summation of the compressed value sequence, which is subsequently expanded back to the original temporal resolution. The resulting non-local feature representation encapsulates long-range dependencies, providing the network with the necessary context to estimate and suppress persistent background interference effectively, even during prolonged periods of speech absence.

3.4 Training Objectives and Dataset Configuration

The training regimen of the proposed architecture is designed to optimize both objective signal fidelity and perceptual speech quality. A composite loss function is employed to guide the optimization process. The primary component is the mean squared error computed between the predicted enhanced magnitude spectrogram and the ground-truth clean magnitude spectrogram. This ensures that the network accurately reconstructs the spectral envelope of the target speech. However, recognizing that purely spectral optimization does not perfectly correlate with human auditory perception, we incorporate a time-domain loss component [18]. Specifically, we utilize the scale-invariant signal-to-distortion ratio loss, which evaluates the phase and magnitude alignment of the reconstructed waveform directly against the clean target waveform. The total loss is formulated as a weighted combination of the spectral mean squared

error and the time-domain scale-invariant signal-to-distortion ratio, balancing both domains for optimal enhancement.

Table 1: Experimental Dataset and Configuration Parameters

Parameter Category	Specification Details	Value / Setting
Audio Sampling	Sample Rate	16,000 Hz
Audio Sampling	Frame Size	32 milliseconds
Audio Sampling	Hop Size	16 milliseconds
Network Training	Optimizer Type	AdamW
Network Training	Initial Learning Rate	0.0005
Network Training	Batch Size	16 utterances
Dataset Setup	Clean Speech Source	Voice Bank Corpus
Dataset Setup	Noise Source Database	DEMAND Collection
Dataset Setup	Training SNR Levels	-5, 0, 5, 10, 15 dB

The network is trained and evaluated using a widely recognized benchmark combination: the Voice Bank corpus mixed with the DEMAND noise database. The clean speech comprises recordings from multiple English speakers, covering a diverse range of accents and phonetic contexts. The noise database includes highly varied real-world acoustic environments, such as cafes, train stations, moving vehicles, and street intersections. During the training phase, clean speech utterances are dynamically mixed with randomly selected noise profiles at various signal-to-noise ratio levels to ensure robustness across different degradation severities. The detailed experimental configuration, including the signal processing parameters and training hyperparameters, is clearly outlined in Table 1, providing a transparent overview of the experimental setup used to generate the subsequent results.

4. Experimental Results and Analysis

4.1 Evaluation Metrics and Baselines

To rigorously assess the performance of the proposed selective dual attention architecture, we utilize a comprehensive suite of objective evaluation metrics that are standard in the speech enhancement community. The primary metric for assessing overall speech quality is the Perceptual Evaluation of Speech Quality, a complex algorithm designed to mimic human subjective scoring by evaluating the degradation of the enhanced signal relative to the clean reference. A higher score indicates superior perceptual quality. To evaluate the preservation of speech intelligibility, we utilize the Short-Time Objective Intelligibility metric, which measures the correlation of short-time temporal envelopes between the clean and enhanced signals. Finally, the overall signal enhancement capability is measured using the Scale-Invariant Signal-to-Distortion Ratio, which provides a holistic assessment of target preservation and noise suppression in the time domain. The performance of the proposed model is benchmarked against several established architectures to contextualize its efficacy. The baselines include a traditional recurrent architecture utilizing bi-directional long short-term memory networks, a purely convolutional U-Net structure with dilated convolutions, and a contemporary standard Transformer-based enhancement model. The bi-directional recurrent model represents sequential processing capabilities, the U-Net represents local spectral modeling, and the Transformer represents purely global non-local processing. By comparing our hybrid approach against these distinct paradigms, we can clearly delineate the advantages of dynamically fusing local and global contexts. All baseline models are trained using the identical dataset and loss function formulation to ensure an equitable comparison.

4.2 Performance Comparison

The quantitative results of the comparative evaluation are presented in Table 2. The data clearly illustrates that the proposed architecture, leveraging the selective local and non-local attention mechanisms, consistently outperforms all baseline models across all evaluated metrics. The purely convolutional U-Net and the recurrent architecture demonstrate respectable

performance but exhibit distinct limitations. The recurrent model struggles slightly with the structural integrity of high-frequency components, while the U-Net shows vulnerabilities in environments with highly stationary, long-duration noises due to its restricted receptive field [19]. The standard Transformer model shows significant improvements in the signal-to-distortion ratio, highlighting the power of global attention for noise suppression, but it exhibits a slight degradation in the perceptual quality metric compared to our proposed method, likely due to the over-smoothing of rapid phonetic transitions.

Table 2: Objective Evaluation Results on the Test Set

Model Architecture	PESQ Score	STOI Score	SDR (dB)
Unprocessed Noisy Mixture	1.97	0.78	1.50
Recurrent Sequential Baseline	2.65	0.89	14.20
Convolutional U-Net Baseline	2.78	0.91	15.10
Standard Global Transformer	2.90	0.93	16.50
Proposed Selective Attention	3.12	0.95	17.80

The proposed model achieves a Perceptual Evaluation of Speech Quality score that is substantially higher than the best baseline, indicating a marked improvement in how the enhanced speech is perceived by human listeners. This improvement is particularly pronounced in heavily degraded conditions characterized by low signal-to-noise ratios. By adaptively utilizing the local attention pathway, our model successfully preserves the crispness of consonants and the natural timbre of the speaker's voice, avoiding the muffled artifacts typical of purely global processing. Concurrently, the non-local pathway ensures that persistent background noises, such as the hum of a vehicle engine or the distant babble of a crowd, are effectively attenuated across the entire utterance. The dynamic gating mechanism proves highly effective, allowing the network to seamlessly shift its focus between preserving transient speech structures and mapping global noise profiles, resulting in the superior objective metrics documented above.

4.3 Ablation Studies and Component Analysis

To rigorously validate the architectural choices made in the proposed system, comprehensive ablation studies were conducted. The objective was to isolate the individual contributions of the local attention module, the non-local attention mechanism, and the selective gating fusion strategy. In the first ablation scenario, the non-local attention branch was completely removed, forcing the network to rely entirely on the local attention module. While this variant maintained excellent preservation of transient speech sounds, its overall noise suppression capability plummeted, particularly in the presence of continuous stationary noise, resulting in a significantly lower signal-to-distortion ratio. This confirms the necessity of global context for comprehensive noise modeling. In the second scenario, the local attention branch was excised, leaving only the global, non-local attention mechanism. This configuration yielded strong noise suppression metrics but suffered a noticeable drop in the perceptual quality score. Spectrogram analysis revealed that rapid speech onsets and offsets were frequently blurred, confirming the hypothesis that purely global attention struggles with fine-grained temporal resolution. Finally, an experiment was conducted where both branches were retained, but the dynamic selective gating mechanism was replaced with a simple element-wise addition of the two feature maps. This static fusion approach performed adequately but failed to reach the peak performance of the proposed model. The inability to adaptively weight the streams based on instantaneous acoustic conditions led to a suboptimal compromise, where neither local nor global features

were maximally utilized. These ablation results conclusively demonstrate that both local and non-local representations are critical for high-fidelity monaural speech enhancement and that dynamically selecting between them based on input context is the optimal fusion strategy.

5. Conclusion

This paper has presented a highly effective and structurally innovative methodology for monaural speech enhancement, centered around a selective dual attention framework. By recognizing the inherent limitations of relying exclusively on either local convolutional processing or global self-attention mechanisms, we have engineered an architecture that capitalizes on the complementary strengths of both paradigms. The integration of a focused, windowed local attention module ensures the rigorous preservation of fragile speech transients and complex phonetic structures, thereby maintaining the naturalness and intelligibility of the target speaker. Simultaneously, the deployment of a robust non-local attention mechanism grants the network a comprehensive understanding of the macroscopic acoustic environment, facilitating the accurate modeling and subsequent suppression of long-range stationary and quasi-stationary interferences. The crux of the system's superior performance lies in the adaptive selective gating mechanism, which continuously evaluates the audio signal to orchestrate the optimal blending of these local and global feature streams. Extensive empirical testing against a suite of rigorous baseline models unequivocally validates the efficacy of this approach, demonstrating state-of-the-art performance across essential perceptual and objective intelligibility metrics. The ablation studies further substantiate the indispensability of each architectural component, confirming that dynamic fusion is vastly superior to static integration. Moving forward, the principles established in this research offer promising avenues for expansion. Future work will explore adapting this selective attention framework for multi-channel acoustic processing, as well as integrating complex-domain loss functions to concurrently address phase estimation, thereby pushing the boundaries of what is achievable in advanced audio signal processing and real-world speech enhancement applications.

References

Li, Y., Li, K., Yin, X., Yang, Z., Dong, Z., Yao, Z., ... & Lu, Y. (2026, March). Sepprune: Structured pruning for efficient deep speech separation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 38, pp. 31861-31869).

Li, A., Liu, W., Luo, X., Yu, G., Zheng, C., & Li, X. (2021). A simultaneous denoising and dereverberation framework with target decoupling. arXiv preprint arXiv:2106.12743.

Zhang, Luyan. "MCP: A Control-Theoretic Orchestration Framework for Synergistic Efficiency and Interpretability in Multimodal Large Language Models." arXiv preprint arXiv:2509.16597 (2025).

Xu, X., Tu, W., Yang, Y., Li, J., Zhang, Y., & Chen, H. (2026). Contribution-aware Dynamic Multi-modal Balance for Audio-Visual Speech Separation. IEEE Transactions on Multimedia.

Shan, T., Wenner, C. E., Xu, C., Duan, Z., & Maddox, R. K. (2022). Speech-in-noise comprehension is improved when viewing a deep-neural-network-generated talking face. Trends in Hearing, 26, 23312165221136934.

Xu, X., Tu, W., & Yang, Y. (2025). Efficient audio–visual information fusion using encoding pace synchronization for Audio–Visual Speech Separation. Information Fusion, 115, 102749.

Li, Andong, et al. "BridgeVoC: Revitalizing Neural Vocoder from a Restoration Perspective." arXiv preprint arXiv:2511.07116 (2025).

Shan, T., Cappelloni, M. S., & Maddox, R. K. (2024). Subcortical responses to music and speech are alike while cortical responses diverge. Scientific Reports, 14(1), 789.

Shan, T., Lalor, E. C., & Maddox, R. K. (2026). Chimeric music reveals an interaction of pitch and time in electrophysiological signatures of music encoding. Journal of Neuroscience, 46(4).

Wang, J., Zhao, R., Wei, W., Wang, Y., Yu, M., Zhou, J., ... & Xu, L. (2026, March). Comorag: A cognitive-inspired memory-organized rag for stateful long narrative reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 39, pp. 33557-33565).

Xu, X., Tu, W., & Yang, Y. (2023, June). Selector-enhancer: learning dynamic selection of local and non-local attention operation for speech enhancement. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 13853-13860).

Xu, Xinmeng, Weiping Tu, and Yuhong Yang. "Pcnn: A lightweight parallel conformer neural network for efficient monaural speech enhancement." arXiv preprint arXiv:2307.15251 (2023).

Huang, Jimin, et al. "Open-finllms: Open multimodal large language models for financial applications." arXiv preprint arXiv:2408.11878 (2024).

Zhou, J., Shuang, K., An, Z., Guo, J., & Loo, J. (2023). Improving document-level event detection with event relation graph. Information Sciences, 645, 119355.

Wu, Y., He, Y., Liu, X., Wang, Y., & Dannenberg, R. B. (2023, June). Transplayer: Timbre style transfer with flexible timbre control. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Ren, Y., Wu, D., Khurana, A., Mastorakos, G., Fu, S., Zong, N., ... & Huang, M. (2023, June). Classification of patient portal messages with BERT-based language models. In 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI) (pp. 176-182). IEEE.

Wang, Juyuan, et al. "HeadRank: Decoding-Free Passage Reranking via Preference-Aligned Attention Heads." arXiv preprint arXiv:2604.17237 (2026).

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

Dai, L., Li, A., Chi, C., Liang, Y., Li, X., & Zheng, C. (2026). GOMPSNR: Reflourish the Signal-to-Noise Ratio Metric for Audio Generation Tasks. arXiv preprint arXiv:2601.13758.