eng

MDPI

Review

Advances in Speech Enhancement: A Comprehensive Review of Noise Suppression Techniques

Pushpraj Tanwar *, Ajay Somkuwar and Rakesh Kumar Gumasta

Department of Electronics and Communication Engineering, Maulana Azad National Institute of Technology, Bhopal 462003, India

* Correspondence: pushpraj2404@gmail.com

Abstract

Over the past several decades, numerous methods have been developed to improve the signal-to-noise ratio, perceptual quality, and intelligibility of speech. In practice, no single method is universally optimal, as each category exhibits distinct strengths and limitations under specific acoustic conditions. The proposed taxonomy classifies speech enhancement methods according to their dominant signal modeling and enhancement mechanism, ranging from the classical signal processing approaches to modern machine learning techniques and advanced hybrid frameworks. The mathematical formulation illustrates the theoretical foundations of the different approaches, providing a clear understanding of their underlying principles and the evolution of performance across successive generations of speech enhancement methods. The comparative analysis demonstrates that statistical and subspace-based approaches offer low operational complexity but show limitations under highly nonstationary acoustic conditions. Adaptive filtering, transform-domain, and speech model-based methods exploit temporal, spectral, and speech production characteristics to achieve improved noise suppression, while perceptual methods improve subjective listening quality by incorporating psychoacoustic principles, thereby providing a more natural and intelligible listening experience. Machine learning-based techniques achieve excellent performance; however, they incur higher computational complexity and substantial training requirements. More recently, hybrid approaches have integrated complementary techniques from multiple paradigms, achieving robust performance under adverse acoustic environments. Overall, this review provides a unified perspective on speech enhancement techniques, identifies their strengths, and highlights emerging research opportunities for the development of robust, efficient, and intelligent speech enhancement systems.

Keywords: speech enhancement; noise reduction; signal-to-noise ratio; perceptual quality; intelligibility

Check for updates

Academic Editors: Gian Giuseppe Soma, Jia Chen, Jun Yang and Wenchao Huang

Received: 10 June 2026
Revised: 4 August 2026
Accepted: 29 August 2026
Published: 9 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Speech communication plays a vital role in information exchange and facilitates effective human–machine interaction. It forms the backbone of numerous real-world applications, including telecommunications, teleconferencing, hearing aids, voice-assisted intelligent systems, speech coding, and human–computer interaction [1–3]. However, in practical environments, speech signals are often contaminated by different forms of noise that distort the original signal and degrade speech quality and intelligibility [4–6]. Noise sources may be stationary, such as white noise (static), pink noise, or steady engine noise, or they may be nonstationary, such as babble noise, crowd noise, impulsive disturbances, or

background crowd [7,8]. Stationary noise has a constant mean, variance, and autocorrelation over time, resulting in predictable statistical behavior. In contrast, nonstationary noise varies continuously over time, making it highly complex and unpredictable in nature. Noisy environments significantly degrade the overall performance of communication systems. In speech signal processing, noise suppression is a fundamental concern for improving speech quality in harsh acoustic environments. Consequently, suppressing such noise has become a major research focus in speech signal processing. The primary objective is to attenuate undesired noise components while preserving important speech features; however, this remains challenging due to the diverse and dynamic nature of real-world acoustic environments. Earlier classical methods are based on the statistical assumptions of noisy signals and provide a computationally simple solution. Such techniques are analytically efficient for stationary noise but are not effective under rapidly varying noisy environments. To compensate for the limitations of earlier methods, more advanced speech enhancement techniques have been developed based on different processing principles [9–11]. Transform-domain methods represent speech signals in different domains where speech and noise components can be more easily separated. This time-frequency representation allows localized processing of speech and noise and enables frequency-dependent noise attenuation. Speech model-based methods exploit prior knowledge of speech characteristics and rely on accurate model development; however, their performance deteriorates if there is a mismatch between the developed model and the real speech signal. More recently, deep learning approaches not only improve the signal-to-noise ratio but also enhance perceptual quality while preserving temporal dynamics and formant structure [12–14]. To overcome the limitations of individual methods, a hybrid speech enhancement approach has emerged as a promising alternative. It integrates the strength of classical statistical methods with modern data-driven approaches to develop a robust and unified framework.

Several comprehensive surveys on speech enhancement have been reported in the literature, each addressing specific aspects of noise suppression, such as statistical signal processing, adaptive filtering, transform-domain techniques, deep learning, or hybrid approaches. However, most existing reviews primarily concentrate on a limited subset of enhancement algorithms or emphasize recent developments in deep learning without providing a unified perspective covering the entire evolution of speech enhancement. Moreover, comparatively few studies systematically correlate the underlying mathematical principles, representative algorithms, computational complexity, performance characteristics, and practical application scenarios within a single framework [15–18]. To address these limitations, the present review proposes a unified taxonomy based on dominant signal modeling and enhancement mechanisms. The proposed framework integrates classical signal processing, auditory-inspired methods, spatial processing, data-driven learning, and hybrid approaches within a single coherent classification. In addition, the review provides a mathematical interpretation of representative algorithms along with a cross-category comparative analysis. It also presents a critical analysis of the strengths and limitations of each category, together with structured future research directions for speech enhancement.

The vast field of speech processing encompasses numerous techniques for noise reduction. The comprehensive classification of speech enhancement methods, ranging from statistical methods to hybrid frameworks, explores their underlying working mechanism, strengths, limitations, and key advantages. The study emphasizes that selecting an appropriate strategy depends on the specific needs of the application. Depending on the underlying principle, speech enhancement methods are categorized into five major domains. Some algorithms may incorporate concepts from multiple domains; for example, transform-domain statistical filtering or perceptually weighted deep learning. Therefore, each method is classified according to its primary operational principle. This well-defined

classification identifies the research gaps between existing techniques and provides direction for future research. The key contributions of this literature review are as follows:

- ✓ Unified taxonomy: A unified taxonomy is proposed that systematically classifies speech enhancement algorithms according to their dominant signal modeling and enhancement mechanisms, providing a consistent framework for understanding the evolution and relationship among classical signal processing, auditory-inspired methods, spatial processing, data-driven techniques, and hybrid approaches.
- ✓ Comprehensive framework: Unlike many existing surveys that primarily focus on algorithm descriptions, this review presents a comprehensive mathematical interpretation of representative techniques together with their theoretical foundations, enabling readers to understand the underlying principles governing each method.
- ✓ Cross-domain performance: A cross-category comparative analysis is presented to analyze the strengths, limitations, computational complexity, robustness, and application applicability of different enhancement paradigms, thereby facilitating informed algorithm selection for various practical scenarios.
- ✓ Hybrid framework: The review discusses recent advances in hybrid frameworks that combine conventional signal processing with modern deep learning techniques to improve speech quality under challenging acoustic environments.
- ✓ Future research directions: Emerging research directions are identified, including end-to-end optimization, self-supervised learning, explainable artificial intelligence, foundational models, lightweight edge implementation, and multimodal speech enhancement, providing a roadmap for future research.

2. Performance Index

In practical communication systems, speech signals may be susceptible to various acoustic disturbances arising from environmental noise, transmission channels, recording devices, and acoustic propagation conditions. From a signal modeling perspective, these degradations can be broadly categorized into additive and multiplicative distortions, each exhibiting distinct mathematical characteristics and requiring different enhancement strategies. The majority of conventional speech enhancement methods assume an additive observation model, as shown in Equation (1). This model accurately represents most real-world acoustic interferences encountered in practical communication systems, including white Gaussian noise, babble noise, traffic noise, factory noise, and crowd interference. Typical frequency characteristics and relative energy distribution of common environmental noises encountered in speech enhancement applications are summarized in Table 1. Depending on their statistical characteristics, these noises may be classified as stationary or nonstationary. Stationary noises exhibit relatively constant statistical properties over time and are comparatively easier to estimate and suppress. In contrast, nonstationary noise exhibits continuously varying temporal and spectral characteristics, frequently overlaps with speech components, and therefore constitutes the principal challenge in modern speech enhancement.

In addition to additive environmental noise, speech signals may also undergo multiplicative or signal-dependent distortions. A simplified multiplicative degradation can be expressed as $y(n) = h(n)s(n)$, where $h(n)$ represents a time-varying gain, fading coefficient, or other signal-dependent disturbances. In practical systems, additive and multiplicative degradation may occur simultaneously and can therefore be represented by the generalized model $y(n) = h(n)s(n) + v(n)$. The fundamental difference is that additive noise introduces an additional unwanted component without directly modifying the underlying speech amplitude, whereas multiplicative degradation scales or modulates the speech signal itself. Consequently, its effect depends on both the instantaneous characteristics of

the speech signal and the variation in the multiplicative factor $h(n)$. This signal-dependent interaction makes the estimation and suppression of multiplicative disturbances considerably more challenging because the unwanted component cannot generally be isolated and directly subtracted in the same manner as additive noise. Multiplicative effects may arise in communication systems because of time-varying channel gain, fading, automatic gain variations, amplitude fluctuations, or modulation-related distortions. However, not all channel-induced speech degradation can be described adequately by a pure multiplicative model. For example, acoustic reverberation is more appropriately represented as a convolution between the clean speech signal and the room impulse response, $y(n) = h(n) * s(n) + v(n)$, where * denotes convolution. Similarly, microphone frequency-response effects and propagation characteristics may introduce filtering or convolutional distortion rather than the simple sample-wise multiplication. Therefore, additive, multiplicative, and convolutional degradation should be distinguished according to their physical origin and mathematical interaction with the speech signal. This review primarily focuses on additive noise because it represents the most common degradation encountered in telecommunication, hearing aids, speech coding, automatic speech recognition, and mobile communication systems. While multiplicative and channel-induced distortions are discussed as related degradation mechanisms, their mitigation generally requires specialized techniques such as channel estimation, equalization, dereverberation, or joint enhancement strategies.

Table 1. Typical frequency characteristics and relative energy distribution of common environmental noises encountered in speech enhancement applications.

Noise Type	Frequency Range	Relative Energy	Typical Application
White Noise	Entire spectrum (0–Fs/2)	Uniform	Laboratory
Pink Noise	Low-frequency dominant	Medium	Environmental
Babble Noise	300–4000 Hz	High	Crowd
Traffic Noise	20–1200 Hz	High	Outdoor
Engine Noise	20–800 Hz	Very High	Vehicles
Machinery	100–5000 Hz	High	Industrial
Wind Noise	<500 Hz	Moderate	Outdoor microphones

Human speech primarily occupies frequencies below 8 kHz (telephone speech: 300–3400 Hz; wideband speech: approximately 50–8000 Hz). Most environmental noises overlap significantly with this frequency range, making complete noise suppression difficult without affecting speech components. In practice, there is no universal upper energy bound because noise energy depends on acoustic environment and source intensity; consequently, speech enhancement algorithms typically estimate noise statistics adaptively rather than assuming fixed energy limits.

To measure the effectiveness of speech enhancement algorithms, performance evaluation metrics are essential. This section provides a brief overview of the most commonly used metrics for assessing speech quality, intelligibility, noise suppression capability, and robustness. No single evaluation metric can comprehensively characterize speech enhancement performance; therefore, multiple complementary objective and subjective measures are required to assess signal quality, perceptual characteristics, and intelligibility. These evaluation metrics facilitate a fair and meaningful comparison among different speech enhancement methods.

For signal assessment, the speech model is initially modeled as shown in Figure 1.

$$y(n) = s(n) + v(n) \tag{1}$$

where $s(n)$ is the clean speech signal in the time domain, $v(n)$ is additive noise, and $y(n)$ is the noisy signal. The frequency domain representation of the signal [19]

$$Y(k,m) = S(k,m) + V(k,m) \tag{2}$$

where $S(k,m)$ is the magnitude spectra of the clean speech signal, $V(k,m)$ is the magnitude spectra of additive noise, and $Y(k,m)$ is the magnitude spectrum of the noisy speech signal. After applying the speech enhancement algorithm for noise suppression, the recovered speech signal is further evaluated. The signal-to-noise ratio (SNR) is the most widely used objective metric for measuring the relative strength of the recovered speech signal [20].

$$SNR\ (\text{dB}) = 10log10\frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N}(s(n) - \hat{s}(n))^2} \tag{3}$$

where $\hat{s}(n)$ is the enhanced speech signal, n is the discrete time sample index, and N is the total number of samples. The speech-to-noise ratio (IS/TN ratio) evaluates the selective suppression of transient noise [21]:

$$\frac{IS}{WS}Ratio\ (\text{dB}) = 10log_{10}\frac{\sum_{\eta \in S} s^2(n)}{\sum_{\eta \in T} v^2(n)} \tag{4}$$

Log-Spectral Distance (LSD) measures the spectral distortion between clean and enhanced speech signals [22]. A lower value of LSD indicates better preservation of spectral information.

$$LSD = \sqrt{\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\left[10log_{10}\left(|S(k,m)|^2\right) - log_{10}\left(|\hat{S}(k,m)|^2\right)\right]^2} \tag{5}$$

where $S(k,m)$ is the clean speech spectrum $\hat{S}(k,m)$ is the enhanced speech spectrum, k is the frequency bin index, m is the time frame index, K is the total number of frequency bins, and M is the total number of frames. The speech distortion index (SDI) quantitatively measures the amount of speech distortion introduced by the noise reduction algorithm [23]. A lower SDI indicates better speech preservation.

$$SDI = \frac{\sum_{n=1}^{N}(s(n) - \hat{s}(n))^2}{s^2(n)} \tag{6}$$

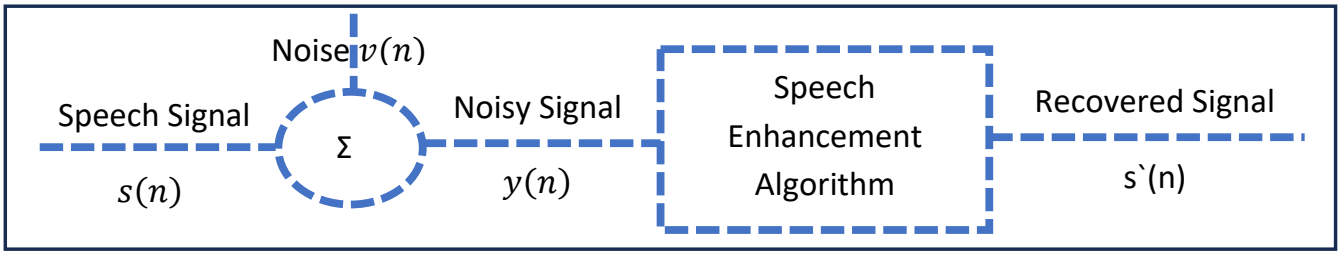


Figure 1. Flow diagram of speech enhancement algorithm.

Different evaluation metrics emphasize different aspects of speech enhancement performance and therefore should be selected according to the target application. SNR improvement and SI-SDR measure the effectiveness of noise suppression and signal reconstruction.

In addition, the Perceptual Evaluation of Speech Quality (PESQ) is a widely accepted objective metric for predicting human perception of speech quality. PESQ scores range from −0.5 to 4.5, with higher scores indicating better perceptual quality [24]. The Mean Opinion Score (MOS) is a standard subjective analysis metric for perceptual quality, based on ratings provided by human listeners on a scale of 1 to 5. PESQ and MOS are primarily used to assess perceptual speech quality in telecommunications, hearing aids, and multimedia applications. Furthermore, PESQ is often correlated with MOS to provide a more comprehensive evaluation of speech quality [25]. The short-time objective intelligibility (STOI) metric is widely adopted to quantify speech intelligibility, especially for hearing-impaired communication and automatic speech recognition front-end evaluation. STOI determines the speech intelligibility by comparing the short-time temporal envelopes of the enhanced signal and the original clean speech signal. It demonstrates a strong correlation with human intelligibility scores within the range of 0 to 1, where the value closer to one indicates better intelligibility [26]. In automatic speech recognition systems, word error rate (WER) evaluates the influence of enhancement algorithms on speech recognition performance. WER measures the percentage of words that are incorrectly recognized compared to the reference transcript.

$$WER = \frac{S + D + I}{N} \times 100\% \tag{7}$$

where S is the number of substitutions (incorrect words), D is the number of deletions (missing words), I is the number of insertions (extra words), and N is the total number of words in the reference transcript. For speech enhancement algorithms intended for automatic speech recognition applications, lower WER indicates better preservation of speech intelligibility and linguistic information. However, computational complexity and latency are equally important for real-time implementation in communication systems. Table 2 summarizes the performance evaluation parameters.

Table 2. Performance evaluation parameters to measure the speech quality.

Parameter	Measures	Range	Key Advantage
SNR/ΔSNR	Noise suppression	dB	Simple, intuitive
PESQ	Perceptual quality	−0.5 to 4.5	human-centric evaluation
STOI	Intelligibility	0 to 1	High correlation
MOS	Subjective quality	1 to 5	Human judgment
LSD	Spectral distortion	≥ 0	Model evaluation

3. Comprehensive Classification of Noise Suppression Algorithms

A speech signal in practical application is affected by a variety of noisy environments, and noise reduction algorithms are therefore required to recover the original speech signal. No single technique performs optimally under all noisy conditions, as different algorithms follow different processing strategies. Depending on the underlying principle, the existing methods are systematically categorized within a unified framework. Methods ranging from classical statistical approaches to modern data-driven techniques are classified in Table 3 and comparatively summarized in Table 4, highlighting their fundamental mechanisms, strengths, and suitability for different applications. This section reviews the evolution of speech enhancement methods from the earliest classical approach to the most recent advances.

Table 3. Classification of speech noise reduction methods.

Category	Underlying Principle	Major Algorithms	Key Strengths	Limitations
Statistical	Treats speech/noise as stochastic processes; estimates noise statistics and applies frequency-dependent gains	Spectral subtraction, Wiener filter, MMSE-LSA	Low complexity, Analytical gain functions are interpretable, perform well for stationary noise	Musical noise, under-estimation of noise produces perceptual artifacts
Subspace	Assumes clean speech occupies a low-dimensional subspace	PCA, SVD, ICA	Effective separation for structured/coloured noise. Preserves major speech components	Rank selection & computational cost. covariance estimation is computationally heavier
Adaptive filtering	Iteratively adapts filter coefficients to minimize error using a reference correlated noise input	LMS, NLMS, RLS	Excellent tracking of non-stationary interference when reference available	Requires reference, Performance collapses without a reliable reference
Transform-domain	Represents signals in time–frequency or time–scale domains where speech is sparse and localized	STFT, Wavelet thresholding, EMD/VMD	Handles non-stationarity and transients. Localized processing preserves transients and harmonics	phase reconstruction often neglected causing artifacts; computational overhead for multi-resolution methods
Model-based	Uses explicit speech production/state models to predict clean speech dynamics	LPC, AR modelling, Kalman filtering	Preserves formants and temporal structure. Strong priors on speech dynamics enable recovery even under low SNR	Requires accurate parameter estimation and higher computational cost
Perceptual/Auditory	Uses human auditory models (critical bands, masking) to suppress only perceptually audible noise	Bark/Mel filter banks, auditory masking gain controllers	High subjective quality and intelligibility preservation. Avoids over-suppression, resulting in natural sound	Psychoacoustic models are complex, objective metrics may not reflect perceptual improvements
Machine learning	Learns nonlinear mappings from noisy to clean speech	DNN, CNN, LSTM, Transformers models	State-of-the-art performance across noise types. Excellent at modelling complex nonlinearities, large gains in PESQ/STOI	Requires large labelled datasets and heavy compute, low interpretability
Hybrid	Combines complementary paradigms to exploit strengths and mitigate weaknesses	Wavelet + DNN, EMD + PCA, DNN + Kalman	Robustness & balanced trade-offs. improve interpretability, and enhance performance	Design complexity & integration issues. increased latency

The proposed taxonomy (Figure 2) classifies speech enhancement methods according to their dominant signal modeling and enhancement mechanisms, including statistical estimation, subspace projection, adaptive optimization, transform-domain representation, speech production modeling, perceptual and auditory-inspired modeling, data-driven learning, and hybrid integration.

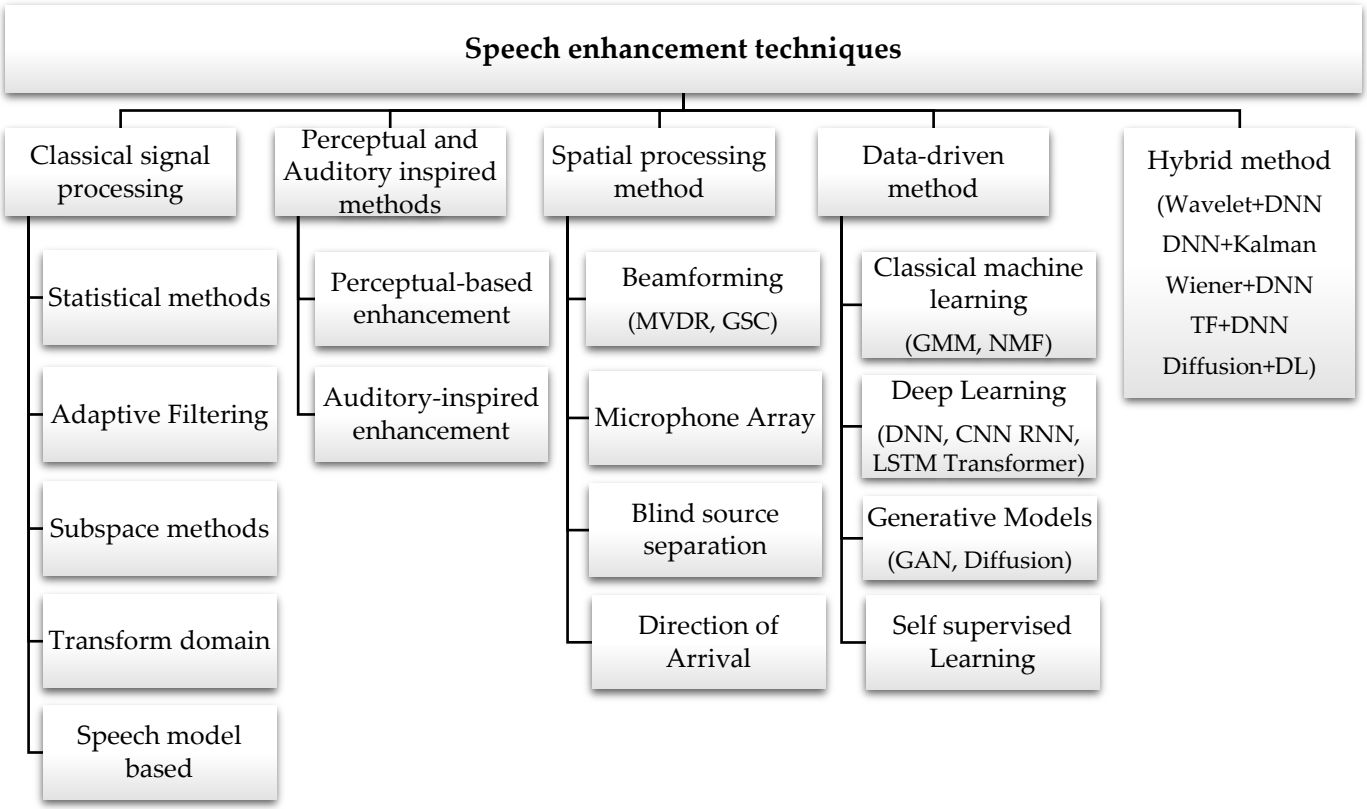


Figure 2. Classification of speech enhancement methods according to their dominant signal modeling and enhancement mechanism.

3.1. Classical Signal Processing

Classical signal processing forms the foundation of the earliest speech enhancement methods for noise suppression, primarily based on mathematical modeling, statistical estimation, signal decomposition, and optimization strategies rather than advanced data-driven learning. Depending on the underlying strategies, this broad category can be re-classified into statistical estimation, subspace projection, adaptive optimization, transform-domain representation, and speech model-based methods. These techniques are fundamentally based on the inherent characteristics of speech and noise signals to estimate clean speech components while suppressing background noise.

Most classical speech enhancement algorithms are developed under a common signal model in which the observed noisy speech signal is represented as the sum of clean speech and additive noise. These models generally assume that the background noise is stationary or slowly varies over short analysis frames, enabling reliable estimation of statistical characteristics under these assumptions. Various estimation techniques derive optimal gain functions to suppress noise while preserving speech. Although these assumptions simplify algorithm design and enable efficient real-time implementation, their validity decreases in rapidly changing acoustic environments, thereby motivating the development of more advanced techniques.

3.1.1. Statistical Model

The statistical approaches are among the earliest and most analytically grounded frameworks, based on the principle of statistical modeling of speech and noise signals in the time-frequency domain. The background noise is estimated independently, assuming it is additive in nature, slowly varying compared to the speech signal, and uncorrelated with speech. Noise suppression is formulated as a statistical estimation problem in which

the clean speech spectrum is estimated from noisy observations by minimizing the statistical cost function, such as log-spectral distortion, mean square error, etc., which rely on probability theory rather than explicit speech models. In the statistical approach, the noisy speech signal is first segmented into overlapping frames (framing) and then transformed into the frequency domain using the short-time Fourier transform (STFT). Next, the power spectral density (PSD) of the noise is estimated during speech-absent intervals using voice activity detection (VAD) or the minimum statistics method. The gain function is then computed using various algorithms, including spectral subtraction, Wiener filtering, and MMSE estimation, all of which rely heavily on accurate estimation of noise statistics.

The first practical frequency domain noise suppression technique was introduced by Boll (1979), who proposed the spectral subtraction method [27,28]. He demonstrated that the noise estimated during the silent intervals could be subtracted from the noisy spectrum (Equation (8)) to improve the signal-to-noise ratio.

$$\left|\hat{S}(k,m)\right| = \left|\hat{Y}(k,m)\right| + \alpha\left|\hat{V}(k,m)\right| \tag{8}$$

where $\left|\hat{S}(k,m)\right|$ is the estimated clean speech magnitude, $\left|\hat{V}(k,m)\right|$ is the estimated noise magnitude, and α is the spectral over subtraction factor. The spectral over-subtraction factor (α) controls how aggressively the estimated noise spectrum is removed from the noisy spectrum. However, although the method is computationally simple and efficient, it is prone to musical noise artifacts. The Wiener filtering method was introduced by Lim et al. (1977), who proposed an optimal linear filter (Equation (9)) that computes the optimal gain from the power spectral density of the clean speech and noise, thereby minimizing the mean square error (MSE) [29].

$$G(k,m) = \frac{P_S(k,m)}{P_S(k,m) + P_V(k,m)} \tag{9}$$

$$\hat{S}(k,m) = G(k,m)Y(k,m) \tag{10}$$

where $G(k,m)$ is the Weiner gain, $P_S(k,m)$ is the power spectral density, and $P_V(k,m)$ is the PSD of the noise, $Y(k,m)$ is the magnitude spectrum of the noisy signal, $\hat{S}(k,m)$ is the estimated clean speech magnitude. Wiener filtering is optimal in the mean square error sense but requires accurate noise estimation. MSE quantifies the average squared difference between the reference clean speech signal and an estimated enhanced speech signal.

$$\begin{aligned} MSE &= \frac{1}{N}\sum_{n=1}^{N}(s(n) - \hat{s}(n))^2 \\ \hat{A}(k,m) &= E\{A(k,m)|Y(k,m)\} \end{aligned} \tag{11}$$

where $A(k,m)$ is the true speech amplitude, $E\{.\}$ is an expectation operator, and $\hat{A}(k,m)$ is the MMSE estimate of the speech spectral amplitude. This work demonstrated that the optimal statistical filtering can outperform simple subtraction when noise statistics are accurately estimated, thereby establishing an important theoretical benchmark. The most significant contribution to statistical speech enhancement was made by Ephraim, who developed MMSE estimators (1990) [30] and MMSE-log spectral amplitude (LSA) estimators (2003) [31]. These methods derive the gain from the probabilistic distribution of speech spectral amplitude. Their work introduced a probabilistic spectral amplitude model (Equation (11)) instead of a deterministic formulation and optimized the estimator using perceptually relevant distortion measures. Both contributions established MMSE estimation as a standard baseline in speech enhancement, and modern deep learning methods are still frequently compared against the MMSE-LSA algorithm.

Under stationary or slowly varying noise environments, these methods provide efficient performance by achieving SNR improvements in the range of 5–12 dB and PESQ scores of approximately 2–2.8, indicating moderate perceptual improvements. Wiener filtering generally outperforms spectral subtraction by providing smoother spectral estimates. Furthermore, MMSE-based methods improve perceptual quality by optimizing log spectral distortion rather than linear error. Overall, statistical estimation methods are analytically efficient for real-time applications involving stationary and slowly varying noise because of their low computational complexity and effective speech enhancement performance. However, their performance deteriorates significantly under highly nonstationary noise conditions. The dependence on accurate estimation of the noise power spectral density remains a fundamental limitation of these methods. When the underlying statistical assumptions are violated in dynamic acoustic environments, incorrect gain estimation may occur, resulting in residual noise and musical noise artifacts. Consequently, although statistically optimal under their underlying assumptions, these methods are often outperformed by adaptive and data-driven approaches in highly dynamic acoustic scenarios.

3.1.2. Subspace-Based Methods

The concept of subspace-based methods is fundamentally different from statistical estimation and is based on the linear algebraic perspective and matrix decomposition theory, which exploits the low-rank nature of speech signals. It relies on the assumption that the clean speech exhibits strong correlation and redundancy, whereas noise tends to spread across the remaining dimensions. This assumption allows speech signals to be represented in a low-dimensional signal subspace and enables the noisy signal to be decomposed into the speech and noise subspaces through projection. To implement this approach, speech frames are stacked to form a covariance or data matrix (Equation (12)). Various matrix decomposition techniques have been developed, including principal component analysis (PCA) based on eigenvalue decomposition (Equation (12)), the low-rank speech approximation method, singular value decomposition (SVD), and independent component analysis (ICA)-based methods for blind source separation. After matrix decomposition, the dominant large eigenvalue or singular values correspond to speech-dominant components through signal subspace selection. Subsequently, noise-dominant components are discarded, and the speech is reconstructed from the retained subspace (Equation (13)).

$$R_y = E\left[yy^T\right] \tag{12}$$

where R_y is the covariance matrix of noisy speech, y is the speech frame vector, and $(.)^T$ denotes the transpose operator.

$$R_y = U\Lambda U^T \tag{13}$$

where R_y is the covariance matrix of noisy speech, U is the eigenvector matrix, and Λ is the diagonal matrix of eigenvalues.

$$\hat{s} = U_s U_s^T y \tag{14}$$

where $\hat{s}$ is the estimated clean speech vector and U_s represents the signal subspace eigenvectors. The earliest work on subspace-based enhancement was reported by Ephraim and Van Trees (1995), who demonstrated that the speech covariance matrix contains a few dominant eigenvalues, thereby confirming the low-rank nature of speech [32]. Hansen (1991) applied an SVD-based method and demonstrated improved robustness against colored noise compared with spectral subtraction [33,34]. Furthermore, Bell and Sejnowski introduced ICA (1996), establishing the foundation for blind source separation in speech processing and demonstrating that statistical independence provides better separation than orthog-

onality [35–39]. The studies validated the robustness of the eigen-based enhancement as a powerful alternative to the statistical estimation. The approach provides improved performance for certain types of colored or structured noise; however, its effectiveness depends on the accurate subspace estimation and rank selection. Subspace-based methods generally outperform statistical methods under structured or colored noise conditions, achieving improvement in the range of 7–11 dB.

3.1.3. Adaptive Filtering Methods

Adaptive filtering methods are based on adaptive system theory and are specially designed for noise that varies over time. The theory behind the adaptive filtering is based on the observation that noise often exhibits predictable behavior and temporal correlation, whereas speech is relatively random with respect to the reference input signal.

In this method, the filter parameters are continuously updated in real-time to track variation in the noise. When a reference signal correlated with the noise is available, adaptive estimation can effectively separate speech from noise. This concept formulates adaptive noise reduction as an error-minimization algorithm, in which the filter coefficients are iteratively updated to minimize the mean square error between the desired signal and the filter output. Through this iterative optimization process, the algorithm continuously minimizes the error and converges toward an optimal noise estimate under time-varying conditions. Consequently, adaptive filters are highly suitable for suppressing nonstationary interferences, such as background hum and engine noise. The implementation begins by primary and reference input signals. The primary input contains speech corrupted by noise, whereas the reference input contains noise that is correlated with the noise but uncorrelated with the speech component. Next, an adaptive filter is applied, where the filter coefficients are initialized and iteratively updated to model the noise path.

Several adaptive filtering algorithms have been developed, including the least mean squares (LMS) algorithm, which updates coefficients using instantaneous gradient descent; the normalized LMS (NLMS) algorithm, which normalizes the coefficient updates according to the input signal power; and the recursive least squares (RLS) algorithm, which provides faster convergence by minimizing the cumulative squared error. Among these algorithms, RLS generally achieves superior performance because of its second-order statistical optimization, although it requires higher computational complexity. Finally, the estimated noise is subtracted from the primary signal, and the residual signal represents the enhanced speech.

$$\hat{v}(n) = w^T(n)x(n) \tag{15}$$

where $\hat{v}(n)$ is the estimated noise, $w(n)$ is the adaptive filter coefficient vector, and $x(n)$ is the reference noise vector. The error signal can be estimated as

$$e(n) = y(n) - \hat{v}(n) \tag{16}$$

where $e(n)$ is the error signal represents the enhanced speech estimate. The LMS update rule is

$$w(n+1) = w(n) + \mu e(n)x(n) \tag{17}$$

where μ is the step size parameter. The RLS update rule is

$$w(n) = w(n-1) + K(n)e(n) \tag{18}$$

where $K(n)$ is the Kalman gain vector. Widrow et al. (1976) [40] introduced the LMS algorithm, establishing the foundation of adaptive noise cancellation. Their work demonstrated that iteratively updated filter coefficients can suppress noise without requiring prior

statistical knowledge of the interference [40]. Haykin subsequently provided a comprehensive theoretical treatment of the LMS, NLMS, and RLS algorithms [41–49]. These studies demonstrated that LMS is computationally efficient, whereas RLS offers faster convergence by exploiting second-order statistical information. Adaptive filtering techniques have been successfully applied in hands-free communication systems. However, their applicability in single-channel speech enhancement remains limited because they require a reference noise signal, which is often unavailable in practical application.

3.1.4. Transform-Domain Method

The transform-domain method explores the time-frequency representation of the signal to differentiate between speech-dominant and noise-dominant components in the noisy signal. Unlike adaptive filtering methods, transform-domain methods do not require a reference noise signal. The theory behind the transform-domain method is based on the observation that, in the transform domain, the speech signal exhibits a sparse and localized representation. Speech energy is concentrated in a small number of coefficients, whereas noise is distributed more uniformly across the transform domain. This characteristic facilitates the identification and separation of noise while preserving the underlying speech components. The transform-domain approach begins by transforming the noisy signal into an alternative representation that reveals its time-frequency/time-scale characteristics. The short-time Fourier transform (STFT) provides time-frequency information and reveals distinct spectral features (Equation (19)). The wavelet transform offers simultaneous localization in both the time and frequency domains, enabling the analysis of transients and slowly varying components while decomposing the speech signal into multiple resolution levels (Equation (20)). Similarly, empirical mode decomposition (EMD) decomposes the noisy signal into a finite set of intrinsic mode functions (IMFs) (Equation (21)). After signal transformation, noise suppression is performed in the transformed domain. In STFT-based methods, spectral masking or gain estimation techniques are employed to suppress noise. In wavelet-based methods, coefficient thresholding attenuates wavelet coefficients below a predefined or adaptive threshold, thereby reducing noise while preserving important speech components. In the EMD-based method, speech-dominant IMFs are retained or thresholded, whereas noise-dominant IMFs and high-frequency disturbances are attenuated. Finally, the enhanced speech signal is reconstructed using the corresponding inverse transform, such as inverse STFT, inverse wavelet transforms, or IMF summation. Consequently, transform-domain methods effectively suppress noise while preserving the spectral structure of the speech signal.

$$Y(k,m) = \sum_n y(n)w(n-m)e^{-j2\pi kn/N} \tag{19}$$

where $\mathrm{w}(n)$ is the analysis window, j is the imaginary unit, N is the FFT size.

$$\hat{w}_i = \begin{cases} 0, & (|w_i| < T) \\ w_i, & (|w_i| \geq T) \end{cases} \tag{20}$$

where w_i is the wavelet coefficient, $\hat{w}_i$ is the threshold coefficient, and T is the threshold value.

$$y(n) = \sum_{i=1}^{K} IMF_i(n) + r(n) \tag{21}$$

where $IMF_i(n)$ is the ith intrinsic mode function, $r(n)$ is the residual signal, and K is the number of IMFs. Allen (1982) introduced the STFT framework, which laid the foundation for transform-domain spectral masking and demonstrated that the short-time spectral representation can preserve the perceptual speech quality [50]. Donoho (1993) demonstrated

that sparse signal representation enable effective noise suppression through coefficient shrinkage [51], thereby establishing theoretical basis for wavelet-based speech denoising. Rilling et al. (2003) initially proposed EMD for nonstationary signals and later extended its application to speech enhancement [52–57]. Their work demonstrated that IMFs effectively separate speech and noise components under real-world acoustic conditions. Transform-domain methods provide enhanced robustness by exploiting localized time-frequency representations. However, their performance remains sensitive to parameter selection, and issues such as mode mixing, phase distortion, and decomposition errors may degrade enhancement performance.

Overall transform-domain methods provide an excellent compromise between noise suppression and speech preservation because they exploit localized time-frequency representations. Nevertheless, their performance depends strongly on parameter selection, decomposition accuracy, and threshold estimation. Methods such as EMD and VMD may suffer from mode mixing and are computationally more demanding, whereas wavelet-based methods require careful selection of basis functions and decomposition level. Consequently, algorithm selection should balance enhancement performance with computational efficiency and implementation requirements.

3.1.5. Speech Model-Based Method

The core idea behind speech model-based methods is based on the assumption that clean speech follows a well-defined mathematical model. Consequently, an explicit mathematical model of the speech production mechanism is developed, rather than modeling only the noise characteristics. In linear predictive coding (LPC) and autoregressive (AR) models, noise is treated as an additive disturbance. When combined with a speech production model, the estimation framework enables the reconstruction of clean speech even when the noise statistics are unknown or inaccurately estimated. Speech model-based methods operating well in the time or state-space domain are capable of preserving pitch, formant structure, and temporal dynamics, unlike statistical methods that primarily operate in the spectral domain. These methods perform effectively when the speech models accurately capture the vocal tract dynamics.

In this approach, the speech signals are modeled using the LPC or AR approach, in which each sample is represented as a linear combination of its past samples. The model parameters can be estimated either directly from the noisy speech signal or derived from pre-trained speech statistics, allowing the model to capture the spectral envelope and temporal structure of clean speech. A state-space framework (Equation (22)) is used to model speech, where the clean speech is treated as a dynamic system represented by

$$x(n) = Ax(n-1) + w(n) \tag{22}$$

where $x(n)$ is the speech state vector, A is the state transition matrix, and $w(n)$ is the excitation (process noise). The noisy model (Equation (23)) is obtained by incorporating additive noise into the measurement equation:

$$y(n) = Cx(n) + v(n) \tag{23}$$

where C is the observation matrix and $v(n)$ denotes observation noise. Given this state-space formulation, Kalman filtering is employed to recursively estimate the clean speech from the noisy signal.

$$\hat{x}(n|n) = \hat{x}(n|n-1) + K(n)[y(n) - C\hat{x}(n|n-1)] \tag{24}$$

where $\hat{x}(n|n)$ is the updated state estimate, and $K(n)$ is the Kalman gain. The Kalman filter operates through alternating prediction and update stages, during which incoming measurements refine the prior state estimates, thereby optimally balancing model predictions and observed data. Finally, the enhanced signal is reconstructed from the estimated speech states, preserving the temporal and spectral characteristics of the original speech. Atal (1982) [58] introduced one of the earliest speech model-based enhancement methods using LPC, laying the foundation for subsequent speech model-based speech enhancement techniques. He demonstrated that the speech spectrum could be accurately represented using a relatively small number of LPC parameters [58–62]. Kalman (1960) developed the Kalman filtering framework for dynamic systems based on state-space estimation theory [63]. Later, studies demonstrated that if speech is treated as a dynamic system, a Kalman filter can outperform the spectral subtraction while preserving formant structure [64–69]. These studies also demonstrated that the explicit speech modeling reduces speech distortion, and subsequent research further improved parameter estimation and enhanced the robustness of the Kalman-based filter under nonstationary noise conditions.

The speech model-based approach provides a strong prior on speech structure and can achieve high enhancement performance. However, its effectiveness depends on the accuracy of the underlying speech model, and any mismatch between the actual speech signal and the assumed model may degrade performance. Compared with purely spectral enhancement methods, speech model-based techniques explicitly exploit the temporal correlation and speech production mechanism. Consequently, LPC-based methods are particularly suitable for speech coding, low-bit-rate communication, and speech synthesis, whereas Kalman filtering provides superior performance for tracking rapidly varying speech dynamics under nonstationary noise conditions. Spectral enhancement methods, however, remain computationally simpler and are generally preferred for real-time applications requiring low computational complexity.

Although classical signal processing techniques provide mathematically elegant and computationally efficient solutions, their performance gradually deteriorates under highly dynamic acoustic environments because they primarily rely on statistical assumptions regarding speech and noise characteristics. In most of the classical methods, the enhancement parameters are based on predefined mathematical models. However, in practical scenarios such as babble noise, traffic, crowd conversation, and rapidly changing background interferences, these assumptions are often violated, resulting in inaccurate noise estimation, residual noise, musical noise artifacts, and speech distortion. Consequently, improvement in objective measures such as SNR does not always translate into better perceived speech quality or intelligibility. These limitations have motivated the development of perceptually inspired methods that exploit the characteristics of the human auditory system to improve subjective listening quality. Instead of treating all frequency components equally, perceptual methods exploit psychoacoustic principles such as auditory masking, critical band analysis, and perceptual weighting to selectively suppress only perceptually significant noise while preserving speech-dominant components, as discussed in the next section.

3.2. Perceptual and Auditory-Inspired Methods

Perceptual and auditory-inspired speech enhancement methods aim to improve speech quality by incorporating the characteristics of human hearing. However, they are based on different underlying principles. Perceptual approaches exploit psychoacoustic phenomena such as auditory masking, critical bands, and perceptual weighting to suppress perceptually relevant noise while preserving speech naturalness [70]. In contrast, auditory-inspired approaches attempt to model the physiological processing of the human auditory system, including cochlear frequency selectivity and auditory filter bank analysis using Gammatone,

Gammachirp, or equivalent cochlear models [71]. Although both approaches are motivated by human hearing, perceptual methods emphasize perceptual optimization of speech quality, whereas auditory-inspired methods focus on biologically inspired representation of speech signals. Because of their complementary characteristics, both paradigms play a significant role in modern speech enhancement systems, improving perceptual quality while preserving important auditory information.

3.2.1. Perceptual-Based Enhancement

Perceptual-based speech enhancement techniques are developed based on psychoacoustic principles that describe the characteristics of human hearing rather than solely relying on objective signal-level optimization. These techniques are based on the premise that not all noise components are equally audible to human ears. Therefore, suppressing only perceptually significant noise preserves speech naturalness while minimizing unnecessary speech distortion. Perceptual algorithms aim to optimize subjective listening quality by exploiting critical band analysis, auditory masking, and perceptual weighting. The enhancement process begins by transforming the noisy speech signal into a perceptually meaningful frequency representation using critical-band analysis. The Bark scale, which closely approximates the frequency resolution of human hearing, divides the audible frequency range into critical bands corresponding to the filtering characteristics of the human cochlea. The energy within each critical band is estimated as

$$E_b = \sum_{k \in \Omega_b} |Y(k)|^2 \tag{25}$$

where E_b represents the energy of the b^{th} critical band, $Y(k)$ denotes the noisy speech spectrum, and Ω_b represents the frequency bins belonging to the corresponding Bark band. Based on the critical band energy distribution, an auditory masking threshold is estimated for each perceptual band. This threshold represents the minimum sound level that remains audible in the presence of stronger neighboring speech components. The masking threshold may be represented as

$$T_b = f(E_b) \tag{26}$$

where T_b is the masking threshold, and $f(.)$ is a psychoacoustic masking function. Noise components whose energy remains below the threshold are considered perceptually insignificant and therefore require little or no attenuation, whereas components above the threshold are selectively attenuated. Subsequently, a perceptual gain function is computed to attenuate only those noise components whose energy exceed the audibility threshold while preserving the speech-dominant regions essential for maintaining speech intelligibility and naturalness.

$$G_b = \begin{cases} 1, & N_b < T_b \\ \beta, & N_b \geq T_b \end{cases} \tag{27}$$

where N_b is the estimated noise energy in the b^{th} band, and β is the attenuation factor. Finally, the enhanced speech signal is reconstructed through inverse spectral transformation. This method avoids unnecessary suppression of inaudible noise, thereby reducing speech distortion and auditory artifacts. The processed perceptual-domain representation is transformed back into the time domain using inverse filter-bank synthesis, resulting in enhanced speech with improved subjective quality. Johnston (2002) developed a psychoacoustic model describing auditory masking and perceptual threshold [72]. His work demonstrated that removing inaudible noise does not necessarily improve perceived speech quality, thereby motivating the development of perceptually weighted enhancement techniques. Wang and Brown (2006) integrated the auditory filter banks into speech

enhancement systems and demonstrated that perceptual processing significantly improves speech intelligibility and naturalness compared with conventional spectral enhancement methods [73–75]. Subsequent studies have demonstrated that perceptual approaches are highly effective in listening devices such as hearing aids, where subjective speech quality is often more important than SNR improvement. These studies have established perceptually based enhancement as an important framework for hearing-impaired communication and other human-centered speech-processing applications.

The main challenge of the perceptual-based method lies in the complexity of the auditory model and the difficulty of the design of universally optimal perceptual criteria. Although these methods generally achieve higher subjective listening quality, reflected by improved PESQ and MOSs, their performance depends strongly on accurate estimation of auditory masking threshold and psychoacoustic parameters, which may become unreliable under highly dynamic acoustic environments.

3.2.2. Auditory-Inspired Enhancement

Auditory-inspired speech enhancement methods differ fundamentally from perceptual approaches by attempting to imitate the physiological processing performed within the human auditory system. Rather than relying solely on psychoacoustic masking, these techniques model cochlea frequency selectivity, basilar membrane vibration, and auditory nerve response to obtain a biologically meaningful representation of speech signals [76]. Such representations preserve important temporal and spectral information that is often lost in conventional Fourier-based analysis. The processing begins by decomposing the noisy speech signal into multiple auditory frequency channels using cochlear filter banks. Commonly adopted auditory models include Gammatone, Gammachirp, and equivalent rectangular bandwidth (ERB) filter banks. The Gammatone filter models the impulse response of the basilar membrane and is expressed as

$$g(t) = at^{n-1}t^{-2\pi bt}\cos 2\pi f_c t + \varphi \tag{28}$$

where a is the amplitude constant, n denotes the filter order, b is the bandwidth parameter, f_c is the center frequency, and φ represents the initial phase.

The Gammachirp filter extends the Gammatone model by incorporating frequency-dependent chirp modulation to provide a more accurate representation of the asymmetric cochlear response.

$$g(t) = at^{n-1}t^{-2\pi bt}\cos 2\pi f_c t + c\ln t + \varphi \tag{29}$$

where c denotes the chirp parameter controlling the frequency glide. Owing to its improved physiological accuracy, the Gammachirp filter has demonstrated superior modeling of human auditory frequency selectivity. The center frequencies of auditory filters are generally distributed according to the equivalent ERB scale,

$$ERB = 24.7(4.37\frac{f}{1000} + 1) \tag{30}$$

which approximates the frequency resolution on the human cochlea across the audible spectrum. This representation provides non-uniform frequency spacing that closely resembles biological hearing and has become a standard front-end representation in many speech processing applications. Following auditory decomposition, each frequency channel is processed independently using adaptive gain estimation, noise suppression, and neural network-based enhancement before reconstructing the enhanced speech signal. Because auditory representations preserve fine temporal structure and cochlear frequency selectivity,

they exhibit improved robustness under highly nonstationary noise conditions compared with conventional spectral analysis.

The theoretical foundation of auditory-inspired processing originated from cochlear modeling developed by Patterson and Holdsworth through the introduction of the Gammatone auditory filter bank [77]. Irino and Patterson later introduced the Gammachirp filter, providing a more accurate representation of cochlear frequency responses [78]. More recently, auditory-inspired front-end processing has been integrated with conventional neural networks, transformers, and hybrid speech enhancement architectures, where biologically motivated representations improve feature extraction and generalization across diverse acoustic environments [79–81]. Auditory-inspired enhancement provides superior robustness for complex acoustic environments because it preserves perceptually relevant speech features while modeling physiological auditory processing. Nevertheless, its higher computational complexity and the need for accurate auditory modeling remain important challenges for real-time implementation, motivating continued research on efficient auditory-impaired and hybrid speech enhancement frameworks.

Despite their ability to improve perceptual speech quality, perceptual and auditory-inspired methods face several practical challenges. Their performance depends on accurate estimation of auditory masking threshold and psychoacoustic parameters, which may vary considerably among listeners and across acoustic environments. In addition, the computational complexity associated with auditory modeling can limit real-time implementation in resource-constrained devices. Furthermore, developing a universally applicable psychoacoustic model remains difficult because human auditory perception is inherently subjective and influenced by individual hearing characteristics and environmental conditions. Consequently, further research is required to develop more adaptive, computationally efficient, and generalizable perceptual enhancement frameworks.

Although perceptual and auditory-inspired methods improve the speech quality by exploiting psychoacoustic principles and modeling the human auditory system, they generally operate on a single microphone and therefore have limited capability to distinguish sound sources based on their spatial locations. In practical acoustic environments, speech and interfering noise often originate from different directions, making it difficult for a single-channel algorithm to effectively suppress directional interference. Modern communication systems such as microphone arrays, hearing aids, teleconferencing systems, and voice control devices commonly employ multiple microphones to capture both temporal and spatial characteristics of the acoustic field. By utilizing differences in the direction of arrival, phase, and time delay of the received signals, spatial processing techniques can selectively enhance the desired speech source while suppressing interfering sound from other directions. This capability has led to the development of spatial processing methods, which exploit spatial diversity to achieve improved robustness and speech intelligibility in complex acoustic environments.

3.3. Spatial Processing Methods

Spatial processing methods exploit the spatial diversity of sound sources by arranging multiple microphones in an array. Single-channel methods depend on the spectral, statistical, and temporal characteristics of the received signals, whereas spatial methods exploit differences in the direction of arrival (DOA), spatial covariance, and propagation delay of acoustic signals. Special processing is based on the fundamental principle that the speech and noise arriving from different spatial locations exhibit distinct spatial characteristics, which can be exploited to identify and suppress interfering sound sources. These techniques are widely used in hearing aids, conference systems, teleconferencing, and automatic speech recognition systems in noisy environments.

3.3.1. Beamforming

Beamforming is a widely used strategy among special filtering techniques. Its operation is based on applying appropriate time delays and weights to microphone signals before combining them. Signals arriving from the desired direction are constructively combined to enhance the target speech, whereas signals arriving from other directions are destructively combined to suppress interference. Delay-and-sum beamforming is the simplest approach, in which the microphone signals are initially aligned according to the expected propagation delay and then averaged. This technique is computationally efficient but provides limited interference suppression in highly noisy environments. The minimum variance distortion response (MVDR) beamformer minimizes the output power while preserving the desired signals without distortion. The MVDR beamformer provides substantial interference suppression and is widely employed as a standard component in modern noise suppression systems. The generalized sidelobe canceller (GSC) is based on decomposing the system into fixed beamforming and adaptive noise cancelation stages. The GSC effectively improves noise suppression in dynamic acoustic environments; however, its performance may degrade because of speech leakage in the adaptive branch.

3.3.2. Multichannel Approach

Multichannel enhancement techniques exploit the correlation between microphone signals by estimating spatial covariance matrices and computing optimal gain functions for noise suppression. Subsequently, speech signals are separated from noise through statistical spatial filtering. Advanced multichannel systems combine beamforming with spectral post-filtering to further suppress the residual noise and reverberation.

Despite their excellent directional noise suppression capability, spatial processing methods require multiple synchronized microphones and accurate knowledge of the array geometry. Their performance is sensitive to microphone mismatch, calibration errors, reverberation, and inaccuracies in direction-of-arrival estimation. Furthermore, the additional hardware increases system cost and limits deployment in compact or low-cost communication devices.

Although single-channel algorithms rely primarily on spectral, temporal, and statistical characteristics of the received speech signal, multichannel systems traditionally exploit spatial diversity through microphone arrays. Beamforming provides effective suppression of directional interference before subsequent speech enhancement processing, whereas single-channel algorithms further suppress residual background noise and improve perceptual quality. Consequently, modern communication systems increasingly integrate beamforming with single-channel enhancement techniques to achieve superior robustness in reverberant environments. These approaches have achieved significant improvements; however, their performance remains constrained by predefined mathematical models and assumptions regarding speech production, noise characteristics, and acoustic propagation. Although computationally efficient and theoretically well-established, they often struggle to represent the highly nonlinear, time-varying, and complex nature of real-world acoustic environments. These limitations have motivated the emergence of data-driven speech enhancement, where machine learning models learn the mapping between noisy and clean speech directly from large-scale datasets rather than relying on handcrafted signal models.

3.4. Data-Driven Methods

The application of machine learning in noise suppression represents a paradigm shift from the analytical modeling to the data-driven learning. Instead of designing an explicit model of speech and noise statistics, machine learning methods learn a nonlinear mapping between noisy and clean speech using large-scale datasets. These methods

are based on the observation that the complex relationship between speech and noise in nonstationary environments cannot be adequately modeled using fixed mathematical formulations but can instead be learned directly from the data. Machine learning-based algorithms are formulated as supervised learning frameworks in which a model is trained to learn the mapping between noisy speech and the corresponding clean speech. The first step involves feature extraction, where noisy speech signals are transformed into time-frequency representation, including short-time Fourier transform (STFT) spectra, magnitude spectra, log-power spectra, and Mel-frequency features. These representations capture the spectral and temporal characteristics of speech and noise and serve as inputs to the learning models. The learning model can be expressed as shown in Equation (31).

$$\hat{S} = f_\theta(Y) \tag{31}$$

where $f_\theta(.)$ is a neural network with parameter θ, $\hat{S}$ is the estimated clean speech spectrum. At the training stage, the dataset is used to optimize the parameters of the learning model. Depending on the framework, the model can predict the noise spectrum, estimate the clean speech spectrum, or compute a time-frequency mask to attenuate noise-dominant regions. Thereafter the model parameters are optimized by minimizing a loss function (Equation (32)) that quantifies the difference between the estimated and reference clean speech signals using either the mean square error or perceptually motivated loss functions.

$$\zeta = \left\| S - \hat{S} \right\|^2 \tag{32}$$

where ζ is the training loss, and $\|.\|$ denotes the Euclidean norm. During the inference stage, the trained model processes previously unseen noisy speech features and generates an improved spectral estimate or time-frequency mask. The mask-based reconstruction can be expressed as

$$\hat{S}(k,m) = M(k,m)Y(k,m) \tag{33}$$

$M(k,m)$ is the estimated time-frequency mask. The noisy phase information is combined with the estimated magnitude spectrum and transformed back into the time domain using the corresponding inverse transform. This framework demonstrates that machine learning methods can effectively attenuate highly intense, nonstationary, and previously unseen noise types by learning the complex nonlinear relationship directly from data, thereby achieving significant improvements in speech quality and intelligibility.

The classical machine learning (ML) methods are based on shallow learning models [82–84], which assume linear or piecewise linear relationships, such as the Gaussian mixture model (GMM) and non-negative matrix factorization (NMF). The work of Lee and Seung (1999) on NMF demonstrated the feasibility of learning-based decomposition and showed that the speech spectra can be decomposed into additive basis components [85,86]. The literature on the GMM approach demonstrated that, by adapting to the underlying data distribution, statistical learning-based methods can achieve better noise reduction than fixed estimators [87–89].

On the other hand, modern deep learning (DL) approaches employ deep neural networks (DNN), convolutional neural networks (CNN), recurrent neural networks (RNN), particularly long short-term memory (LSTM) networks, and transformers, which are capable of modeling highly nonlinear relationships and long-term temporal dependencies. DNNs learn global nonlinear mappings, CNNs exploit local spectral correlations, LSTM networks model temporal dependencies across consecutive frames, and transformer-based models capture long-range contextual information through attention mechanisms. The work of Xu et al. (2013) on DNN-based speech enhancement marked a breakthrough in

machine learning, demonstrating that deep neural networks can directly learn the mapping from noisy speech to clean speech, thereby achieving unprecedented improvement in PESQ and STOI [90–95]. Subsequently, Wang et al. (2016) combined CNNs and LSTM networks and demonstrated that deep learning methods consistently outperform classical baseline algorithms across different datasets and noise conditions [96–99]. These methods effectively model temporal dependencies, spectral patterns, and contextual information that are difficult to capture using conventional signal processing techniques. These successive advances established deep learning as the state-of-the-art approach for speech enhancement and noise suppression.

Deep learning-based speech enhancement methods have made remarkable progress, but several practical challenges still limit their widespread deployment in real-world applications. These issues are summarized as:

- ✓ Data dependency: Deep learning-based speech enhancement methods rely heavily on the availability of large, diverse, and accurately annotated datasets for effective training. Most supervised models require paired noisy and clean speech recordings collected under various acoustic conditions to achieve robust performance. However, acquiring such dataset is expensive and time-consuming, while limited training diversity may reduce the model's ability to handle previously unseen speakers, noise types, and recording environments.
- ✓ Model interpretability: Another important challenge is the limited interpretability of deep learning models. Unlike classical signal processing techniques, which are based on well-defined mathematical formulations and physically meaningful parameters, deep neural network operates as complex nonlinear mapping functions with millions of trainable parameters. Consequently, understanding the decision-making process and explaining why a particular enhancement result is obtained remain challenging, reducing transparency and limiting their adoption in safety-critical applications.
- ✓ Generalization capability: The generalization capability of deep learning models remains an active research challenge. Although these models perform exceptionally well under acoustic conditions similar to their training data, their performance may degrade when exposed to previously unseen noisy environments, speakers, languages, and recording devices. Improving robustness through domain adaptation, self-supervised learning, continual learning, and large-scale foundation models has therefore become an important direction for the next-generation of speech enhancement.

However, deep learning approaches have achieved remarkable performance by learning complex nonlinear relationships from large-scale datasets; however, they often require substantial computational resources, large annotated datasets, and high-performance hardware for effective training and deployment. Furthermore, their performance may degrade when encountering acoustic conditions that differ substantially from the training data, limiting their generalization and real-time applicability in resource-constraint environments. To address these challenges, recent research has increasingly adopted hybrid speech enhancement frameworks that integrate the computational efficiency, interpretability, and domain knowledge of classical signal processing with the powerful representation-learning capability of artificial intelligence.

3.5. Hybrid Approach

Since the early development of speech enhancement methods, numerous approaches have been developed, leading to continuous improvement in performance. However, no single method can effectively suppress all types of noise across diverse acoustic environments. To overcome the limitation of individual techniques, the hybrid approach combines two or more fundamentally different methods into a unified framework. The

hybrid framework leverages the complementary strength of different approaches while mitigating the individual limitations to achieve optimal performance. The literature indicates that classical methods improve interpretability, transform-domain methods effectively handle nonstationary noise, and machine learning offers superior enhancement performance. By integrating these complementary techniques, hybrid frameworks can achieve higher enhancement performance, improved perceptual quality, and greater robustness. The working mechanism of a generic hybrid framework follows a multistage cascaded processing architecture (Equation (34)), where each stage performs a specific enhancement task. The generic hybrid speech enhancement framework can be represented as a cascade of multiple enhancement operators, where the output of one stage serves as the input to the subsequent stage.

Each processing block performs a complementary task such as preliminary noise suppression, feature refinement, adaptive estimation, or deep learning-based enhancement, thereby progressively improving speech quality while preserving speech intelligibility. The overall framework is mathematically expressed as

$$\hat{s}(n) = E_2(E_1(y(n))) \tag{34}$$

where $\hat{s}(n)$ is the enhanced speech signal, $y(n)$ is the noisy speech signal, $E_1(.)$ is the first-stage enhancement operator, and $E_2(.)$ is the second-stage enhancement operator. The initial stage involves front-end signal processing techniques such as transform-domain decomposition (STFT, wavelet transform, and EMD) or classical statistical filters to perform initial noise reduction. This stage reduces the noise variance and improves the SNR before further processing. At the intermediate step, more sophisticated methods, such as subspace techniques, speech modeling, and perceptual weighting, are incorporated to further attenuate residual noise and refine speech-dominant components by exploiting the structural and perceptual characteristics of the speech signal. In some hybrid frameworks, these stages are followed by a learning-based module, such as DNN, which processes the pre-enhanced speech to remove the remaining noise artifacts and estimate the optimal gain function or time-frequency mask. Finally, a post-processing stage applies additional filtering techniques, such as Wiener filtering, Kalman filtering, or temporal smoothing, to further reduce artifacts and improve speech quality. The enhanced speech signal is then reconstructed in the time domain using the corresponding inverse transform.

Table 4. Recent Hybrid Noise Reduction Methods.

Hybrid Method	Core Idea	Test Noise	Key Result	Ref
Wiener + DNN	Classical Wiener front-end for coarse denoising + DNN for refinement	Babble, street	Improved PESQ/reduced musical noise vs. standalone stages	[100]
DNN + Kalman	DNN estimates AR/PSD parameters; colored-noise Kalman filter for final smoothing	Colored/ real noises	Significant ΔSNR PESQ gains vs. Kalman alone	[101]
Beamforming + DNN post-filter	Multi-mic beamformer for spatial selectivity + DNN post-filter for spectral cleanup & phavee correction	Reverberan, multi-talker	Robust multi-channel improvement in ASR & PESQ	[102]
Wavelet front-end + DNN	Wavelet/subband decomposition feeds DNN per subband (or DNN acts on wavelet coeffs)	Street/babble	Better transient preservation and reduced models size	[103]

Table 4. *Cont.*

Hybrid Method	Core Idea	Test Noise	Key Result	Ref
EMD + DNN/PCA	Decompose into IMFs (or modes), n apply subspace filtering or DNN per component	Nonstationary/ impulsive	Robustness to impulsive and time-varying noise	[104]
VMD + DNN	Unsupervised decomposition (NMF/VMD) followed by supervised DNN to refine masks/coefs	Babble/music	Improves separation alongside fewer labels	[105]
DNN + WPE + Beamforming	Joint optimization: DNN for mask estimation → WPE dereverberation → beamformer weights, trained end-to-end	Reverberant rooms	Strong ASR improvement and dereverberation	[106]
Diffusion + feature-guided pretraining	Pretrain diffusion model alongside feature guidance (hybrid of generative pretrain + predictive features)	Babble	Improved efficiency & enhanced PESQ/STOI	[107]
Transformer TF-mask + perceptual weighting	Transformer long-context attention for mask estimation + perceptual weighting/post-filter	Music/babble	Better long-context coherence and MOS	[108]

Earlier hybrid studies by Hu and Loizou (2007) [109] combined wavelet decomposition with Wiener filtering and demonstrated that the hybrid framework suppresses the musical noise while preserving speech quality. Zhao (2015) [110] integrated EMD with a subspace model, exhibiting improved robustness against the nonstationary noise. More recently, deep learning models combined with classical signal processing methods, guided by signal processing principles, have achieved superior performance under low-SNR conditions.

The hybrid approach effectively suppresses highly nonstationary noise while maintaining high perceptual speech quality. Table 4 summarizes representative hybrid speech enhancement frameworks that integrate multiple signal-processing paradigms to exploit their complementary advantages.

Hybrid approaches employ multistage processing in which classical signal-processing methods perform coarse noise suppression or feature extraction, while advanced learning-based methods further refine the enhanced speech by suppressing residual noise and reducing speech distortion. This collaborative processing strategy enables hybrid systems to overcome many of the limitations associated with individual techniques, including musical noise artifacts, inaccurate noise estimation, limited adaptability, and poor generation under highly dynamic environments. The comparison reveals a clear evolution of hybrid frameworks from the integration of traditional signal-processing algorithms to modern deep learning architectures. These systems benefit from both the interpretability and computational efficiency of classical signal processing and the powerful nonlinear modeling capability of deep learning.

Table 5 presents a comparative overview of the representative speech enhancement algorithms based on their fundamental operating principles, key strengths, and typical application domains, while Table 6 compares the performance of major speech enhancement methods. The comparison highlights the continuous evolution of speech enhancement from mathematically driven signal processing approaches to data-driven intelligence systems. Statistical estimation methods, including spectral subtraction, Wiener filter, and MMSE

estimators, provide computationally efficient solutions and remain attractive for real-time applications under stationary or slowly varying noise environments.

Table 5. Comprehensive analysis of noise reduction methods.

Method	Core Principle	Key Strength	Ref
Spectral Subtraction	Noise spectrum subtraction	Simple, low complexity	[28]
Wiener Filter	MMSE-based spectral gain	Smooth suppression	[111]
MMSE-LSA	Log-spectral MMSE estimation	Improved perceptual quality	[31]
PCA-Based Enhancement	Signal subspace projection	Preserves dominant speech	[32]
SVD-Based Enhancement	Low-rank approximation	Better noise separation	[112]
ICA-Based Enhancement	Statistical independence	Effective source separation	[113]
LMS Adaptive Filter	Gradient adaptation	Tracks slow non-stationarity	[114]
NLMS Adaptive Filter	Normalized adaptation	Faster convergence	[49]
RLS Adaptive Filter	Recursive least squares	Rapid tracking	[115]
STFT Spectral Masking	TF masking	Localized suppression	[116]
Wavelet Thresholding	Multi-resolution thresholding	Preserves transients	[117]
EMD-Based Enhancement	IMF decomposition	Handles nonstationary noise	[118]
LPC-Based Enhancement	Speech production modeling	Formant preservation	[119]
Kalman Filter	State-space estimation	Temporal smoothing	[120]
Auditory Masking	Psychoacoustic masking	High MOS	[72]
Auditory Filter Bank	Cochlear analysis	Improved intelligibility	[121]
NMF-Based Enhancement	Parts-based decomposition	Music/noise modeling	[86]
GMM-Based Enhancement	Probabilistic modeling	Flexible noise modeling	[122]
DNN-Based Enhancement	Nonlinear spectral mapping	Strong PESQ/STOI	[91]
CNN-Based Enhancement	Local TF learning	Captures spectral patterns	[1]
LSTM-Based Enhancement	Temporal modeling	Long-term context	[123]
Transformer-Based Enhancement	Attention-based modeling	Global context	[124]

Adaptive filtering methods further improve robustness by exploiting signal correlations and adaptive optimization, making them suitable for applications involving structured or time-varying interference. Transform-domain and speech-model-based techniques preserve the spectral and temporal characteristics of the speech more effectively by utilizing sparse signal representations or explicit speech production models, resulting in improved perceptual quality and reduced speech distortion. Auditory-inspired and perceptually motivated algorithms enhance the subjective listening experience by incorporating psychoacoustic characteristics rather than focusing solely on SNR improvement.

More recently, data-driven approaches based on machine learning and deep neural networks have significantly advanced the state of the art by learning complex nonlinear mappings directly from large speech datasets. These methods consistently achieve superior objective and perceptual performance, particularly under highly nonstationary noise conditions, but require substantially greater computational resources, large annotated datasets, and dedicated hardware for training and inference.

Therefore, the selection of an appropriate speech enhancement technique depends not only on enhancement performance but also on computational complexity, latency

requirements, availability of training data, robustness to environmental variations, and the target application.

Table 6. Representative performance comparison of major speech enhancement methods.

Algorithm Category	Typical SNR Improvement	PESQ	Computational Cost	Key Trade-Off	Ref
Statistical Methods	5–12 dB	2.0–2.8	Low Real-time capable, minimal memory	Fast and interpretable, degrades sharply under nonstationary noise	[28,125]
Subspace Methods	7–11 dB	2.1–2.9	Moderate Eigen decomposition adds cost	Good for structured/colored noise; sensitive to rank selection	[32]
Adaptive Filtering	8–15 dB	2.2–3.0	Low (LMS/NLMS), moderate (RLS)	Needs a correlated reference signal; RLS converges faster	[49,114]
Transform-Domain	6–12 dB	2.2–3.1	Moderate decomposition/reconstruction overhead	Better spectral/temporal preservation; performance tied to model accuracy	[117,126]
Perceptual/Auditory	4–10 dB	2.5–3.4	Moderate to high auditory filterbank/masking-threshold estimation adds cost	Prioritizes subjective quality over raw SNR; ideal for hearing aids	[72,121]
Deep Learning	12–22 dB	2.8–4.2	High large datasets, memory training, heavy resource demand	Best objective/perceptual scores, weak generalization to unseen noise	[90]
Hybrid Frameworks	14–25 dB	3.0–4.4	High multistage cascade compounds cost of constituent stages	Highest robustness/perceptual quality; increased system complexity, latency	[109]

3.6. Complementary Acoustic Enhancement Techniques

The speech enhancement algorithms for suppressing background noise and recovering clean speech signals have been discussed in the preceding section. In addition, acoustic signal processing encompasses several complementary techniques that address related problems, such as active noise control (ANC) for adaptive noise cancellation and acoustic echo cancellation (AEC) for echo suppression, thereby contributing to the development of advanced acoustic signal processing systems. Likewise, sparse subband adaptive filtering has attracted considerable attention because of its fast convergence, reduced computational complexity, and improved performance for echo cancellation and nonstationary noise suppression. Although these domains address objectives that differ from those of the speech enhancement algorithms discussed earlier, they share common theoretical foundations, including adaptive filtering, optimization, and signal estimation, and are increasingly integrated into modern hybrid speech enhancement frameworks. This convergence reflects the growing trend toward unified and intelligent acoustic signal processing systems capable of jointly addressing noise suppression, echo cancellation, and adaptive acoustic control.

Active noise control. Active noise control (ANC) is a widely used and effective technique in headphones, automotive cabins, and other acoustic applications for suppressing unwanted, repetitive, low-frequency noise [127–129]. The ANC system suppresses the unwanted sound by generating an anti-noise signal with the same amplitude and opposite phase, thereby achieving destructive interference.

Adaptive filtering methods are the predominant approach for ANC, particularly the feedforward filtered-x least mean square (FxLMS), which has gained widespread adoption because of real-time feasibility, computational efficiency, and implementational simplicity. The FxLMS method minimizes the second-order error moment, i.e., $J(n) = E\{e^2(n)\}$, where $e(n)$ is the residual error signal and $E\{.\}$ denotes expectationoperator [130–133].

Another widely used variant, the filtered-x normalized least mean square (FxNLMS) algorithm, addresses the sensitivity of FxNLMS to variations in input signal power. It further incorporates normalization, leakage, and zero attraction mechanisms to better handle filter sparsity, reduce coefficient drift, and improve computational efficiency [134–136]. The variable step size (VSS)-FxLMS approach adapts the learning rate online to balance convergence speed and steady-state error. VSS-FxLMS variants, including heuristic and data-driven schemes, improve adaptation under stationary conditions but retain the MSE objective, making them vulnerable to impulsive outliers [137–140]. Robust adaptive filtering for impulsive environments has been extensively studied outside the ANC framework, with least audible deviation (LAD), Huber, and maximum correntropy criteria (MCC)-based methods demonstrating superior robustness against outliers [141–143]. Generalized filtered-x MCC and maximum correntropy criterion-based FxLMS variants have been proposed to address non-Gaussian noise in ANC. However, existing formulations often suffer from either high steady-state misalignment or the inability to jointly address adaptive kernel-width estimation and secondary-path compensation [144–146]. Correntropy is a localized similarity measure induced by a positive definite kernel that emphasizes small informative errors while suppressing the influence of large outliers. More recently, the distribution-aware, risk-sensitive, filtered-x NLMS (DA-RS-FxNLMS) method replaces the traditional MSE objective with a correntropy-based, risk-sensitive cost and combines it with an adaptive kernel width estimator and filtered reference compensation [147]. These developments indicate that future ANC systems are likely to become an integral component of intelligent speech communication and acoustic signal processing systems.

Acoustic echo cancellation. Acoustic echo cancellation (AEC) is another technique that aims to estimate the acoustic echo path and subtract the estimated echo from the microphone signal before subsequent speech processing. Conventional AEC systems are based on adaptive filtering methods such as least mean square (LMS), normalized LMS, affine projection algorithm (APA), recursive least square (RLS), and Kalman filtering. Among these, NLMS is one of the most widely adopted algorithms because it offers an attractive balance between computational complexity and convergence performance. Most recent proportionate adaptive filters and sparse adaptive algorithms have demonstrated superior performance in identifying long acoustic impulse responses in practical environments. Modern neural network-based echo cancellation systems estimate nonlinear echo components, improve double-talk robustness, and compensate for loudspeaker distortions. Contemporary conferencing platform and full-duplex communication systems combine acoustic echo cancellation with speech enhancement, dereverberation, beamforming, and noise separations to provide an integrated front-end speech processing framework.

Sparse subband adaptive filters. Adaptive filtering remains the fundamental computational framework for many speech enhancement, ANC and AEC algorithms. Under long acoustic impulse-response conditions, broadband adaptive filters often suffer from high computational complexity and slow convergence. Sparse subband adaptive filters decompose broadband signals into multiple frequency bands, allowing adaptive processing to be performed independently within each subband. This decomposition reduces eigenvalue spread, accelerates convergence, and lowers computational complexity. Furthermore, sparse adaptive filtering exploits the observation that many acoustic impulse responses contain only a small number of dominant coefficients. Algorithms such as proportionate

NLMS, improved PNLMS, and sparsity-aware LMS-based subband adaptive filters selectively update significant coefficients, thereby achieving faster convergence and improved steady-state performance. Owing to their computational efficiency and rapid adaptation capability, sparse subband adaptive filters have become an attractive option for real-time speech enhancement, acoustic echo cancellation, and active noise control systems.

4. Critical Assessment: Merits, Challenges, and New Perspectives

This study presents a unified, systematic, and comprehensive framework for noise suppression techniques for speech enhancement by integrating contributions from diverse research domains, thereby bridging classical signal processing techniques with modern machine learning approaches, unlike conventional survey articles, which are limited to partial categorization. The proposed five-category taxonomy provides a comprehensive classification of speech enhancement methods and facilitates an understanding of the evolution of noise separation techniques from traditional analytical models to data-driven frameworks, which is often fragmented in existing literature. Furthermore, beyond providing a descriptive summarization, this review presents a comparative analysis across all categories to explain why specific methods perform better under different acoustic conditions. The insight-oriented comparison provides a deeper understanding of algorithm selection, which is typically missing in previous review articles. The methods are evaluated from the perspectives of performance, computational complexity, and practical applicability, where different techniques are assessed not only in terms of noise suppression performance using objective metrics such as SNR, PESQ, and STOI but also with respect to computational complexity, real-time feasibility, and practical deployment constraints. The multidimensional evaluation provides valuable guidance for selecting an appropriate speech enhancement algorithm for applications such as speech recognition, telecommunication systems, hearing aids, etc.

Classical signal processing methods play a crucial role in establishing the foundation of speech enhancement because of their computational efficiency. Statistical and adaptive filtering techniques require relatively low computational resources. However, transform-domain and speech model-based methods exhibit moderate computational complexity because of additional signal decomposition and reconstruction processes they employ. These approaches are mathematically well established and are suitable for real-time applications because they rely on well-established analytical formulations. Under stationary or slowly varying noise conditions, they provide effective enhancement while maintaining low computational complexity. However, the performance of these approaches depends on the accurate noise estimation and assumptions regarding the characteristics of the speech and noise, including whether they are stationary or nonstationary and linear or nonlinear. In practical acoustic environments where noise is highly nonstationary and rapidly varying, these assumptions may no longer hold. Under such conditions, speech distortion, residual noise, and musical noise artifacts may occur, thereby degrading speech quality. Furthermore, these methods exhibit limited capability in modeling the complex nonlinear relationships between speech and noise. Despite these advantages and limitations, classical methods remain highly effective for practical applications such as telecommunication, embedded systems, and low-power speech processing. Future work should focus on integrating classical signal processing techniques with modern machine learning approaches to improve noise suppression while preserving speech quality under dynamic acoustic environments. Such hybrid frameworks combine the interpretability and computational efficiency of classical methods with the powerful learning capability of modern machine learning techniques.

The auditory-inspired approach provides valuable insights into human auditory perceptions and shifts speech enhancement techniques from a purely signal-based approach to human-centered frameworks. The incorporation of psychoacoustic principles, such as frequency selectivity, auditory masking, and critical-band analysis, enhances speech quality while preserving naturalness. The approach effectively suppresses perceptually disturbing artifacts while maintaining the listening comfort required for hearing aids and related assistive listening devices. However, the performance of auditory-inspired techniques depends on the accuracy of the psychoacoustic model. It is known that human auditory perception varies among individuals and is influenced by several factors, such as age group, hearing ability, and listening conditions. Hence, developing a universally optimal auditory masking model that accurately represents these variations remains highly challenging. In addition, these methods can improve subjective quality as reflected by MOS and listener preference but may not always achieve comparable improvements in objective performance metrics such as SNR. Despite these strengths and shortcomings, several challenges remain unresolved. First, the development of auditory-inspired models and perceptually meaningful objective evaluation metrics remains an important research direction. Second, future systems could incorporate personalized auditory models that adapt to the hearing characteristics of individual users.

Unlike single-channel techniques that depend on the spectral, temporal, and statistical characteristics of the received speech signal, spatial processing approaches provide a fundamentally different perspective on speech enhancement. These methods exploit the spatial diversity of sound sources by using multiple microphones, utilizing DOA information, propagation delay, and spatial covariance matrices to distinguish desired speech from interfering noise. This strategy significantly improves speech intelligibility and noise separation under dynamic acoustic environments, particularly in multi-speaker scenarios. However, accurate microphone array geometry and sensor calibration remain essential for reliable estimation of spatial parameters. In a highly reverberant environment, reflections can distort the directional information required for source separation, thereby degrading enhancement performance. In addition, these systems require multiple microphones, which increases hardware complexity, power consumption, and implementational cost. Furthermore, signal leakage and estimation error remain challenging issues in adaptive beamforming techniques. The computational complexity varies among different beamforming strategies, resulting in a trade-off between enhancement performance and computational efficiency. Delay-and-sum beamforming is particularly suitable for real-time applications because of its low computational requirements. In contrast, advanced methods, including MVDR beamforming, generalized sidelobe cancellers, and multichannel Wiener filters, require sophisticated matrix computations. The integration of a spatial-spectral joint framework could open a promising direction for improving robustness.

Data-driven approaches have transformed the field of speech enhancement by replacing analytical modeling with learning-based techniques trained on large-scale datasets. Both machine learning and deep learning methods can model the complex and nonlinear relationship between speech and noisy signals, making them capable of performing effectively under highly challenging acoustic conditions. Advanced speech enhancement architectures such as DNN, CNN, LSTM networks, Transformers, and diffusion models provide excellent objective and subjective performance metrics. Despite their state-of-the-art performance, these approaches still face several practical challenges. Deep learning architectures require large datasets for training, resulting in substantial computational and memory requirements. Furthermore, these architectures are generally developed as black-box models with limited interpretability, making it difficult to explain their decision-making process. Moreover, when deployed in previously unseen acoustic environments that differ

significantly from the training data, their performance may degrade, raising concerns about their robustness and generalization capability. These limitations restrict their reliability in resource-constraint environments. However, different machine learning architectures exhibit different computational complexities. Classical machine learning methods generally require moderate computational resources, whereas deep learning models demand substantially greater memory and computational power. Future research should therefore focus on lightweight network architectures, data-efficient training strategies, self-supervised learning, and foundation models to improve robustness, computational efficiency, and real-time performance in previously unseen acoustic environments.

The critical analysis of the different categories of speech enhancement methods indicates that no single method performs optimally in all acoustic conditions, thereby motivating the development of hybrid frameworks. The integration of well-established noise suppression methods within a single framework provides a promising direction for advancing speech enhancement. The unified framework provides flexibility and adaptability, allowing the complementary strengths of individual techniques to be effectively exploited and thereby achieving superior performance compared with standalone methods. The multistage processing framework integrates these complementary strengths to achieve improved performance. For example, classical signal processing provides interpretable and computationally efficient front-end processing; machine learning models further refine residual distortions through nonlinear modeling; and transform-domain processing isolates speech-dominant components from background noise before subsequent enhancement. However, the integration of multiple processing stages requires an optimized system architecture and may increase overall computational complexity. Furthermore, these interconnected frameworks require careful parameter tuning and substantial training resources. In addition, the increased number of interconnected components demands greater memory than conventional approaches. Consequently, multistage processing results in higher computational complexity. These limitations continue to restrict the deployment of hybrid frameworks in resource-constrained devices in real-time applications and highlight the need for more efficient interaction among individual processing stages. Nevertheless, the substantial improvement in enhancement performance achieved by hybrid frameworks generally outweighs these computational challenges, making them a promising direction for future speech enhancement research.

5. Challenges and New Perspectives

This study demonstrates that significant progress has been made in the field of speech enhancement and noise reduction. The overall framework organizes the existing techniques into five major categories based on their underlying operating principles. The mathematical formulation provides a coherent understanding of how speech and noise signals are modeled to recover the original signal. The analysis shows that classical and subspace-based methods offer low computational complexity and are easy to implement; however, their performance is largely limited to stationary or slowly varying noise environments. Adaptive filtering methods perform effectively when reference signals are available. On the other hand, transform-domain methods enhance the robustness by exploiting the time-frequency and time-scale localization properties of speech signals. Speech model-based approaches preserve the speech structure and formant information by incorporating the explicit speech production models. Furthermore, perceptual and auditory-inspired methods improve subjective listening quality by aligning noise suppression techniques with psychoacoustic characteristics of human hearing. The emergence of machine learning methods significantly advanced the modeling of complex nonlinear systems, enabling effective enhancement of degraded speech signals. Machine learning-based methods achieve excellent perfor-

mance in terms of objective and perceptual evaluation metrics but require large datasets for training, resulting in enhanced computational cost. The comprehensive classification of the speech enhancement methods explains their operating mechanism, strengths, limitations, and key application advantages. Overall, this review provides a balance comparison of representative speech enhancement methods under diverse acoustic environments and serves as a useful reference for selecting appropriate techniques for different practical applications.

Despite these advances, several fundamental challenges remain unresolved under adverse acoustic conditions. The primary concern is the robust handling of unpredictable and intense nonstationary noise. Existing methods are effective; however, they still struggle in complex, rapidly changing, and reverberant environments. Accurately modeling speech and noise under such highly dynamic acoustic conditions remains challenging. Another major issue lies in the trade-off between speech distortion and noise suppression. Algorithms designed to aggressively suppress noise may inadvertently remove important speech components, including formant structure and temporal continuity, thereby compromising speech intelligibility and naturalness. Speech model-based and perceptual methods partially mitigate this issue by maintaining an optimal balance between noise suppression and speech preservation. Furthermore, objective evaluation metrics do not always correlate well with subjective listening quality, complicating the assessment of the speech enhancement algorithms. The dependency of deep learning methods on large annotated datasets also remains a significant challenge because it requires substantial computational resources, limiting their deployment in low-power and resource-constrained applications, such as hearing aids. Several promising research directions have recently emerged to address these challenges. Advanced hybrid frameworks combine conventional signal processing techniques and modern approaches by providing a unified perspective that facilitates the development of robust, computationally efficient, and practically deployable speech enhancement systems for real-world applications.

Future Research Directions

The study establishes that speech enhancement has achieved remarkable progress over the past few decades; however, numerous challenges remain unresolved, particularly under highly dynamic noisy environments and real-time implementation constraints. Future research should move beyond incremental improvements to individual algorithms and focus on developing more robust, efficient, and generalizable speech enhancement frameworks. Several promising research directions are discussed below.

End-to-end hybrid cascades. The hybrid speech enhancement framework has demonstrated superior performance by combining the complementary advantages of classical signal processing and deep learning. Most existing hybrid systems consist of independently optimized stages. As a result, errors introduced during earlier processing stages may propagate to subsequent stages, thereby limiting the overall enhancement performance. Future end-to-end trainable hybrid architecture should focus on joint optimization of the multiple processing stages, including statistical filtering, adaptive processing, beamforming, feature extraction, and neural network-based enhancement. Such unified optimization allows each module to adapt according to the requirement of the entire system rather than operate independently. Consequently, end-to-end hybrid cascades are expected to provide greater robustness and generalization across diverse acoustic environments.

Self-supervised and unsupervised learning. Current supervised deep learning methods primarily rely on massive datasets containing paired noisy and clean speech recordings. The collection of such datasets is highly expensive and sometimes impractical for many languages and acoustic environments, thereby limiting the generalization capability of

supervised models when deployed under previously unseen noise conditions. In contrast, self-supervised learning offers a promising alternative by learning meaningful speech representations directly from a large collection of unlabeled speech data. These models learn latent acoustic representations through pretext tasks and are subsequently fine-tuned for speech enhancement using only a limited amount of labeled data. The development of self-supervised and unsupervised learning frameworks is expected to significantly improve adaptation to previously unseen acoustic environments while reducing dependence on large annotated datasets.

Foundation models for speech processing. Artificial intelligence has enabled the development of large-scale foundation models based on learning generalized representation from enormous speech corpora. These foundation models learn universal acoustic representations that can be efficiently transferred across multiple speech processing tasks, including speech enhancement, speaker identification, emotion recognition, and speech recognition. Such approaches could substantially reduce training time while improving robustness across diverse acoustic environments. Moreover, multimodal foundation models can jointly learn speech, text, and visual information, thereby enabling more comprehensive and robust speech enhancement systems.

Lightweight age AI implementation. Deep neural networks have achieved state-of-the-art performance, but their deployment on resource-constrained devices such as hearing aids, smartphones, and wearable electronics remains limited because of their high computational complexity, power consumption, and memory requirements. Therefore, the design of lightweight neural network architecture is highly desirable to achieve comparable performance with significantly lower computational complexity. Model compression techniques such as network pruning, weight quantization, low-rank decomposition, knowledge distillation, and efficient neural architectures are becoming increasingly important for enabling real-time speech enhancement on edge devices.

Multimodal speech enhancement. Human speech perception integrates multiple sensory cues, including auditory, contextual, and visual information. Future multimodal speech enhancement systems are expected to integrate audio signals with complementary modalities such as facial expression, lip movements, and microphone arrays. This visual speech information provides additional constraints for recovering corrupted speech. The integration of multimodal information is expected to significantly improve speech intelligibility and speaker robustness and speaker discrimination.

Uncertainty-aware speech enhancement. Most speech enhancement algorithms produce deterministic estimates of the enhanced speech signal without quantifying the confidence associated with their predictions. Under unseen acoustic conditions or severe environmental mismatches, their performance may deteriorate because the algorithm cannot distinguish reliable estimates from uncertain predictions. Research on uncertainty-aware enhancement methods, such as Bayesian deep learning, probabilistic graphical models, Monte Carlo inference, and confidence estimation frameworks, enables the simultaneous estimation of enhanced speech signals and their associated prediction uncertainty. By explicitly modeling prediction and uncertainty, these approaches can support more reliable decision-making, reduce the risk of erroneous predictions, and improve the robustness of safety-critical communication systems.

Fuzzy logic. Fuzzy logic has emerged as an effective computational intelligence approach for handling nonlinear systems and uncertain environments where precise mathematical modeling is difficult. Unlike conventional speech enhancement algorithms, fuzzy inference systems employ linguistic rules and membership functions to make adaptive decisions based on the characteristics of the input speech signals. This capability enables these systems to effectively manage uncertainties associated with time-varying noise and

changing acoustic conditions. Fuzzy logic has been investigated for adaptive threshold selection in wavelet-based denoising, dynamic adjustment of noise suppression parameters, voice activity detection, spectral gain estimation, and adaptive filtering. By incorporating expert knowledge through fuzzy IF-THEN rules, these systems can continuously modify enhancement parameters according to instantaneous acoustic conditions, thereby improving speech quality while reducing speech distortion. Furthermore, fuzzy inference systems provide greater flexibility in balancing noise suppression and speech preservation, particularly under highly nonstationary acoustic environments. Recent advances have extended fuzzy logic by integrating it with machine learning and adaptive signal processing techniques. Neuro-fuzzy systems combine the interpretability of fuzzy reasoning with the learning capability of the neural networks, while fuzzy adaptive filters dynamically optimize filter parameters according to environmental variations.

The development of synchronous fuzzy control methods and the theory of fuzzy Markov jump systems have demonstrated powerful capabilities for modeling uncertain nonlinear dynamic systems. Although these methods have been primarily investigated in control engineering, their underlying principles provide promising opportunities for speech enhancement systems operating under rapidly changing acoustic conditions.

Moreover, the future of speech enhancement lies in the convergence of classical signal processing, adaptive optimization, auditory perception, and modern artificial intelligence. The emerging directions are expected to define the next generation of speech enhancement systems. These developments will not only improve objective performance metrics such as PESQ, STOI, and SI-SDR but also facilitate robust, human-centric speech communication in complex real-world acoustic environments.

6. Conclusions

Herein, this study presents a comparative analysis of the different noise reduction approaches across multiple paradigms and provides a comprehensive classification of contemporary speech enhancement techniques. The analytical assessment reveals that no single method is universally optimum across all acoustic environments. Therefore, hybrid speech enhancement frameworks integrate complementary techniques from multiple paradigms to achieve improved robustness, speech quality, and perceptual performance under adverse acoustic conditions. Overall, this review provides a unified framework for understanding the evolution, strength, limitation, and future research directions of speech enhancement methods. It serves as a reference for researchers in designing application-specific speech management algorithms for diverse real-world acoustic environments.

Author Contributions: Conceptualization, A.S.; methodology, A.S. and P.T.; software, P.T.; validation, A.S.; formal analysis, A.S. and P.T.; investigation, A.S. and P.T.; original draft preparation, A.S. and P.T.; writing—review and editing, P.T., A.S. and R.K.G.; supervision, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study or available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the Department of Electronics and Communication Engineering, Maulana Azad National Institute of Technology, Bhopal, India, for providing the support to carry out the research work.

Conflicts of Interest: There is no conflict of interest.

References

1. Wang, D.; Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1702–1726. [CrossRef]
2. Richard, G.; Smaragdis, P.; Gannot, S.; Naylor, P.A.; Makino, S.; Kellermann, W.; Sugiyama, A. Audio signal processing in the 21st century: The important outcomes of the past 25 years. *IEEE Signal Process. Mag.* **2023**, *40*, 12–26. [CrossRef]
3. Tanwar, P.; Somkuwar, A. Hard component detection of transient noise and its removal using empirical mode decomposition and wavelet-based predictive filter. *IET Signal Process.* **2018**, *12*, 907–916. [CrossRef]
4. Stanchev, G. Technical obstacles to speech intelligibility. In Proceedings of the International Scientific and Practical Conference. Environment. Technology. Resources, Rezekne, Latvia, 24–26 June 2024; Volume 3, pp. 300–305.
5. Hepsiba, D.; Justin, J. Enhancement of single channel speech quality and intelligibility in multiple noise conditions using wiener filter and deep CNN. *Soft Comput.* **2022**, *26*, 13037–13047. [CrossRef]
6. Mahmmod, B.M.; Abdulhussain, S.H.; Ali, T.M.; Alsabah, M.; Hussain, A.; Al-Jumeily, D. Speech Enhancement: A Review of Various Approaches, Trends, and challenges. In Proceedings of the 2024 17th International Conference on Development in eSystem Engineering (DeSE), Bucharest, Romania, 10–12 November 2024; pp. 31–36.
7. Chhetri, S.; Joshi, M.S.; Mahamuni, C.V.; Sangeetha, R.N.; Roy, T. Speech enhancement: A survey of approaches and applications. In Proceedings of the 2023 2nd International Conference on Edge Computing and Applications (ICECAA), Namakkal, India, 19–21 July 2023; pp. 848–856.
8. Iqbal, Y.; Zhang, T.; Gunawan, T.S.; Pratondo, A.; Zhao, X.; Geng, Y.; Kartiwi, M.; Saleem, N.; Bourious, S. A Hybrid Speech Enhancement Technique Based on Discrete Wavelet Transform and Spectral Subtraction. *IEEE Access* **2025**, *13*, 39765–39781. [CrossRef]
9. Mansour, N.; Marschall, M.; May, T.; Westermann, A.; Dau, T. A method for realistic, conversational signal-to-noise ratio estimation. *J. Acoust. Soc. Am.* **2021**, *149*, 1559–1566. [CrossRef]
10. Hao, X.; Su, X.; Wang, Z.; Zhang, H. UNetGAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition. *arXiv* **2020**, arXiv:2010.15521.
11. Graetzer, S.; Hopkins, C. Intelligibility prediction for speech mixed with white gaussian noise at low signal-to-noise ratios. *J. Acoust. Soc. Am.* **2021**, *149*, 1346–1362. [CrossRef]
12. Kodrasi, I. Temporal envelope and fine structure cues for dysarthric speech detection using CNNs. *IEEE Signal Process. Lett.* **2021**, *28*, 1853–1857. [CrossRef]
13. Alohali, M.A.; Saleem, N.; Rhouma, D.; Medani, M.; Elmannai, H.; Bourouis, S. Temporally dynamic spiking transformer network for speech enhancement. *IEEE Access* **2024**, *12*, 146513–146526. [CrossRef]
14. O'Shaughnessy, D. Speech enhancement—A review of modern methods. *IEEE Trans. Hum.-Mach. Syst.* **2024**, *54*, 110–120. [CrossRef]
15. Das, N.; Chakraborty, S.; Chaki, J.; Padhy, N.; Dey, N. Fundamentals, present and future perspectives of speech enhancement. *Int. J. Speech Technol.* **2021**, *24*, 883–901. [CrossRef]
16. Ramonaitė, J.; Korvel, G.; Tamulevičius, G. Generative adversarial networks in speech enhancement: A survey. *IEEE Access* **2026**, *14*, 29048–29071. [CrossRef]
17. Naik, D.; Murthy, A.S.; Nuthakki, R. A literature survey on single channel speech enhancement techniques. *Int. J. Sci. Technol. Res.* **2020**, *9*, 5082–5091.
18. Drgas, S. A survey on low-latency DNN-based speech enhancement. *Sensors* **2023**, *23*, 1380. [CrossRef]
19. Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.-K.; Ren, F. Learning in the frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1740–1749.
20. Papadopoulos, P.; Tsiartas, A.; Gibson, J.; Narayanan, S. A supervised signal-to-noise ratio estimation of speech signals. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 8237–8241.
21. Miller, C.W.; Bentler, R.A.; Wu, Y.-H.; Lewis, J.; Tremblay, K. Output signal-to-noise ratio and speech perception in noise: Effects of algorithm. *Int. J. Audiol.* **2017**, *56*, 568–579. [CrossRef]
22. Kodrasi, I.; Bourlard, H. Spectro-temporal sparsity characterization for dysarthric speech detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1210–1222. [CrossRef]
23. Kates, J.M.; Arehart, K.H. The hearing-aid speech perception index (HASPI) version 2. *Speech Commun.* **2021**, *131*, 35–46. [CrossRef]
24. Phadke, K.V.; Laukkanen, A.-M.; Ilomäki, I.; Kankare, E.; Geneid, A.; Švec, J.G. Cepstral and perceptual investigations in female teachers with functionally healthy voice. *J. Voice* **2020**, *34*, 485.e433–485.e443. [CrossRef]
25. Rosenbaum, T.; Cohen, I.; Winebrand, E.; Gabso, O. Differentiable mean opinion score regularization for perceptual speech enhancement. *Pattern Recognit. Lett.* **2023**, *166*, 159–163. [CrossRef]

26. Edraki, A.; Chan, W.-Y.; Jensen, J.; Fogerty, D. Speech intelligibility prediction using spectro-temporal modulation analysis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *29*, 210–225. [CrossRef]
27. Boll, S. A spectral subtraction algorithm for suppression of acoustic noise in speech. In Proceedings of the ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing, Washington, DC, USA, 2–4 April 1979; Volume 4, pp. 200–203.
28. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [CrossRef]
29. Kassam, S.A.; Lim, T.L. Robust wiener filters. *J. Frankl. Inst.* **1977**, *304*, 171–185. [CrossRef]
30. Ephraim, Y. A minimum mean square error approach for speech enhancement. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 3–6 April 1990; pp. 829–832.
31. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **2003**, *33*, 443–445.
32. Ephraim, Y.; Van Trees, H.L. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 251–266. [CrossRef]
33. Hansen, P.C. Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. *SIAM J. Sci. Stat. Comput.* **1990**, *11*, 503–518. [CrossRef]
34. Hansen, P.C.; Sekii, T.; Shibahashi, H. The modified truncated SVD method for regularization in general form. *SIAM J. Sci. Stat. Comput.* **1992**, *13*, 1142–1150. [CrossRef]
35. Bell, A.J.; Sejnowski, T.J. Learning the higher-order structure of a natural sound. *Netw. Comput. Neural Syst.* **1996**, *7*, 261. [CrossRef]
36. Bell, A.J.; Sejnowski, T.J. The "independent components" of natural scenes are edge filters. *Vis. Res.* **1997**, *37*, 3327–3338. [CrossRef]
37. Makeig, S.; Bell, A.; Jung, T.-P.; Sejnowski, T.J. Independent component analysis of electroencephalographic data. *Adv. Neural Inf. Process. Syst.* **1995**, *8*, 145–151.
38. Bell, A.J.; Sejnowski, T.J. Learning to Find Independent Components in Natural Scenes. In *Perceptual Learning*; MIT Press: Cambridge, MA, USA, 2002; pp. 355–365.
39. Bell, A.; Sejnowski, T.J. Edges are the'Independent Components' of Natural Scenes. *Adv. Neural Inf. Process. Syst.* **1996**, *9*, 831–837.
40. Widrow, B.; McCool, J.M.; Larimore, M.G.; Johnson, C.R. Stationary and nonstationary learning characteristics of the LMS adaptive filter. *Proc. IEEE* **1976**, *64*, 1151–1162. [CrossRef]
41. Chen, Z.; Gay, S.L.; Haykin, S. Proportionate adaptation: New paradigms in adaptive filters. In *Least-Mean-Square Adaptive Filters*; Wiley: Hoboken, NJ, USA, 2003; pp. 293–334.
42. Beex, A.; Zeidler, J.R.; Haykin, S.; Widrow, B. Steady-state dynamic weight behavior in (N) LMS adaptive filters. In *Least-Mean-Square Adaptive Filters*; Wiley: Hoboken, NJ, USA, 2003; pp. 335–443.
43. Jiang, L.; Wiklund, K.; Haykin, S. A Simulink laboratory package for teaching adaptive filtering concepts. *Int. J. Eng. Educ.* **2005**, *21*, 572.
44. Hassibi, B.; Haykin, S.; Widrow, B. On the robustness of LMS filters. In *Least-Mean-Square Adaptive Filters*; Wiley: Hoboken, NJ, USA, 2003; Volume 4, pp. 105–144.
45. Haykin, S. Statistical learning theory of the LMS algorithm under slowly varying conditions, using the Langevin equation. In Proceedings of the 2006 Fortieth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 29 October–1 November 2006; pp. 229–232.
46. Haykin, S.; Sayed, A.H.; Zeidler, J.R.; Yee, P.; Wei, P.C. Adaptive tracking of linear time-variant systems by extended RLS algorithms. *IEEE Trans. Signal Process.* **2002**, *45*, 1118–1128.
47. Haykin, S.; Sayed, A.; Zeidler, J.; Yee, P.; Wei, P. Tracking of linear time-variant systems. In Proceedings of the Proceedings of MILCOM'95, San Diego, CA, USA, 5–8 November 1995; Volume 2, pp. 602–606.
48. Haykin, S. Adaptive systems for signal process. In *Advanced Signal Processing*; CRC Press: Boca Raton, FL, USA, 2017; pp. 25–78.
49. Haykin, S.S. *Adaptive Filter Theory*; Pearson Education: Tamil Nadu, India, 2008.
50. Allen, J. Applications of the short time Fourier transform to speech processing and spectral analysis. In Proceedings of the ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, France, 3–5 May 1982; Volume 7, pp. 1012–1015.
51. Donoho, D.L. Superresolution via sparsity constraints. *SIAM J. Math. Anal.* **1992**, *23*, 1309–1331. [CrossRef]
52. Flandrin, P.; Rilling, G.; Goncalves, P. Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett.* **2004**, *11*, 112–114. [CrossRef]
53. Rilling, G.; Flandrin, P. One or two frequencies? The empirical mode decomposition answers. *IEEE Trans. Signal Process.* **2007**, *56*, 85–95. [CrossRef]
54. Rilling, G.; Flandrin, P.; Gonçalves, P.; Lilly, J.M. Bivariate empirical mode decomposition. *IEEE Signal Process. Lett.* **2007**, *14*, 936–939. [CrossRef]

55. Flandrin, P.; Gonçalves, P.; Rilling, G. EMD equivalent filter banks, from interpretation to applications. In *Hilbert–Huang Transform and Its Applications*; World Scientific: London, UK, 2014; pp. 99–116.
56. Flandrin, P.; Goncalves, P.; Rilling, G. Detrending and denoising with empirical mode decompositions. In Proceedings of the 2004 12th European Signal Processing Conference, Vienna, Austria, 6–10 September 2004; pp. 1581–1584.
57. Rilling, G.; Flandrin, P. Sampling effects on the empirical mode decomposition. *Adv. Adapt. Data Anal.* **2009**, *1*, 43–59. [CrossRef]
58. Atal, B.; Remde, J. A new model of LPC excitation for producing natural-sounding speech at low bit rates. In Proceedings of the ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, France, 3–5 May 1982; Volume 7, pp. 614–617.
59. Atal, B. Efficient coding of LPC parameters by temporal decomposition. In Proceedings of the ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing, Boston, MA, USA, 14–16 April 1983; Volume 8, pp. 81–84.
60. Atal, B.S. A model of LPC excitation in terms of eigenvectors of the autocorrelation matrix of the impulse response of the LPC filter. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK, 23–26 May 1989; pp. 45–48.
61. Singhal, S.; Atal, B. Improving performance of multi-pulse LPC coders at low bit rates. In Proceedings of the ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing, San Diego, CA, USA, 19–21 March 1984; Volume 9, pp. 9–12.
62. Caspers, B.E.; Atal, B.S. Changing pitch and duration in LPC synthesized speech using multipulse excitation. *J. Acoust. Soc. Am.* **1983**, *73*, S5. [CrossRef]
63. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
64. Gannot, S.; Burshtein, D.; Weinstein, E. Iterative and sequential Kalman filter-based speech enhancement algorithms. *IEEE Trans. Speech Audio Process.* **2002**, *6*, 373–385.
65. Goh, Z.; Tan, K.-C.; Tan, B. Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model. *IEEE Trans. Speech Audio Process.* **1999**, *7*, 510–524. [CrossRef]
66. Paliwal, K.; Basu, A. A speech enhancement method based on Kalman filtering. In Proceedings of the ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, USA, 6–9 April 1987; Volume 12, pp. 177–180.
67. Gannot, S.; Yeredor, A. The Kalman filter. In *Springer Handbook of Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 135–160.
68. Wu, W.-R.; Chen, P.-C. Subband Kalman filtering for speech enhancement. *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.* **1998**, *45*, 1072–1083. [CrossRef]
69. So, S.; Paliwal, K.K. Modulation-domain Kalman filtering for single-channel speech enhancement. *Speech Commun.* **2011**, *53*, 818–829. [CrossRef]
70. Gustafsson, S.; Martin, R.; Jax, P.; Vary, P. A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. *IEEE Trans. Speech Audio Process.* **2002**, *10*, 245–256. [CrossRef]
71. Isoyama, T.; Kidani, S.; Unoki, M. Computational models of auditory sensation important for sound quality on basis of either gammatone or gammachirp auditory filterbank. *Appl. Acoust.* **2024**, *218*, 109914. [CrossRef]
72. Johnston, J.D. Transform coding of audio signals using perceptual noise criteria. *IEEE J. Sel. Areas Commun.* **1988**, *6*, 314–323. [CrossRef]
73. Brown, G.J.; Wang, D. Modelling the perceptual segregation of double vowels with a network of neural oscillators. *Neural Netw.* **1997**, *10*, 1547–1558. [CrossRef]
74. Wang, D.L.; Brown, G.J. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. Neural Netw.* **1999**, *10*, 684–697. [CrossRef]
75. Brown, G.J.; Wang, D. Separation of speech by computational auditory scene analysis. In *Speech Enhancement*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 371–402.
76. Kim, Y.; Kim, J.-S.; Kim, G.-W. A novel frequency selectivity approach based on travelling wave propagation in mechanoluminescence basilar membrane for artificial cochlea. *Sci. Rep.* **2018**, *8*, 12023. [CrossRef]
77. Patterson, R.D.; Holdsworth, J. A functional model of neural activity patterns and auditory images. *Adv. Speech Hear. Lang. Process.* **1996**, *3*, 547–563.
78. Irino, T.; Patterson, R.D. Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. *Speech Commun.* **2002**, *36*, 181–203. [CrossRef]
79. Li, K.; Zaman, K.; Li, X.; Akagi, M.; Dang, J.; Unoki, M. Machine anomalous sound detection using spectral-temporal modulation representations derived from machine-specific filterbanks. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 2059–2073. [CrossRef]
80. Peng, Z.; Li, X.; Zhu, Z.; Unoki, M.; Dang, J.; Akagi, M. Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access* **2020**, *8*, 16560–16572. [CrossRef]

81. Zohar, E.; Nelken, I.; Rafaely, B. An Efficient Neural Network for Modeling Human Auditory Neurograms for Speech. In Proceedings of the ICASSP 2026—2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2026; pp. 15617–15621.
82. Ososkov, G.; Goncharov, P. Shallow and deep learning for image classification. *Opt. Mem. Neural Netw.* **2017**, *26*, 221–248. [CrossRef]
83. Xu, S.; Song, Y.; Hao, X. A comparative study of shallow machine learning models and deep learning models for landslide susceptibility assessment based on imbalanced data. *Forests* **2022**, *13*, 1908. [CrossRef]
84. Akinci, T.C.; Topsakal, O.; Akbas, M.I. Machine Learning Methods from Shallow Learning to Deep Learning. In *Shallow Learning vs. Deep Learning: A Practical Guide for Machine Learning Solutions*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–28.
85. Lee, D.; Seung, H.S. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* **2000**, *13*, 556–562.
86. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef]
87. Povey, D.; Burget, L.; Agarwal, M.; Akyazi, P.; Feng, K.; Ghoshal, A.; Glembek, O.; Goel, N.K.; Karafiát, M.; Rastrow, A. Subspace Gaussian mixture models for speech recognition. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 4330–4333.
88. Ververidis, D.; Kotropoulos, C. Emotional speech classification using Gaussian mixture models. In Proceedings of the 2005 IEEE International Symposium on Circuits and Systems, Kobe, Japan, 23–26 May 2005; pp. 2871–2874.
89. Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* **2000**, *10*, 19–41. [CrossRef]
90. Xu, Y.; Du, J.; Dai, L.-R.; Lee, C.-H. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process. Lett.* **2013**, *21*, 65–68. [CrossRef]
91. Xu, Y.; Du, J.; Dai, L.-R.; Lee, C.-H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *23*, 7–19. [CrossRef]
92. Xu, Y.; Du, J.; Dai, L.-R.; Lee, C.-H. Dynamic noise aware training for speech enhancement based on deep neural networks. In Proceedings of the Interspeech, Singapore, 14–18 September 2014; pp. 2670–2674.
93. Zhao, Y.; Xu, B.; Giri, R.; Zhang, T. Perceptually guided speech enhancement using deep neural networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5074–5078.
94. Xu, L.; Zhang, T. Fractional feature-based speech enhancement with deep neural network. *Speech Commun.* **2023**, *153*, 102971. [CrossRef]
95. Du, J.; Xu, Y. Hierarchical deep neural network for multivariate regression. *Pattern Recognit.* **2017**, *63*, 149–157. [CrossRef]
96. Wang, J.; Yu, L.-C.; Lai, K.R.; Zhang, X. Dimensional sentiment analysis using a regional CNN-LSTM model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, 7–12 August 2016; pp. 225–230.
97. Song, X.; Yang, F.; Wang, D.; Tsui, K.-L. Combined CNN-LSTM network for state-of-charge estimation of lithium-ion batteries. *IEEE Access* **2019**, *7*, 88894–88902. [CrossRef]
98. Lu, W.; Li, J.; Li, Y.; Sun, A.; Wang, J. A CNN-LSTM-based model to forecast stock prices. *Complexity* **2020**, *2020*, 6622927. [CrossRef]
99. Li, C.; Wang, P.; Wang, S.; Hou, Y.; Li, W. Skeleton-based action recognition using LSTM and CNN. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017; pp. 585–590.
100. Coto-Jimenez, M.; Goddard-Close, J.; Di Persia, L.; Rufiner, H.L. Hybrid speech enhancement with wiener filters and deep lstm denoising autoencoders. In Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), San Carlos, Costa Rica, 18–20 July 2018; pp. 1–8.
101. Yu, H.; Zhu, W.-P.; Champagne, B. Speech enhancement using a DNN-augmented colored-noise Kalman filter. *Speech Commun.* **2020**, *125*, 142–151. [CrossRef]
102. Cheng, R.; Bao, C. Speech Enhancement Based on Beamforming and Post-Filtering by Combining Phase Information. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 4496–4500.
103. Kuwałek, P.; Jęśko, W. Speech Enhancement Based on Enhanced Empirical Wavelet Transform and Teager Energy Operator. *Electronics* **2023**, *12*, 3167. [CrossRef]
104. Lan, C.; Chen, H.; Zhang, L.; Zhao, S.; Guo, R.; Fan, Z. Research on Speech Enhancement Algorithm by Fusing Improved EMD and GCRN Networks. *Circuits Syst. Signal Process.* **2024**, *43*, 4588–4604. [CrossRef]
105. Du, K.-L.; Swamy, M. Nonnegative matrix factorization. In *Neural Networks and Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 427–445.
106. Noh, K.; Chang, J.-H. Joint optimization of deep neural network-based dereverberation and beamforming for sound event detection in multi-channel environments. *Sensors* **2020**, *20*, 1883. [CrossRef]

107. Yang, Y.; Trigoni, N.; Markham, A. Pre-training feature guided diffusion model for speech enhancement. *arXiv* **2024**, arXiv:2406.07646.
108. Wu, T.; He, S.; Zhang, H.; Zhang, X. ScaleFormer: Transformer-based speech enhancement in the multi-scale time domain. In Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, China, 31 October–3 November; pp. 2448–2453.
109. Hu, Y.; Loizou, P.C. A comparative intelligibility study of speech enhancement algorithms. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 15–20 April 2007; pp. IV-561–IV-564.
110. Zhao, N.; Li, R. EMD method applied to identification of logging sequence strata. *Acta Geophys.* **2015**, *63*, 1256–1275. [CrossRef]
111. Lim, J. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE* **1962**, *67*, 1689–1697.
112. Hansen, P.C. The truncated SVD as a method for regularization. *BIT Numer. Math.* **1987**, *27*, 534–553. [CrossRef]
113. Bell, A.J.; Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **1995**, *7*, 1129–1159. [CrossRef]
114. Widrow, B.; Glover, J.R.; McCool, J.M.; Kaunitz, J.; Williams, C.S.; Hearn, R.H.; Zeidler, J.R.; Dong, J.E.; Goodlin, R.C. Adaptive noise cancelling: Principles and applications. *Proc. IEEE* **1975**, *63*, 1692–1716. [CrossRef]
115. Eleftheriou, E.; Falconer, D. Tracking properties and steady-state performance of RLS adaptive filter algorithms. *IEEE Trans. Acoust. Speech Signal Process.* **2003**, *34*, 1097–1110.
116. Scalart, P. Speech enhancement based on a priori signal to noise estimation. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 7–10 May 1996; Volume 2, pp. 629–632.
117. Donoho, D.L. De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **1995**, *41*, 613–627. [CrossRef]
118. Rilling, G.; Flandrin, P.; Goncalves, P. On empirical mode decomposition and its algorithms. In Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03, Grado, Italy, 8–11 June 2003.
119. Atal, B.S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Acoust. Soc. Am.* **1974**, *55*, 1304–1312. [CrossRef]
120. Koo, B.; Gibson, J.D.; Gray, S.D. Filtering of colored noise for speech enhancement and coding. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Glasgow, Scotland, UK, 23–26 May 1989; pp. 349–352.
121. Wang, D.; Brown, G.J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*; Wiley-IEEE Press: Hoboken, NJ, USA, 2006.
122. Srinivasan, S.; Samuelsson, J.; Kleijn, W.B. Codebook-based Bayesian speech enhancement for nonstationary environments. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 441–452. [CrossRef]
123. Weninger, F.; Hershey, J.R.; Le Roux, J.; Schuller, B. Discriminatively trained recurrent neural networks for single-channel speech separation. In Proceedings of the 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, GA, USA, 3–5 December 2014; pp. 577–581.
124. Wang, K.; He, B.; Zhu, W.-P. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7098–7102.
125. Ephraim, Y.; Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 1109–1121. [CrossRef]
126. Allen, J.B.; Rabiner, L.R. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE* **1977**, *65*, 1558–1564. [CrossRef]
127. Lu, L.; Yin, K.-L.; de Lamare, R.C.; Zheng, Z.; Yu, Y.; Yang, X.; Chen, B. A survey on active noise control in the past decade—Part I: Linear systems. *Signal Process.* **2021**, *183*, 108039. [CrossRef]
128. Luo, Z.; Shi, D.; Gan, W.-S. A hybrid sfanc-fxnlms algorithm for active noise control based on deep learning. *IEEE Signal Process. Lett.* **2022**, *29*, 1102–1106. [CrossRef]
129. Tang, Y.; Zhang, H.; Zhang, Y. Stability guaranteed active noise control: Algorithms and applications. *IEEE Trans. Control Syst. Technol.* **2023**, *31*, 1720–1732. [CrossRef]
130. Tian, X.; Huang, J.; Feng, X.; Shen, Y. An intermittent FxLMS algorithm for active noise control systems with saturation nonlinearity. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2347–2356. [CrossRef]
131. Munir, M.W.; Abdulla, W.H. On FxLMS scheme for active noise control at remote location. *IEEE Access* **2020**, *8*, 214071–214086. [CrossRef]
132. Wu, Z.; Yang, J.; Yan, L.; Niu, A. Regularized Autocorrelated Errors Variable Step Size FxLMS Algorithm-Based Active Noise Control. *IEEE Trans. Ind. Electron.* **2025**, *73*, 4850–4861. [CrossRef]
133. Ardekani, I.T.; Abdulla, W.H. Theoretical convergence analysis of FxLMS algorithm. *Signal Process.* **2010**, *90*, 3046–3055. [CrossRef]

134. Song, P.; Zhao, H. Filtered-x least mean square/fourth (FXLMS/F) algorithm for active noise control. *Mech. Syst. Signal Process.* **2019**, *120*, 69–82. [CrossRef]
135. Pang, Y.; Ou, S.; Cai, Z.; Gao, Y. Posterior Error Energy Minimization Based Combined FxNLMS and FxLMS Algorithm for Active Noise Control. *IEEE Access* **2024**, *12*, 30754–30764. [CrossRef]
136. Wang, X.; Ou, S.; Pang, Y. Adaptive combination of filtered-X NLMS and affine projection algorithms for active noise control. In Proceedings of the CAAI International Conference on Artificial Intelligence, Beijing, China, 5–6 August 2022; pp. 15–25.
137. Zhang, X.; Yang, S.; Liu, Y.; Zhao, W. Improved variable step size least mean square algorithm for pipeline noise. *Sci. Program.* **2022**, *2022*, 3294674. [CrossRef]
138. Althahab, A.Q.J.; Ma, H.; Vuksanovic, B. Addressing modelling errors in feedforward ANC systems: A new normalised semi-variable step size FxLMS algorithm. *Appl. Acoust.* **2025**, *233*, 110602. [CrossRef]
139. Huang, B.; Xiao, Y.; Sun, J.; Wei, G. A variable step-size FXLMS algorithm for narrowband active noise control. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *21*, 301–312. [CrossRef]
140. Wang, J.; Liao, J.; He, L.; Tan, X.; Chen, Z. Variable Step-Size FxLMS Algorithm Based on Cooperative Coupling of Double Nonlinear Functions. *Symmetry* **2025**, *17*, 1222. [CrossRef]
141. Chien, Y.-R.; Yu, C.-H.; Tsao, H.-W. Affine-projection-like maximum correntropy criteria algorithm for robust active noise control. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2255–2266. [CrossRef]
142. Liu, Z.; Zhao, Z.; Qiu, Y.; Jing, B.; Yang, C. State of charge estimation for Li-ion batteries based on iterative Kalman filter with adaptive maximum correntropy criterion. *J. Power Sources* **2023**, *580*, 233282. [CrossRef]
143. Zhou, E.; Xia, B.; Li, E.; Wang, T. An efficient algorithm for impulsive active noise control using maximum correntropy with conjugate gradient. *Appl. Acoust.* **2022**, *188*, 108511. [CrossRef]
144. Kumar, K.; Karthik, M.L.S.; George, N.V. A robust active noise control system based on an exponential hyperbolic cosine norm. *Signal Process.* **2024**, *221*, 109469. [CrossRef]
145. Long, X.; Zhao, H.; Hou, X. Hybrid geometric algebra correntropy: Definition and application to robust adaptive filtering. *IEEE Trans. Circuits Syst. II Express Briefs* **2023**, *71*, 952–956. [CrossRef]
146. Mu, Z.; Gao, Y.; Guo, X.; Ou, S. Variable Step-Size Hybrid Filtered-x Affine Projection Generalized Correntropy Algorithm for Active Noise Control. *Sensors* **2025**, *25*, 1881. [CrossRef]
147. Tanwar, P.; Somkuwar, A.; Gumasta, R.K. Distribution-Aware, Risk-Sensitive (DA-RS-FxNLMS) Active Noise Control for Non-Gaussian Acoustic Environments. *Acoustics* **2026**, *8*, 36. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.