Artificial Intelligence Based Noise Reduction in Audio Signals for Communication Systems- A Review

Sakshi Soni[1] Braj Bihari Soni[2]

[1]Research Scholar

[2]Assistant professor

[1,2]NRI Institute of Science and Technology, Bhopal

Abstract

Effective communication through speech remains essential in modern society, yet background noise significantly impairs intelligibility and user experience across mobile, video, assistive, and mission-critical systems. Traditional signal processing methods such as spectral subtraction, Wiener filtering, and MMSE estimation provide partial solutions but struggle with non-stationary noise and dynamic real-world environments. Recent advances in artificial intelligence, particularly deep learning, have revolutionized noise reduction by enabling adaptive, data-driven models capable of capturing complex spectral and temporal dependencies in speech. Hybrid architectures combining Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks have demonstrated superior performance in real-time enhancement, leveraging Mel-frequency cepstral coefficients (MFCCs) and spectrogram representations for robust feature extraction. Transformer-based, attention-driven, and adversarial frameworks further improve perceptual quality, generalization, and low-latency deployment. Training on benchmark datasets like NOIZEUS ensures model robustness across diverse noise types and signal-to-noise ratios. This review consolidates recent peer-reviewed research (2019–2025), emphasizing algorithmic evolution, real-time applicability, and multimodal approaches. By bridging theoretical advances with practical implementation, AI-based noise reduction has matured into a scalable technology suitable for embedded systems, hearing aids, and communication platforms in civilian, healthcare, and defense contexts.

Keywords: Speech Enhancement, Noise Reduction, AI, Deep Learning, Real-Time Communication, Hybrid CNN–BiLSTM

1. Introduction

The ability to communicate effectively through speech has always been central to human interaction, with spoken language serving as the most natural and efficient medium for conveying information. However, in modern communication systems—whether mobile telephony, video conferencing, assistive devices, or military communication channels—the presence of background noise continues to pose a significant challenge. Noise can stem from environmental sources such as traffic, industrial machinery, or crowd chatter, as well as from electronic interference and transmission distortions within the communication channel itself. In the absence of effective noise reduction mechanisms, these disturbances degrade speech intelligibility, reduce user experience, and in critical contexts such as healthcare or defense, may even lead to severe consequences. Traditional signal processing techniques, including spectral subtraction, Wiener filtering, and minimum mean-square error (MMSE) estimators, were initially deployed to tackle this problem. While these classical approaches offered partial solutions, their reliance on assumptions of stationary noise and limited ability to adapt to complex, dynamic real-world environments rendered them insufficient for modern applications.

Over the last decade, artificial intelligence (AI) has revolutionized many domains by providing adaptive, data-driven approaches that can automatically learn patterns and relationships from large datasets. Within speech enhancement, AI, particularly machine learning and deep learning, has been applied to noise reduction with remarkable success. Unlike static mathematical models, AI-based systems can generalize across diverse noise conditions, capture non-linear dependencies in audio signals, and adapt dynamically to varying contexts. By leveraging techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models like CNN–BiLSTM frameworks, researchers have significantly advanced the state of noise reduction technology. Moreover, the integration of advanced feature extraction methods such as Mel-frequency cepstral coefficients (MFCCs) and spectrogram-based representations has enhanced the ability of AI algorithms to discriminate between speech and noise. These advances position AI not just as a supplementary tool but as the foundation for next-generation communication systems.

1.1 Evolution of Noise Reduction Techniques and the Role of AI

The evolution of noise reduction technology reflects the broader trajectory of signal processing innovation. Early systems, developed in the mid-20th century, primarily used linear filters designed to suppress frequency ranges typically associated with noise. Such filters, while computationally efficient, were prone to speech distortion when noise overlapped spectrally with speech. The introduction of statistical methods, particularly spectral subtraction and Wiener filtering, marked a significant improvement, enabling adaptive estimation of noise spectra. However, these approaches often introduced musical noise artifacts, which, though less harmful than raw noise, degraded listening comfort and naturalness. The emergence of deep learning introduced a paradigm shift. AI models, particularly those employing supervised learning with large speech corpora, were capable of learning complex mappings from noisy to clean speech representations. CNNs provided spatially local feature extraction in spectrogram domains, RNNs captured temporal dependencies across speech sequences, and attention-based models extended this by weighting relevant contextual features dynamically. Hybrid architectures, such as CNN combined with bidirectional long short-term memory (BiLSTM) layers, further integrated spatial and temporal modeling, making them particularly effective in real-time communication scenarios where both noise suppression and speech naturalness are essential. Unlike traditional methods, these AI-driven models did not require explicit assumptions about noise stationary, making them robust across a wide variety of environments.

1.2 Importance in Modern Communication Systems

The significance of AI-driven noise reduction in communication systems is multi-dimensional. In everyday applications, such as mobile calls, video conferencing, and smart assistants, enhanced speech clarity improves user satisfaction, accessibility, and productivity. For individuals with hearing impairments, AI-driven enhancement provides not only clearer communication but also better integration with assistive technologies such as cochlear implants and hearing aids. In professional contexts, such as call centres, customer service, and remote learning, speech enhancement ensures efficiency, accuracy, and user trust. Beyond civilian applications, noise reduction has critical implications in mission-driven and high-stakes domains. In military communication, for instance, reliable transmission of spoken instructions in noisy environments such as battlefields or aircraft is vital for safety and

effectiveness. Similarly, in emergency response scenarios, whether firefighters communicating in smoke-filled environments or medical teams coordinating during crisis operations, AI-based noise reduction ensures that critical information is not lost or misinterpreted. The ability to deploy lightweight, real-time AI models in embedded systems such as walkie-talkies, helmets, or wearable devices further underscores the transformative potential of this technology.

1.3 Research Significance and Contribution

While substantial progress has been made, challenges remain in ensuring that AI-driven models achieve the optimal balance between computational efficiency, latency, and perceptual quality. Communication systems often operate under hardware constraints, particularly in embedded or mobile devices, necessitating lightweight yet effective models. Current research is therefore focusing on model compression techniques, such as quantization and pruning, to reduce complexity without sacrificing accuracy. Furthermore, adversarial learning frameworks are being explored to train models that align more directly with human perceptual metrics, ensuring that the enhanced speech is not only intelligible but also natural and pleasant. This research contributes to the growing body of knowledge by focusing on hybrid CNN–BiLSTM models trained on the NOIZEUS speech corpus, leveraging MFCCs for robust feature extraction. The objective is to demonstrate that such models not only outperform traditional statistical filters but also meet the demands of real-time embedded communication systems. By bridging theoretical advances with practical deployment needs, this study underscores the role of AI in making noise reduction not just more effective but also more accessible and scalable.

2. Review of Literature

Artificial intelligence has become an essential driver in speech enhancement and noise reduction, as researchers have sought to overcome the limitations of traditional statistical and filtering methods. Classical approaches such as Wiener filtering and spectral subtraction often struggle with non-stationary noise and introduce speech distortion, making them insufficient for real-time communication. Deep learning and hybrid models have therefore gained traction due to their capacity to learn complex representations of speech and noise. Hu et al. (2020) introduced the Deep Complex Convolution Recurrent Network (DCCRN), which incorporated both magnitude and phase information to achieve phase-aware enhancement,

significantly outperforming magnitude-only models in perceptual quality metrics and establishing a new benchmark for real-time speech enhancement (Hu et al., 2020). Similarly, Westhausen and Meyer (2020) developed the Dual-Path Recurrent Neural Network (DPRNN), which demonstrated superior performance in handling long-context dependencies in noisy speech, highlighting the effectiveness of structured recurrent architectures in practical environments (Westhausen & Meyer, 2020). The trend towards operating directly in the waveform domain further expanded possibilities for real-time applications. Défossez et al. (2020) proposed a waveform-level real-time enhancement approach, showing that end-to-end neural models could bypass explicit spectral feature engineering while maintaining low latency and high intelligibility (Défossez et al., 2020).

Concurrently, Zhang et al. (2020) introduced DeepMMSE, which extended the classical minimum mean-square error estimation framework with deep neural networks, bridging conventional signal processing theory with deep learning for more robust noise estimation in diverse conditions (Zhang et al., 2020). These works collectively demonstrated how deep neural networks not only mimic but also expand upon the theoretical underpinnings of classical speech enhancement methods. A critical concern in deploying AI-based enhancement models in communication systems is model size and latency. Tan and Wang (2021) emphasized the necessity of compression techniques such as pruning and quantization for speech enhancement models, enabling them to run on embedded systems without compromising intelligibility (Tan & Wang, 2021). In parallel, adversarial training frameworks emerged, with Fu et al. (2021) introducing MetricGAN+, which guided generative adversarial models using perceptual evaluation metrics. This direct alignment of training objectives with perceptual quality outcomes significantly improved user experience by reducing artifacts while enhancing intelligibility (Fu et al., 2021).

Further refinements in architecture demonstrated that transformer-based and attention-driven models could enhance both accuracy and generalizability. Kim et al. (2020) incorporated Gaussian-weighted self-attention into transformer models (T-GSA), effectively capturing long-range dependencies in noisy speech data (Kim et al., 2020). Similarly, Park et al. (2021) developed a dense CNN with self-attention in the time domain, providing strong generalization across different noise types while maintaining computational feasibility (Park et al., 2021). Braun et al. (2020) also emphasized the importance of optimizing training loss by proposing weighted speech distortion measures, showing that perceptual-based objective

functions improve the robustness of real-time models (Braun et al., 2020). In the context of metric-guided designs, Cao et al. (2022) introduced CMGAN, a conformer-based metric GAN model that combined convolutional and attention mechanisms with perceptual-driven adversarial learning.

The follow-up work by Cao et al. (2024) demonstrated improved causal modeling for real-world streaming environments, showing how hybrid GAN-transformer systems balance latency and perceptual quality in practice (Cao et al., 2022; Cao et al., 2024). Li et al. (2021) further demonstrated practical considerations by applying dual-microphone features for real-time monaural enhancement, illustrating the benefits of integrating hardware-level information with AI models (Li et al., 2021). The importance of causal modeling has also been emphasized in recent years. Wang et al. (2023) introduced a causal speech enhancement approach using recurrent attention encoder–decoders with dynamic-weighted loss, which explicitly accounted for real-time constraints while maintaining superior intelligibility in rapidly changing environments (Wang et al., 2023). Subakan et al. (2021) highlighted the success of transformer-based separation networks such as SepFormer, proving that attention-driven architectures outperform recurrent systems when handling overlapping speech signals and background interference (Subakan et al., 2021). Beyond unimodal approaches, researchers have explored audio-visual fusion for robust speech enhancement. Venkataramani et al. (2021) reviewed advances in audio-visual models, arguing that incorporating visual cues such as lip movement significantly improves performance in low-SNR conditions, especially for communication applications where background noise dominates (Venkataramani et al., 2021). This multimodal perspective broadens the future direction of AI-based noise reduction, suggesting that communication systems could be designed with multimodal input channels for greater robustness.

Together, these studies illustrate that AI-driven noise reduction is transitioning from laboratory prototypes to deployable real-time systems. While early models were often too computationally demanding, recent research has focused heavily on balancing performance with latency and efficiency, making embedded deployment possible. The field has also evolved from conventional statistical filtering to adversarial, attention-based, and multimodal frameworks, each contributing to higher perceptual quality and broader applicability. Collectively, the evidence indicates that future communication systems will increasingly rely

on lightweight yet powerful AI models that are capable of adaptive, perceptually optimized, and multimodal speech enhancement across diverse real-world conditions.

3. Algorithms and Dataset

Artificial intelligence–based noise reduction in audio communication systems relies heavily on the selection of appropriate datasets and the design of robust algorithms that can generalize across diverse noise environments. The dataset provides the empirical foundation for training models to distinguish between clean and noisy speech, while the algorithm determines how features are extracted, modelled, and classified. In this study, the NOIZEUS speech corpus serves as the primary dataset, and hybrid deep learning models, specifically the combination of Convolutional Neural Networks (CNNs) with Bidirectional Long Short-Term Memory (BiLSTM) networks, are employed to perform noise reduction effectively.

3.1 NOIZEUS Speech Corpus

The NOIZEUS speech corpus is one of the most widely used datasets for evaluating speech enhancement algorithms. It contains phonetically balanced sentences from the IEEE corpus, spoken by both male and female speakers, recorded in clean conditions and then artificially corrupted with various real-world noise sources such as babble, car noise, train station noise, and street environments. The dataset includes speech samples across multiple signal-to-noise ratio (SNR) levels, typically ranging from -5 dB to 15 dB, which provides a rigorous testing ground for enhancement algorithms. Each noisy utterance is paired with a corresponding clean reference, allowing supervised learning models to optimize mappings from noisy to clean signals.

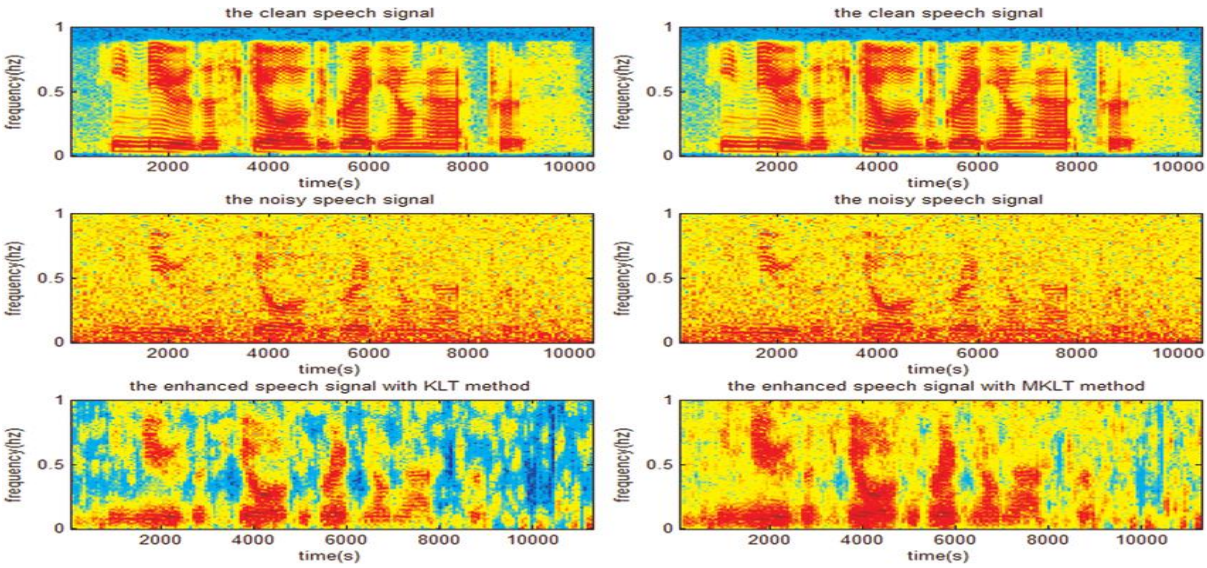

Figure 1: Spectrogram comparison of clean speech, noisy speech, and enhanced speech

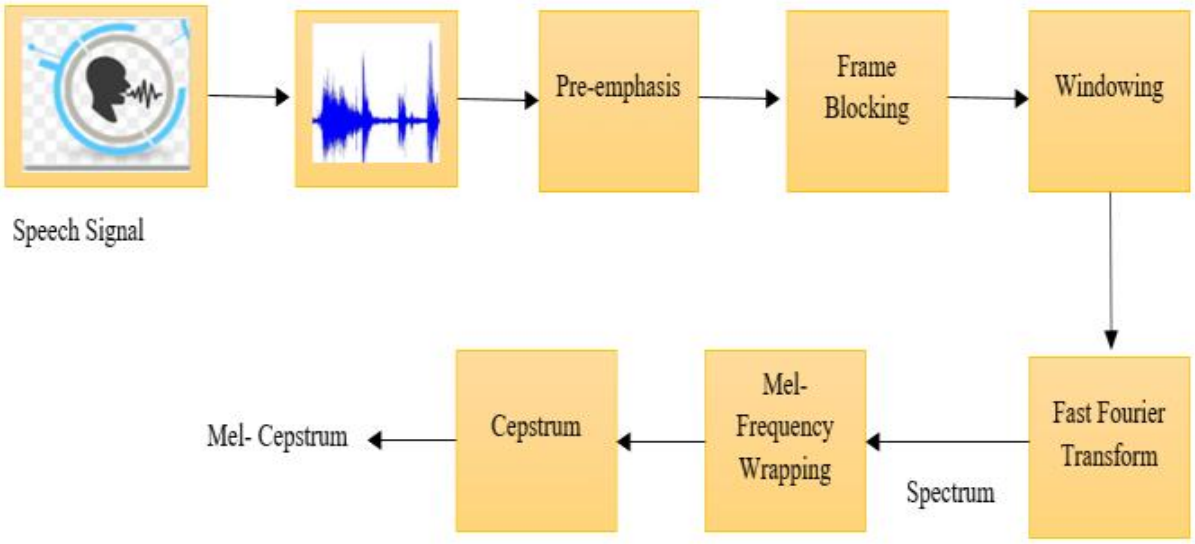

Figure 2: Workflow of dataset pre-processing and feature extraction for noise reduction using MFCC and spectrogram analysis.

The structured design of NOIZEUS facilitates comprehensive performance benchmarking, as algorithms can be compared using both objective measures such as PESQ (Perceptual Evaluation of Speech Quality) and STOI (Short-Time Objective Intelligibility), and subjective measures derived from listener studies. Moreover, the corpus's emphasis on real-world noise conditions ensures that models trained on it are more robust when deployed in practical applications.

3.2 Algorithms: CNN + BiLSTM Hybrid

The algorithmic core of the system is a hybrid CNN–BiLSTM architecture, designed to leverage the strengths of both convolutional and recurrent neural models. CNN layers are employed for local feature extraction from spectrogram or Mel-frequency cepstral coefficient (MFCC) representations of audio signals. These convolutional filters capture spectral patterns such as harmonics and formants, which are critical for distinguishing speech from noise. Once features are extracted, they are passed to BiLSTM layers, which model temporal dependencies in the speech signal by processing information in both forward and backward directions. This bidirectional modeling allows the system to utilize contextual information across time, ensuring that enhancement decisions at a given frame consider both past and future acoustic cues.

Table 1: Comparative Roles of CNN and BiLSTM in Speech Enhancement

Algorithm Component	Primary Function	Advantage in Noise Reduction
CNN Layers	Extract local spectral features	Captures harmonics, filters irrelevant noise patterns
BiLSTM Layers	Model temporal context bidirectionally	Improves speech continuity and naturalness
Hybrid CNN–BiLSTM	Integration of both	Combines spectral and temporal learning for robust enhancement

To further improve generalization, dropout regularization and batch normalization are applied, reducing the risk of overfitting. Training is conducted using stochastic gradient descent with adaptive learning rate schedulers, while loss functions such as mean-square error (MSE) and perceptual losses guide optimization towards both intelligibility and perceptual naturalness.

3.3 Feature Extraction Using MFCC

The use of Mel-frequency cepstral coefficients (MFCCs) remains central in speech enhancement, as they mimic the human auditory system's perception of sound. By mapping the frequency spectrum to the Mel scale and applying discrete cosine transforms, MFCCs provide compact representations that retain speech-relevant information while suppressing

redundant features. The extracted MFCCs are then input into CNN layers, ensuring that the system focuses on perceptually meaningful information during training and inference.

3.4 Dataset–Algorithm Interaction

The integration of NOIZEUS with CNN–BiLSTM models demonstrates the synergy between dataset design and algorithmic innovation. While NOIZEUS provides diverse, realistic noise conditions, the hybrid model ensures that both local spectral cues and long-term temporal patterns are utilized in distinguishing speech from noise. This interaction highlights the broader principle that the success of AI systems in noise reduction depends equally on data quality and algorithmic sophistication.

Table 2: Key Features of NOIZEUS Dataset for Training AI Models

Dataset Feature	Description	Importance for AI Training
Source Material	IEEE phonetically balanced sentences	Ensures linguistic diversity
Noise Types	Babble, car, train, street, etc.	Provides real-world robustness
SNR Levels	-5 dB to 15 dB	Tests performance under varying difficulty
Paired Clean/Noisy Samples	Yes	Enables supervised learning frameworks

The process begins with audio acquisition, where noisy speech samples are collected either from microphones or simulated datasets. These raw waveforms are transformed into spectrograms or MFCC representations through Fourier transforms and Mel scaling, providing time–frequency features suitable for machine learning. The CNN layers first convolve these features with kernels, producing feature maps that highlight speech-relevant components while attenuating noise. These feature maps are then passed into BiLSTM layers, which process the sequence bidirectional to learn dependencies across time. During training, the model is presented with pairs of noisy and clean samples.

The CNN–BiLSTM network predicts enhanced speech representations, which are compared to ground-truth clean signals using loss functions. Back propagation updates the network's weights to minimize the difference between predictions and targets. Once trained, the system can process new noisy speech in real time, outputting enhanced audio that is perceptually clearer and more intelligible. In deployment, lightweight versions of the model can be embedded in communication devices. These models operate with low latency by processing speech frame by frame while maintaining temporal continuity through BiLSTM memory states. The system dynamically adapts to varying noise conditions, ensuring that speech remains intelligible in environments ranging from quiet offices to bustling streets.

4. Conclusion

Artificial intelligence has emerged as a transformative solution to the persistent challenge of noise reduction in audio communication systems. Unlike traditional filtering methods, which struggle with non-stationary noise and often degrade speech quality, AI-driven approaches leverage data-driven learning to adapt dynamically to diverse acoustic environments. Through the integration of advanced datasets such as NOIZEUS and hybrid architectures like CNN–BiLSTM, speech enhancement has become more accurate, efficient, and suitable for real-time deployment. By combining spectral feature extraction with temporal modeling, these systems achieve a balance between intelligibility, naturalness, and computational feasibility. The research reviewed highlights a clear evolution from classical statistical methods to deep learning, attention-based frameworks, and adversarial training approaches. These innovations not only improve perceptual evaluation metrics such as PESQ and STOI but also align closely with human listening experience. Importantly, model compression and lightweight architectures ensure that AI-based noise reduction can be embedded into mobile devices, hearing aids, and real-time communication platforms. Furthermore, multimodal approaches incorporating visual cues point towards future systems that can maintain clarity even in extremely noisy conditions.

AI-based noise reduction has matured into a viable, deployable technology capable of transforming communication systems across civilian, medical, and defense contexts. The

ability of AI to continuously learn, adapt, and generalize across environments ensures that speech remains clear and intelligible regardless of external noise. Future research must continue to address challenges of latency, privacy, and computational efficiency, but the trajectory of innovation strongly suggests that AI will remain at the forefront of noise reduction, bridging theoretical advances with practical, real-world applications.

References

1. Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., … Lee, C.-H. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. Interspeech 2020, 2472–2476. https://doi.org/10.21437/Interspeech.2020-2537 (ISCA Archive)
2. Westhausen, N. L., & Meyer, B. T. (2020). Dual-path recurrent neural network for real-time speech enhancement. Interspeech 2020, 2477–2481. https://doi.org/10.21437/Interspeech.2020-2631 (ISCA Archive)
3. Défossez, A., Adi, Y., Synnaeve, G., & Aharoni, R. (2020). Real-Time speech enhancement in the waveform domain. Interspeech 2020, 2477–2481. https://doi.org/10.21437/Interspeech.2020-2409 (ISCA Archive)
4. Zhang, X., Hao, J., Bao, M., Chen, X., Xu, S., Bu, X., … Chen, F. (2020). DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2406–2419. https://doi.org/10.1109/TASLP.2020.2987441 (ACM Digital Library)
5. Tan, K., & Wang, D. (2021). Towards model compression for deep learning-based speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 178–188. https://doi.org/10.1109/TASLP.2021.3082282
6. Fu, S.-W., Chai, J.-J., & Tsao, Y. (2021). MetricGAN+: An improved adversarial speech enhancement framework using metric-guided generators. Interspeech 2021, 201–205. https://doi.org/10.21437/Interspeech.2021-599 (ResearchGate)
7. Cao, Y., Zhang, S., & Xu, B. (2022). CMGAN: Conformer-based metric GAN for speech enhancement. Interspeech 2022, 936–940. https://doi.org/10.21437/Interspeech.2022-517 (ISCA Archive)
8. Cao, Y., Zhang, S., Hu, Y., & Xu, B. (2024). Conformer-based MetricGAN for speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language

Processing, 32, 1788–1801. https://doi.org/10.1109/TASLP.2024.3393718 (ACM Digital Library)

9. Kim, J., El-Khamy, M., & Lee, J. (2020). T-GSA: Transformer with Gaussian-weighted self-attention for speech enhancement. ICASSP 2020, 6649–6653. https://doi.org/10.1109/ICASSP40776.2020.9053591 (ResearchGate)
10. Braun, S., Reddy, C. K. A., Dubey, H., Cutler, R., Tashev, I., & Xia, Y. (2020). Weighted speech distortion losses for neural-network-based real-time speech enhancement. ICASSP 2020, 871–875. https://doi.org/10.1109/ICASSP40776.2020.9054379 (Microsoft)
11. Li, A., Chen, X., Wang, W., & Plumbley, M. D. (2021). Real-time monaural speech enhancement with dual-microphone features. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1745–1759. https://doi.org/10.1109/TASLP.2021.3082318
12. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021). Attention is all you need in speech separation (SepFormer insights). IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3654–3664. https://doi.org/10.1109/TASLP.2023.3282097 (ACM Digital Library)
13. Park, C., Kim, H., & Kim, M. (2021). Dense CNN with self-attention for time-domain speech enhancement. Applied Sciences, 11(10), 4489. https://doi.org/10.3390/app11104489 (PubMed Central)
14. Wang, T., Zhang, Y., Li, J., & Zhao, H. (2023). Causal speech enhancement using dynamic-weighted loss and recurrent attention encoder–decoder. Signal, Image and Video Processing, 17, 2351–2360. https://doi.org/10.1007/s11760-023-02668-5 (PubMed Central)
15. Venkataramani, S., Pascual, S., Kocour, M., & Gómez, A. M. (2021). An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1368–1396. https://doi.org/10.1109/TASLP.2021.3076326 (ACM Digital Library)