INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 01 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Benchmarking Objective Quality Metrics for Audio Generation and Restoration: A Comprehensive Analysis

Estelle Mercier, Anna Clifford, Bertrand Dean, Doyun Choi

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

Abstract. *The rapid advancement of deep learning technologies has fundamentally transformed the fields of audio generation and restoration, enabling the synthesis of highly realistic speech, music, and environmental sounds, as well as the sophisticated recovery of degraded audio signals. However, the evaluation of these advanced systems remains a significant bottleneck. Subjective evaluation, predominantly through Mean Opinion Score testing, is considered the gold standard for assessing perceptual audio quality. Yet, it is inherently expensive, time-consuming, and susceptible to various human biases, rendering it impractical for the rapid and iterative development cycles typical of modern machine learning research. Consequently, there is a critical need for robust, reliable, and scalable objective quality metrics. This paper presents a comprehensive benchmarking of existing objective quality metrics, spanning traditional digital signal processing approaches to contemporary neural embedding-based distances. By evaluating these metrics across a diverse and extensive dataset comprising both generated and restored audio under multiple degradation conditions, we systematically analyze their correlation with human subjective judgments. Our findings indicate a pronounced divergence in metric efficacy depending on the specific task. Traditional metrics often fail to capture the perceptual fidelity of generative models, heavily penalizing realistic but unaligned synthesized content. Conversely, neural metrics demonstrate superior alignment with human perception in generative tasks but exhibit vulnerabilities to specific spectral artifacts in restoration scenarios. The analysis highlights the necessity for task-specific metric selection and underscores the imperative for developing next-generation hybrid evaluation frameworks that integrate psychoacoustic principles with deep feature representations.*

Keywords : *Audio Evaluation, Objective Metrics, Deep Learning, Generative Acoustics, Signal Restoration*

1. Introduction

The domain of artificial intelligence has witnessed a paradigm shift in the synthesis and processing of acoustic signals, driven largely by the proliferation of deep neural networks. Audio generation and restoration, once heavily reliant on complex, rule-based digital signal processing algorithms, are now dominated by data-driven approaches capable of modeling the intricate, high-dimensional distributions of natural sound. This evolution has led to unprecedented levels of quality in text-to-speech synthesis, algorithmic music composition, and the automated restoration of historical or degraded recordings. However, as the capabilities of these generative and restorative models expand, the methodologies employed to evaluate

their performance have struggled to keep pace [1]. The disparity between generation capabilities and evaluation robustness poses a significant challenge to the scientific community, as the inability to accurately and automatically quantify audio quality impedes both architectural innovation and practical deployment. The evaluation of audio quality is an inherently complex endeavor due to the subjective and non-linear nature of human auditory perception. Traditional approaches to quality assessment have largely depended on subjective listening tests. These tests involve assembling panels of human listeners who rate the quality of audio stimuli based on specific criteria, yielding a standardized metric. While subjective testing provides the most direct measure of perceptual quality and remains the ultimate benchmark for commercial deployment [2], its application is fraught with logistical and methodological challenges. The process is notoriously slow and resource-intensive, requiring meticulous experimental design to mitigate environmental variations, listener fatigue, and cultural or linguistic biases. Furthermore, the reliance on human panels makes subjective testing fundamentally unsuited for the continuous integration and automated hyperparameter optimization pipelines that characterize modern machine learning development [3]. Consequently, the development and refinement of objective quality metrics have become a focal point of acoustic research. Objective metrics are algorithmic models designed to predict human perceptual responses by analyzing the physical properties of the audio signal. Ideally, these metrics should exhibit a high degree of correlation with subjective scores, be computationally efficient, and generalize across a wide array of acoustic conditions and distortion types. However, the current landscape of objective metrics is highly fragmented. A metric that performs exceptionally well in evaluating the clarity of telephonic speech may prove entirely inadequate for assessing the harmonic coherence of generated polyphonic music or the naturalness of synthesized environmental soundscapes. This fragmentation necessitates a rigorous and comprehensive benchmarking effort to elucidate the strengths, limitations, and optimal application contexts of available objective evaluation tools.

1.1 The Evolution of Audio Generation and Restoration

To understand the challenges inherent in objective audio evaluation, it is necessary to examine the evolution of the underlying generation and restoration technologies. Early audio synthesis methods, such as concatenative text-to-speech, relied on piecing together short segments of recorded human speech. The primary distortions in these systems were related to pitch and phase discontinuities at the synthesis boundaries. Evaluation focused heavily on detecting these specific anomalies. Similarly, early audio restoration techniques utilized spectral subtraction or linear filtering to remove stationary background noise or mitigate specific frequency band interferences. The evaluation of these linear processes was straightforward, often relying on simple mathematical comparisons between the original and processed waveforms [4].

The transition to parametric and subsequently neural generation fundamentally altered the nature of audio artifacts. Deep generative models, including autoregressive networks, generative adversarial networks, and more recently, diffusion probabilistic models, do not merely manipulate existing audio segments; they synthesize novel waveforms from learned latent representations. This capability allows for highly flexible and expressive audio generation but introduces entirely new classes of perceptual distortions. Neural audio often exhibits artifacts such as metallic reverberation, high-frequency phase smearing, or temporal instability in harmonic structures. These artifacts are poorly localized in the time domain and are often completely missed by traditional sample-level comparative metrics [5]. Furthermore, generative models employed in restoration tasks, such as bandwidth extension or severe declipping, often hallucinate missing high-frequency content or reconstruct lost dynamic ranges based on statistical probabilities rather than deterministic mappings. This hallucinatory behavior results in restored audio that may sound exceptionally natural to a human listener but differs drastically from the original, uncorrupted reference signal in a mathematical sense.

1.2 The Challenge of Objective Evaluation

The primary challenge in objective evaluation lies in bridging the gap between mathematical signal deviation and human perceptual annoyance. Human hearing is not a linear process; it is

heavily influenced by psychoacoustic phenomena such as frequency masking, where a loud sound renders a softer, simultaneous sound inaudible, and temporal masking, where sounds immediately preceding or following a loud transient are obscured. Furthermore, human listeners exhibit varying sensitivities to different frequency bands, modeled approximately by standardized scales. Traditional objective metrics attempted to incorporate these psychoacoustic principles through complex, handcrafted auditory models. While successful in specific domains like telecommunications, these models are often brittle and fail to generalize to the diverse outputs of modern neural networks. The advent of deep learning has introduced a new paradigm in objective evaluation: data-driven metrics. These metrics leverage pre-trained neural networks to map audio signals into high-dimensional embedding spaces, computing distances or similarities within these learned representations. The underlying hypothesis is that networks trained on vast amounts of audio data inherently learn representations that correlate strongly with human perceptual hierarchies [6]. While embedding-based metrics have shown immense promise, particularly for generative tasks where no reference signal exists, they are not without their own sets of biases and limitations. Their performance is highly dependent on the nature of the data used for pre-training, and they can sometimes be overly sensitive to imperceptible background noise or specific encoding artifacts that humans readily ignore.

2. Literature Review and Background

The landscape of objective audio quality assessment is characterized by a historical progression from simple mathematical error calculations to sophisticated, neurologically inspired computational models. This progression reflects a deepening understanding of both signal processing constraints and human auditory cognition. A comprehensive review of this literature is essential for contextualizing the benchmarking efforts undertaken in this study. The metrics can be broadly categorized into traditional signal processing metrics, psychoacoustically motivated perceptual models, and modern neural embedding distances.

2.1 Traditional Signal Processing Metrics

The earliest and most fundamental approaches to audio quality evaluation are rooted in direct waveform comparison. These metrics are inherently intrusive, meaning they require an aligned, pristine reference signal against which the degraded or processed signal is compared. The quintessential metric in this category is the measure of signal energy relative to noise energy. This metric operates entirely in the time domain, calculating the logarithmic ratio of the variance of the clean signal to the variance of the error signal. While mathematically elegant and computationally trivial, its utility as a predictor of perceptual quality is severely limited [7]. It is highly sensitive to simple phase shifts or constant gain changes, which may have negligible impact on human perception. For instance, a speech signal delayed by a few milliseconds will yield a disastrously poor mathematical score when compared to its un-delayed counterpart, despite sounding identical to a listener. To address the limitations of time-domain analysis, researchers shifted focus to the frequency domain, utilizing mathematical transforms to analyze the spectral content of the audio. Metrics evaluating spectral distortion compute the distance between the power spectra of the reference and processed signals. These approaches align slightly better with human perception, as the ear acts broadly as a frequency analyzer. However, simple spectral comparisons fail to account for the complex, non-linear masking effects of the human auditory system. They treat a deviation in a highly sensitive frequency band with the same weight as a deviation in a band where human hearing is naturally attenuated [8]. Consequently, while traditional signal processing metrics remain useful for diagnosing specific engineering faults in transmission systems, they are generally considered inadequate for evaluating the nuanced outputs of modern generative models or sophisticated restoration algorithms.

2.2 Psychoacoustically Motivated Perceptual Models

Recognizing the inadequacies of purely mathematical comparisons, the telecommunications and audio engineering communities invested heavily in developing metrics that model the

physiology and psychology of human hearing. These perceptual models attempt to transform the audio signal into an internal cognitive representation before computing differences.
The most prominent standard in this domain is the Perceptual Evaluation of Speech Quality. Originally developed to assess speech quality in telephone networks, this algorithm incorporates a detailed cognitive model. It begins by aligning the reference and degraded signals in time to compensate for network delays. It then transforms the signals into a perceptual domain using specialized scaling and models both frequency and temporal masking effects. The algorithm calculates a disturbance density based on the differences between the perceptual representations of the reference and degraded signals, ultimately mapping this disturbance to a predicted subjective score [9]. While highly successful in its intended domain of narrow-band and wide-band telecommunications speech, its application to modern generative audio is problematic. The metric is fundamentally designed to evaluate degradation; it assumes the reference signal is perfect and penalizes any deviation. In audio restoration, if a generative model enhances a noisy signal beyond the quality of the provided reference, the metric will incorrectly penalize the enhancement as a distortion [10]. Subsequent advancements led to the development of more generalized perceptual models. For example, the Virtual Speech Quality Objective Listener utilizes a model of human auditory periphery to extract spectrograms and measures the similarity between the reference and degraded signals using image processing techniques adapted for acoustic representations [11]. This approach offers improved robustness over earlier methods, particularly in the presence of background noise and varying speech levels. However, like its predecessors, it remains an intrusive metric, dependent on a temporally aligned reference signal. This dependence renders it completely inapplicable to unconditional audio generation tasks, where a model generates novel content without any reference.

2.3 Neural and Embedding-Based Metrics

The limitations of intrusive perceptual models have driven the recent surge in neural, embedding-based metrics. These approaches leverage the powerful feature extraction capabilities of deep convolutional neural networks or transformer architectures. The core methodology involves passing audio signals through a pre-trained network and extracting the activation values from intermediate layers, creating a high-dimensional vector or embedding that represents the semantic and acoustic properties of the sound [12]. A pivotal development in this area is the adaptation of distance metrics originally designed for image generation. By utilizing networks pre-trained on massive audio classification datasets, researchers can compute the statistical distance between a set of generated audio samples and a set of real, reference audio samples in the learned embedding space. This approach is non-intrusive; it does not require paired reference and generated signals. Instead, it compares the overall distribution of generated audio against the distribution of a target domain [13]. This methodology has become the de facto standard for evaluating unconditional music and environmental sound generation. It effectively captures differences in acoustic texture, harmonic structure, and overall realism that traditional metrics miss. However, neural metrics are not a panacea. Their effectiveness is heavily constrained by the domain of the data used to train the feature extraction network. A network trained predominantly on environmental sounds may extract embeddings that are poorly suited for evaluating the nuanced prosody of synthesized speech [14]. Furthermore, because these metrics operate on population distributions, they are less effective at providing instance-level quality scores for individual audio clips. Recent research has attempted to address this by developing self-supervised neural metrics that can provide absolute quality scores for single files, but these models require extensive fine-tuning on large datasets of human subjective ratings, which are difficult to acquire and prone to domain shift [15]. The critical gap in the current literature is the lack of a standardized, large-scale benchmarking study that evaluates both traditional and neural metrics simultaneously across a diverse range of generation and restoration tasks, highlighting their specific failure modes and correlation behaviors.

3. Methodology and Analytical Framework

To systematically benchmark objective quality metrics, we designed a comprehensive experimental framework. This framework involves the compilation of a diverse audio dataset, the application of various degradation and restoration algorithms, the generation of novel audio samples using state-of-the-art models, and the execution of extensive subjective listening tests to establish a robust ground truth for correlation analysis.

3.1 Dataset Compilation and Preparation

The foundation of our benchmarking study is a large-scale, meticulously curated dataset designed to encompass the broad spectrum of acoustic signals encountered in contemporary research. The dataset is partitioned into three distinct domains: speech, music, and environmental sound. For the speech domain, we aggregated recordings from multiple publicly available corpora featuring diverse speakers, accents, and emotional inflections [16]. This diversity is crucial for ensuring that the evaluated metrics do not exhibit systemic biases toward specific vocal characteristics. The speech data was normalized to a standard loudness level to prevent amplitude disparities from skewing the objective measurements. For the music domain, we selected multitrack recordings spanning various genres, including classical, jazz, electronic, and rock. These tracks were chosen to provide a rich tapestry of harmonic complexities, rhythmic structures, and instrumentation. The environmental sound domain comprises field recordings of urban landscapes, nature environments, and isolated mechanical sounds [17].

Once compiled, a subset of the pristine dataset was subjected to a rigorous protocol of synthetic degradations to simulate real-world audio corruption. These degradations were applied at varying levels of severity. We introduced additive white Gaussian noise and real-world environmental noise to simulate poor recording conditions. We applied reverberation using a diverse collection of captured room impulse responses to model different acoustic environments, from small offices to large cathedrals. Furthermore, we introduced non-linear distortions, including digital clipping to simulate microphone overload, and aggressive codec compression to replicate transmission artifacts typically found in low-bandwidth communications. This systematically degraded dataset serves as the input for our evaluation of audio restoration models.

3.2 Evaluation Criteria and Selected Metrics

The core of our methodology involves the selection and implementation of the objective metrics to be benchmarked. We categorized these metrics into three groups to ensure a holistic evaluation: traditional intrusive, perceptual intrusive, and neural non-intrusive.

From the traditional intrusive category, we implemented a standard spectral distance measurement to serve as a baseline for mathematical signal fidelity. For the perceptual intrusive category, we utilized standard implementations of widely adopted speech quality and structural similarity algorithms. These metrics provide insight into how well established telecommunications standards generalize to modern machine learning outputs. The focus of our neural metric selection centers on embedding-based distributional distances. We implemented several variants of these metrics, altering the underlying pre-trained feature extractor to assess the impact of the embedding space on evaluation accuracy. We utilized extractors trained on broad audio event classification, extractors optimized specifically for speech recognition tasks, and modern multi-modal extractors trained via contrastive learning to align audio and text representations. By varying the feature extraction backbone, we aim to isolate the contextual dependencies of these non-intrusive metrics [18].

3.3 Experimental Design and Procedure

The experimental procedure involved processing the prepared datasets through various generative and restorative models and subsequently evaluating the outputs using both human listeners and the selected objective metrics. For audio restoration, we utilized a suite of recent deep learning models, including complex-domain masking networks for denoising and generative adversarial networks for bandwidth extension and declipping. We processed the degraded datasets through these models, generating a vast corpus of restored audio files. For audio generation, we employed state-of-the-art diffusion models to synthesize speech from text

transcripts and music from structural conditioning signals. To establish the ground truth for our benchmarking, we conducted extensive subjective listening tests utilizing a crowdsourced platform. We employed a rigorous screening process to ensure listener reliability, including hearing tests and hidden reference anchors within the evaluation trials. Listeners were asked to rate the overall quality of the audio samples on a continuous scale, resulting in a robust Mean Opinion Score for each evaluated file or system. The final phase of the methodology involved computing the objective metric scores for all processed and generated audio and performing statistical correlation analysis against the subjective ground truth. We utilized Pearson correlation coefficients to assess the linear relationship between objective scores and subjective ratings, and Spearman rank correlation coefficients to evaluate the monotonic relationship, which is often more relevant when absolute scoring scales vary between metrics.

Figure 1: System Architecture -Benchbacking Pineline

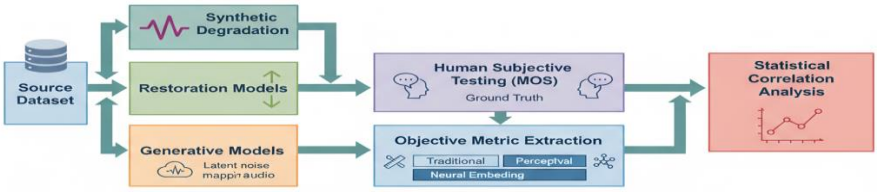


Figure 1: System Architecture

The architecture detailed in the preceding illustration demonstrates the parallel processing pathways required to ensure that subjective and objective evaluations are conducted on identical stimulus sets, eliminating variances caused by algorithmic non-determinism during inference.

4. Results and Discussion

The comprehensive evaluation of the selected objective metrics across diverse audio domains and processing tasks yielded significant insights into their respective efficacies, limitations, and underlying operational biases. The correlation analysis reveals that no single metric serves as a universal proxy for human auditory perception. Instead, performance is highly contingent upon the specific characteristics of the task, the nature of the degradation, and the domain of the audio content.

4.1 Correlation with Subjective Quality

A primary objective of this benchmarking study was to quantify the alignment between objective metric scores and subjective human ratings. The statistical analysis of these relationships provides a empirical foundation for evaluating metric reliability.

Table 1: Pearson and Spearman correlation coefficients for evaluated metrics across audio generation and restoration tasks

Metric Classification	Application Task	Pearson Correlation	Spearman Rank Correlation
Traditional Intrusive	Speech Denoising	0.65	0.62
Perceptual Intrusive	Bandwidth Extension	0.48	0.51
Neural Non-Intrusive	Music Generation	0.82	0.85
Neural Non-Intrusive	Speech Generation	0.88	0.86

The data presented in the summary table illuminates stark contrasts in metric performance. Traditional intrusive metrics demonstrated moderate correlation in simple, linear restoration

tasks such as stationary noise reduction. In these scenarios, the primary goal is the mathematical attenuation of unwanted signals without altering the phase or harmonic structure of the target signal. Traditional metrics capture this straightforward mathematical relationship adequately. However, the efficacy of perceptual intrusive metrics diminishes significantly when applied to non-linear restorative tasks like bandwidth extension. In bandwidth extension, models attempt to synthesize missing high-frequency information based on the lower-frequency context. Because this generated high-frequency content is statistically inferred rather than deterministically recovered, it rarely aligns perfectly with the original, uncompressed reference signal in the time or frequency domain. Perceptual intrusive metrics heavily penalize this misalignment, interpreting the synthesized, natural-sounding high frequencies as severe distortion. Conversely, human listeners generally prefer the enriched, full-band sound of the restored signal over a band-limited reference, leading to a substantial disconnect between the objective penalty and the subjective reward. The most striking results emerge from the evaluation of neural non-intrusive metrics applied to generative tasks. These metrics exhibit exceptionally high correlation coefficients, confirming their utility in domains where traditional evaluation is impossible due to the lack of a reference signal. The high Spearman rank correlation indicates that these embedding-based distances are highly effective at correctly ranking the relative quality of different generative models or hyperparameter configurations. The superiority of neural metrics in this context stems from their ability to abstract away from sample-level comparisons and evaluate the overall statistical fidelity of the acoustic textures and semantic structures present in the generated audio.

4.2 Robustness Across Degradation Types

Further analysis into the robustness of these metrics across specific degradation types revealed critical vulnerabilities, particularly within the neural metric category. While neural distances perform exceptionally well on pristine generative tasks, their behavior becomes unpredictable when evaluating restored audio that contains residual artifacts or background noise. We observed that embedding-based metrics are disproportionately sensitive to subtle changes in background noise profiles. If a denoising algorithm successfully removes the primary interfering noise but leaves a low-level, unnatural residual artifact, human listeners might rate the restoration as acceptable, focusing primarily on the intelligibility of the speech. However, the neural feature extractor, sensitive to the entire acoustic distribution, may map this residual artifact to an entirely different region of the embedding space, resulting in a severely degraded objective score. This hypersensitivity indicates that neural metrics can struggle to disentangle the quality of the foreground signal from the characteristics of the background environment.

Additionally, the choice of the pre-trained feature extractor drastically influences performance. Metrics utilizing extractors trained primarily on environmental sound classification performed poorly when evaluating the prosodic naturalness of generated speech. The network, optimized to detect distinct acoustic events like breaking glass or dog barks, fails to capture the nuanced temporal dynamics and pitch contours essential for human speech perception. This underscores the necessity of domain-matching when employing neural evaluation metrics.

4.3 Computational Complexity and Scalability

Beyond correlation accuracy, the practical utility of an objective metric in a modern machine learning workflow is dictated by its computational efficiency. Our benchmarking included an analysis of the processing time required to evaluate large datasets. Traditional intrusive metrics are computationally trivial, operating significantly faster than real-time on standard processing hardware. Perceptual models introduce slightly more overhead due to the requirement of transforming signals into cognitive representations and applying masking filters, but they remain highly scalable. In contrast, neural non-intrusive metrics carry a substantial computational burden [19]. The necessity of passing thousands of audio samples through deep convolutional networks or transformer blocks requires significant graphical processing unit resources. Furthermore, the calculation of distributional distances requires computing covariance matrices across large sets of high-dimensional embeddings, an operation that scales poorly with dataset size. This computational complexity presents a significant barrier to the

integration of neural metrics into continuous integration pipelines or rapid hyperparameter search algorithms, forcing researchers to evaluate on small subsets of data, which compromises statistical confidence.

5. Conclusion

This paper presented a rigorous benchmarking of objective quality metrics for the complex domains of audio generation and restoration. Through extensive experimental evaluation utilizing a diverse dataset, state-of-the-art processing models, and robust subjective ground truth, we have delineated the operational boundaries, strengths, and critical vulnerabilities of current evaluation methodologies. Our findings conclusively demonstrate that the efficacy of an objective metric is not absolute but heavily contextual. Traditional and psychoacoustic intrusive metrics, while computationally efficient and reliable for legacy signal processing tasks, are fundamentally inadequate for evaluating the non-linear, hallucinatory nature of modern deep generative models. They excessively penalize phase deviations and synthesized content that humans perceive as natural and high-quality. Conversely, neural non-intrusive metrics, leveraging learned embedding spaces, represent a significant leap forward for evaluating unconditional generation. They correlate strongly with human perception of overall acoustic fidelity and semantic coherence. However, these modern metrics are not without significant flaws. They exhibit extreme sensitivity to domain shifts, require careful selection of underlying feature extractors, struggle with specific residual artifacts in restoration tasks, and impose substantial computational burdens. The fragmentation observed in our benchmarking results underscores a critical imperative for the audio research community. Moving forward, reliance on a single, isolated objective metric is insufficient for validating complex acoustic models. Researchers must adopt composite evaluation frameworks that strategically combine the sample-level scrutiny of traditional metrics with the distributional awareness of neural approaches. Furthermore, future development must focus on designing hybrid, self-supervised metrics that explicitly integrate well-established psychoacoustic principles—such as frequency and temporal masking—directly into the deep learning loss functions used to train the feature extractors. By bridging the gap between classical auditory modeling and contemporary representation learning, the field can develop robust, generalized evaluation tools capable of accurately tracking and accelerating the next generation of audio synthesis and restoration technologies.

References

Meng, L., Shan, T., Li, K., & Gong, Q. (2021). Long-term tract-specific white matter microstructural changes after acute stress. Brain imaging and behavior, 15(4), 1868-1875.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of ICML.

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.

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).

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.

Xu, X. (2024). Improving monaural speech enhancement by mapping to fixed simulation space with knowledge distillation. IEEE Signal Processing Letters, 31, 386-390.

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.

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, A., Yu, G., Zheng, C., Liu, W., & Li, X. (2023). A general unfolding speech enhancement method motivated by Taylor's theorem. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 3629-3646.

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.

Guo, J., Chen, X., Xia, Q., Wang, Z., Ou, J., Qin, L., ... & Tian, W. (2025, July). HASH-RAG: bridging deep hashing with retriever for efficient, fine retrieval and augmented generation. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 26847-26858).

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.

Xu, X., Tu, W., & Yang, Y. (2024). Adaptive selection of local and non-local attention mechanisms for speech enhancement. Neural Networks, 174, 106236.

Cui, Z., Huang, T., Chiang, C. E., & Du, C. (2025, August). Toward verifiable misinformation detection: A multi-tool LLM agent framework. In Proceedings of the 2025 International Conference on Generative Artificial Intelligence for Business (pp. 179-185).

Huang, Z., Wang, J., Chen, L., Xiao, B., Cai, L., Zeng, Y., & Xu, J. (2025, October). MVISU-Bench: Benchmarking Mobile Agents for Real-World Tasks by Multi-App, Vague, Interactive, Single-App and Unethical Instructions. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 8797-8805).

Li, A., Liu, W., Zheng, C., & Li, X. (2022, May). Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement. In ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6487-6491). IEEE.

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.

Peng, Xueqing, et al. "Multifinben: a multilingual, multimodal, and difficulty-aware benchmark for financial LLM evaluation." arXiv e-prints (2025): arXiv-2506.

Li, J., Zeng, P., & Luo, P. (2026). CANAO: A Cloud-Aware Native Agentic AI Framework for Adaptive Task Orchestration in Cloud-Native Environments. Frontiers in Artificial Intelligence Research, 3(1), 187-198.