

Review

An In-Depth Review of Speech Enhancement Algorithms: Classifications, Underlying Principles, Challenges, and Emerging Trends

Nisreen Talib Abdulhusein [†] and Basheera M. Mahmmod ^{*,†} 

Department of Computer Engineering, University of Baghdad, Al-Jadriya, Baghdad 10071, Iraq

* Correspondence: basheera.m@coeng.uobaghdad.edu.iq

[†] These authors contributed equally to this work.

Abstract

Speech enhancement aims to improve speech quality and intelligibility in noisy environments and is important in applications such as hearing aids, mobile communications and automatic speech recognition (ASR). This paper shows a structured review of speech enhancement techniques, classified depending on the channel configuration and signal processing framework. Both traditional and modern approaches are discussed, including classical signal processing methods, machine learning techniques, and recent deep learning-based models. Furthermore, common noise types, widely used speech datasets, and standard evaluation metrics for evaluating speech quality and intelligibility are reviewed. Key challenges such as non-stationary noise, data limitations, reverberation, and generalization to unseen noise conditions are highlighted. This review presents the advancements in speech enhancement and discusses the challenges and trends of this field. Valuable insights are provided for researchers, engineers, and practitioners in the area. The findings aid in the selection of suitable techniques for improved speech quality and intelligibility, and we concluded that the trend in speech enhancement has shifted from standard algorithms to deep learning methods that can efficiently learn information regarding speech signals.

Keywords: speech enhancement algorithm (SEA); single- and multi-channel; traditional methods; deep learning; noise; objectives measurements

1. Introduction

Speech is considered the most important form of communication between humans, and it depends on auditory perception. It is used to communicate information from a speaker to one or more listeners, where the speaker produces a speech signal in the form of pressure waves that travel from speaker's mouth to listener's ears. The signal involves a distinction in pressure as a function of time and is measured directly in front of the mouth, the primary sound source [1]. Speech properties are often influenced by the diverse types of noise around us, and this noise distorts and reduces its quality and intelligibility. Therefore, speech enhancement has become crucial and can be defined as a process used to improve perceptual aspects of speech-like quality and intelligibility, or the degree of listener fatigue. Reducing noise, enhancing audio clarity, and avoiding listener fatigue or discomfort are the objectives of speech enhancement [2]. The general process of speech enhancement algorithm is shown in Figure 1.



Academic Editor: Maryam Ravan

Received: 6 January 2026

Revised: 1 February 2026

Accepted: 4 February 2026

Published: 7 February 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

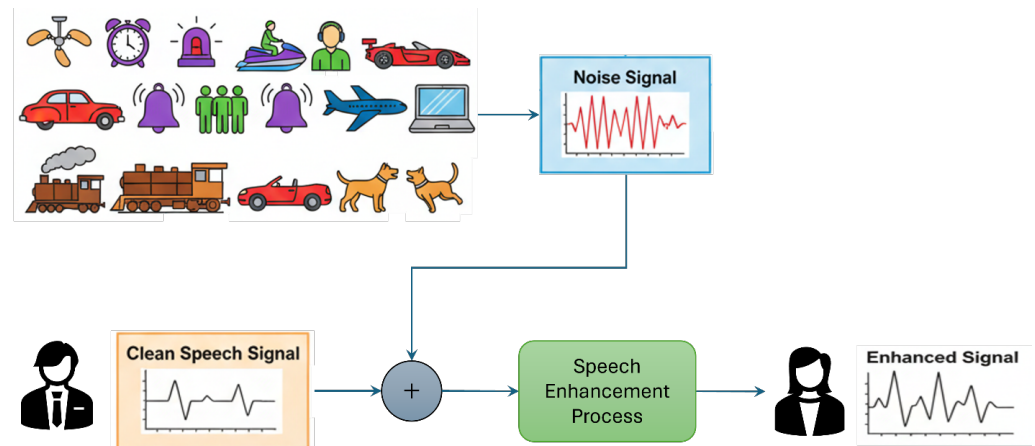


Figure 1. The general process of speech enhancement algorithms.

In the field of speech enhancement, speech quality refers to the perceptual attributes of the enhanced speech signal, such as naturalness, distortion, and residual noise, as perceived by human listeners [3]. It is commonly evaluated using metrics such as the Perceptual Evaluation of Speech Quality (PESQ) and the Mean Opinion Score (MOS). In contrast, speech intelligibility measures the degree to which a listener can correctly understand or recognize spoken words, particularly in noisy environments, and is typically assessed using objective and subjective measures such as the Short-Time Objective Intelligibility (STOI) metric.

Speech enhancement algorithms (SEAs) aim to suppress background noise in order to improve both speech quality and intelligibility. Improving speech quality is especially important for reducing listener fatigue in prolonged noisy conditions, such as industrial or manufacturing environments, while intelligibility is critical for effective communication in daily and professional applications.

Noise refers to unwanted signals that interfere with meaningful speech. Although it is generally considered undesirable, it also conveys information about its source and the surrounding acoustic environment, and is commonly present in everyday environments such as streets, vehicles, restaurants, offices, and public spaces [4]. Environmental noise can be broadly categorized based on temporal characteristics, including continuous, intermittent, and impulsive noise, or based on spectral characteristics, often referred to as noise “colors” such as white, pink, and brown noise. These noise types have different temporal and spectral properties, which significantly affect the performance of speech enhancement algorithms [5]. Consequently, effective noise reduction plays an important role in a wide range of applications [4], including hearing aids, security systems, industrial monitoring, mobile and VoIP communications, and biomedical signal processing [5].

Over the years, numerous speech enhancement techniques have been developed, ranging from classical signal processing methods such as spectral subtraction [6], Wiener filtering [7], MMSE-based estimators [8] and Signal Subspace [9] to modern machine learning and deep learning approaches [10]. Traditional methods rely on mathematical and statistical models of speech and noise, whereas data-driven models, including convolutional and recurrent neural networks, learn complex mappings directly from data and have shown improved performance in non-stationary noise environments. Studies have explained that deep neural networks (DNNs) result cleaner and more natural-sounding speech, which is why they are now widely used in voice assistants (like Siri and Google Assistant), smart hearing aids, and modern communication systems [11]. In addition, various time–frequency transforms and hybrid representations have been employed to facilitate speech–noise separation [12–20].

Speech enhancement serves as a fundamental pre-processing stage for many speech processing applications, including automatic speech recognition (ASR), speaker recognition (SR), speech synthesis or text-to-speech [21], telecommunications, hearing aids, and security systems. Despite significant progress, achieving robust noise suppression without introducing speech distortion remains challenging, especially in highly dynamic or unseen noise conditions. Issues such as statistical modeling accuracy, computational complexity, and real-time applicability continue to limit practical deployment.

Because of the distinctive characteristics of noise over time, it is not easy to enhance speech in a noisy environment. Until now, removing noise from noisy speech has been a difficult issue because spectral packages of non-stationary noise are very sensitive to estimate and prediction. Noise calculation is a dangerous issue when noise power is greater than speech power because speech content may be eliminated when treating it as noise [5]. Choosing the appropriate probability density function (PDF) to represent speech and noise signals plays a critical role in determining the performance of speech enhancement algorithms. Since both speech and noise are random and often non-stationary signals, their statistical modeling must precisely characterize how their energy is distributed across time and frequency domains. In traditional models, the Gaussian probability density function (PDF) has been widely used because of its simplicity and mathematical ease of use.

However, Gaussian models often fail to define the sharp variations and complex behavior of real speech signals, particularly in highly noisy or non-stationary environments. To address this limitation, researchers have approved Laplacian, gamma, and hyper-Gaussian probability density distributions, which are better at taking in dynamic nature of speech components compared to noise [4].

With the increasing interest in SE, several previous articles have attempted to summarize the various methods used for SE. While there have been several previous reviews in the field of SE, most have focused on a specific aspect of algorithms, such as traditional methods or deep learning model, particularly in terms of architecture, and have often been descriptive without providing a comprehensive critical analysis.

This work differs from previous reviews in that it offers a thorough and integrated review of speech enhancement algorithms from multiple perspectives simultaneously. It does not just classify the algorithms, but explains the main principles upon which they are based, analyzes the strengths and weaknesses of each category, and focuses on practical challenges such as generalization over unknown noise, computational complexity, and real-time requirements. However, many existing surveys provide descriptive overviews or focus on specific algorithm families without jointly discussing noise types, benchmark datasets, performance evaluation metrics, and practical challenges.

Contributions of This Review: This paper presents a comprehensive review of classical and modern speech enhancement algorithms, summarizes commonly used noise types, datasets, and performance evaluation metrics, and discusses key challenges and emerging research trends in the field. The aim is to provide researchers and practitioners with a structured and up-to-date reference that highlights current limitations and future research directions.

The subsequent structure of this paper is arranged as follows: Section 2 will review SEAs and clarify the entry point of this paper; Section 3 will introduce definition of noise and its types; Section 4 presents a detailed description of different types of SEA datasets; Section 5 discusses the quality and intelligibility assessment measure in speech enhancement; Section 6 presents the key applications for speech enhancement technologies; Section 7 reviews the limitations and challenges of SEA; and Section 8 summarizes the entire paper and looks forward to future work in the field of speech enhancement.

2. Speech Enhancement Algorithms (SEAs)

SEAs play an important role in the reduction of noise to achieve clear and improved speech, a result that in turn affects many speech applications such as hearing aids and audio processing devices. Increasing the quality and intelligibility of the speech will make the corrupted speech more agreeable to the listener. Many algorithms attempt to improve speech quality at the expense of intelligibility, where an improvement in quality is mostly followed by a decrease in clarity. This is because of the distortion detected on the clean speech signal resulting from the disproportionate suppression of the noisy signal. Here, it is necessary to improve both the quality of the speech and the speech's intelligibility simultaneously in order to make the corrupted speech more pleasant to the listener [4]. Speech enhancement tasks are part of speech signal processing, and tasks such as noise reduction, echo cancellation, dereverberation, source separation, and speech separation are clearly presented within the framework of speech enhancement in general, and aim to improve speech quality and intelligibility. The scope of this review will focus on speech enhancement algorithms (SEA) (noise reduction type), which removes background noise from the speech signal. Based on that, SEAs can be classified into different categorizations. One of the most important categorizations is based on the number of available channels, which can be divided into two categories: single-channel SEA and multi-channel SEA [22,23].

Most single-channel speech enhancement methods have traditionally depended on digital signal processing (DSP) techniques, such as spectral subtraction, Wiener filtering, and mean square error (MMSE) estimators (which use clean speech and noise power spectral density estimates obtained from the noisy signal). These algorithms were effective at reducing stationary noise, but they encountered difficulties when dealing with non-stationary noise environments, such as background speech or traffic noise, because of their strong assumptions about the statistical properties of noise and speech signals [4]. To exceed these limitations, researchers began integrating data-driven learning methods, leading to the appeal of deep neural networks (DNNs) for speech enhancement. Contrary to traditional methods that depend on static mathematical models, DNNs learn directly from large datasets of noisy and clear speech pairs, allowing them to adapt to various noise conditions and form complex nonlinear relationships between noisy and clear signals [10]. This transformation from manual feature extraction to automatic feature learning has led to important enhancements in speech quality and intelligibility under realistic conditions [24]. According to this improvement, SEA-based multi-channel approaches have seen great enhancements in terms of their performance and improved system strength by adding additional microphones. Using multiple channels allows the system to take spatial cues and realistic noise samples from different directions, enhancing the ability to suppress noise more effectively while preserving the speech source [25].

The traditional technologies used for noise removal can be categorized depending on processing domains into transform domain-based SEAs, time domain-based SEAs, and hybrid domain-based SEAs [23]. The transform-based domain has, in turn, several types depending on the processing techniques which are as follows: spectral-subtractive algorithms, methods based on statistical models, and subspace criteria. Statistical methods can be further divided into many types such as the Short-Time Spectral Amplitude Minimum Mean Square Error Estimator (STSA-MMSE), the Maximum A Posteriori Estimator (MAP), the Wiener filter, and the Kalman filter. This paper will deal with single- and multi-channel types in a clear detail, explaining the basic types that deal with these methods, how they work, and their special characteristics, which will be explained in the following paragraphs. A general taxonomy of SEA based on the literature review of this paper is shown in Figure 2.

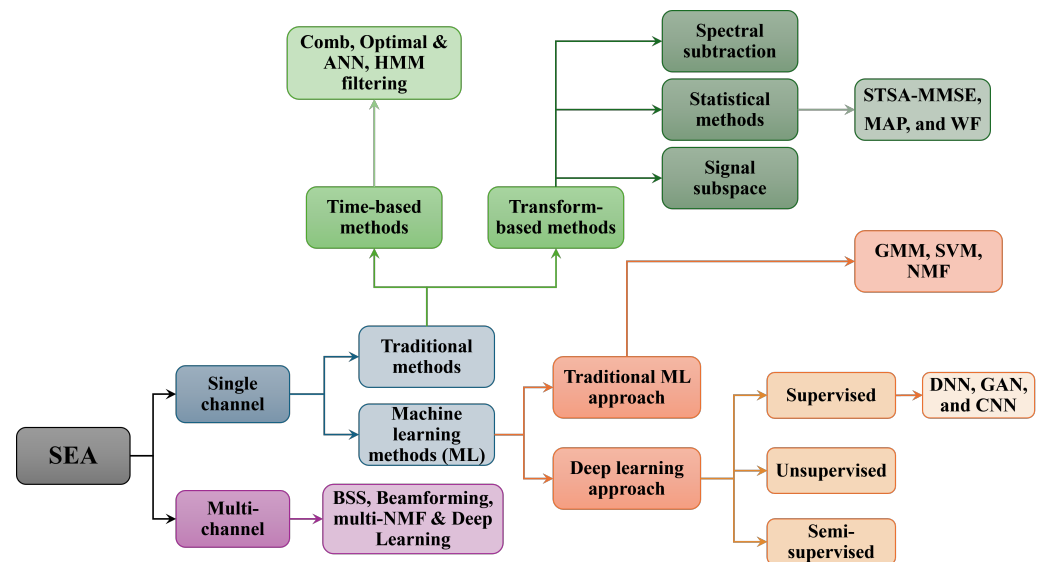


Figure 2. A general taxonomy of speech enhancement algorithms.

2.1. Mathematical Formulation of Speech Enhancement Problem

In single-channel speech enhancement, the observed noisy speech signal $y[n]$ is modeled as the sum of clean speech $x[n]$ and additive noise $v[n]$:

$$y[n] = x[n] + v[n] \quad (1)$$

The objective of SEA is to estimate the clean speech signal $\hat{x}[n]$ from the noisy observation $y[n]$:

$$\hat{x}[n] = f(y[n]). \quad (2)$$

where $f(\cdot)$ is the enhancement algorithm. In multi-channel speech enhancement, the noisy observation is captured by multiple microphones:

$$Y[n] = X[n] + V[n] \quad (3)$$

where $y[n] = [y_1[n], y_2[n], \dots, y_M[n]]^T$, M is the number of microphones. The objective is to estimate the clean speech signal $\hat{X}[n]$:

$$\hat{X}[n] = g(Y[n]) \quad (4)$$

where $g(\cdot)$ is the enhancement algorithm.

2.2. Single-Channel Speech Enhancement Algorithms (SC-SEA)

Single-channel SEA uses only a single audio signal from a single microphone to improve audio quality by suppressing noise and distortion while preserving essential speech information. This approach is widely used in everyday applications such as telephones (VoIP, Zoom, and Skype), voice assistants (Google Assistant and Siri), hearing aids [26], and (ASR). This method depends on only a single channel without relying on spatial information such as in multi-channel systems [4]. Researchers in the field of single-channel speech enhancement (SC-SE) usually divide enhancement algorithms into two main groups: those based on traditional signal processing methods and other based on machine learning methods. This broad classification is commonly used in the literature because it classes algorithms by their underlying modeling philosophy, whether they use analytical signal models or data-driven learning models.

2.2.1. Traditional Speech Enhancement Algorithms

Traditional methods use clear mathematical or statistical models of clean speech and noise signals. For decades, these methods were the most important in this field, and they are still important sources of information and speech enhancement research [4]. They are usually split into two subgroups based on the processing domain:

1. Methods Based on Transform Domain

These methods are based on the transform domain, where the signals are transformed into another domain to show the hidden properties that are used in the desired processing. There are many different transforms that are applied in SE such as STFT, DCT, or wavelet transforms. The most significant SEA that are applied in transform domain are listed below:

A. Spectral Subtraction (SS)

The method of spectral subtraction (SS) is historically one of the first algorithms suggested for noise reduction [27]. Its operation is based on a simple principle: if we assume that noise is simply added to speech, we can obtain pure speech by subtracting the estimated noise spectrum from the noisy spectrum. To do this, the noise spectrum is estimated (and updated) during the moments when no one is talking [4]. This technique performs the reduction of noise based on precise statistics that estimate the noise energy spectrum by finding the smallest value for a smooth energy spectrum of a distorted speech signal. The spectral-subtractive algorithms are evaluated for a wide range of noise types at varying volume levels using a single-channel speech source and they are suitable for voice communication methods [28].

B. Statistical Methods

Statistical speech enhancement methods are a major subclass of transform domain single-channel enhancement algorithms. These methods depend on the statistical modeling of clean speech and noise in the spectral domain (such as STFT or DCT). They estimate the clean speech spectrum by applying optimal estimators derived from Bayesian estimation theory, MMSE theory, or MAP-based inference. These methods are the backbone of classical speech enhancement techniques [4] and include the following:

STSA-MMSE

The Short-Time Spectral Amplitude Minimum Mean Square Error Estimator (STSA-MMSE) treats speech enhancement as a statistical estimation issue in the time–frequency domain. In the STFT domain, the noisy signal is modeled as the sum of clean speech and extra noise. This is based on the idea that noise is commonly assumed to follow a Gaussian distribution, while the speech's spectral amplitude is modeled using appropriate statistical priors. A closed-form optimal gain function was derived that minimizes the predicted squared error between the actual and estimated spectral amplitudes, predicated on this assumption [29]. One of the best things about the STSA-MMSE is that it becomes cleared of noise without adding musical noise artifacts, which is a problem with spectral subtraction algorithms. The estimator combines a priori and a posteriori SNRs to do this. The estimator proved to be of significantly enhanced perceptual quality compared to traditional spectral subtraction methods.

The STSA-MMSE estimator is a basis for numerous prevalent statistical speech enhancement techniques, such as the Minimum Mean Square Error Log-Spectral Amplitude Estimator (MMSE-LSA) [29] and the Optimally Modified LSA estimator [30]. The STSA-MMSE is still one of the most important methods for improving single-channel speech and is thought to be a key part of noise sup-

pression based on statistical models. Mojtaba Bahrami and Neda Faraji [31] presented a model that consider a modern extension of the STSA-MMSE algorithm, which depends on modeling the amplitude of the speech spectrum using the Weibull distribution with the inclusion of uncertainty in the presence of speech, resulting in the improved performance of the STSA-MMSE algorithm compared to traditional statistical models.

Maximum A Posteriori (MAP)

The MAP estimator is a statistical model-based speech enhancement technique that estimates the clean speech spectral components by maximizing the posterior probability depending on the noisy observation. Unlike MMSE, which computes the expectation of the posterior distribution, MAP selects the most probable value. Assuming that noise is Gaussian and speech is a statistical model, the MAP estimator yields gain functions dependent on the a priori and a posteriori SNRs. When the posterior distribution is symmetric, the MAP estimator and the MMSE produce identical results; otherwise, MAP provides another practical choice with a little stronger suppression at low SNR conditions [4]. Thomas Lotter and Peter Vary in [32] presented two speech spectral amplitude estimators based on maximum a posteriori estimation (MAP) using the super-Gaussian statistical modeling of speech Fourier coefficient amplitudes. The model assumes that speech's spectral amplitudes follow distributions such as Laplace or gamma rather than the traditional Gaussian distribution. The results show that the proposed estimators achieve better speech quality optimization compared to the traditional algorithms such as STSA-MMSE and MMSE-LSA, which assume Gaussian speech and noise models. Raziye Ranzbaryan and Hamid Reza Abutalebi [33] proposed a multi-frame MAP algorithm to improve speech quality using a single microphone by utilizing the temporal correlation between speech parameters in the STFT domain. This is realized by using the current frame and a limited number of preceding frames at each time–frequency unit. To form this correlation, a complex factor is employed to express the temporal relationship between the speech parameters in adjacent frames. Multi-frame MAP estimators are then derived to enhance speech and reduce noise. The results prove that this method outperforms conventional MAP estimators and Wiener filters in noise reduction.

Wiener Filtering (WF)

This is statistically the best way to estimate speech and noise when they are modeled as Gaussian [34]. It works like an "optimal" filter, depending on statistical assumptions and past information to estimate the clean speech from a noisy input. This filter minimizes the mean square error (MSE) of the magnitude spectrum between the estimated speech signal and the clean signal. Its main goal is to reduce the difference between the filtered output and the actual clean signal as much as possible to reach the optimum value. The main limitation of the Wiener filter is that it needs an accurate estimates of both the clean signal and the noise spectra, which is not always easy in practice especially when the noise is non-stationary or rapidly changing [35]. Over the years, different modifications of this filter are used for enhancement task [6]. Anil Garg in [36] proposed a speech enhancement model based on a Long Short-Term Memory (LSTM) network to intelligently set the tuning factor in an adaptive Wiener filter according to signal characteristics. The researcher used a combination of Non-Negative Matrix Factorization (NMF), Empirical Mean Decomposition (EMD), and Bark frequency features to train the system, realizing larger performance

compared to traditional models in terms of PESQ and SNR across several noise environments. Rahul Kumar Jaiswal et al. in [37] introduce an implicit WF that adaptively estimates noise using a recursive equation, enabling strong single-channel speech enhancement in both stationary and non-stationary noise conditions. The algorithm is tested on different noises like drones, helicopters, and airplanes, and it is implemented on Raspberry Pi4. This method surpasses the spectral approach in the intelligibility and nature of the speech.

C. Signal Subspace algorithms (SSA)

The signal subspace approach (SSA) provides dimensionality reduction and a better trade-off between residual noise and the signal distortion of the output signal compared to other existing techniques with the proper tuning of parameters like window size, matrix rank, etc. These algorithms were all founded on the opinion that the vector space of the noisy signal can be decomposed into “signal” and “noise” subspaces. The decomposition can be achieved using either the singular value decomposition (SVD) of a Toeplitz-structured data matrix or the eigen-decomposition of a covariance matrix [4].

Chengli Suna and Junsheng Mu proposed in [38] a subspace-based speech enhancement algorithm that relies on eigenvalue filtering. The method first performs generalized eigenvalue decomposition (GEVD) of the covariance matrices for clean speech and noise, and then removes the less significant components with negative eigenvalues, as these primarily represent noise. Because the filtering process makes the eigenvector matrix irreversible, the generalized inverse is used to reconstruct the speech signal. The experimental results show that the proposed method outperforms conventional approaches, particularly under high-noise conditions, in terms of minimizing speech distortion and reducing residual noise. Chengli SUN et al. in [39] presented an improved subspace speech enhancement method based on Joint Low-Range Scattered Decomposition (JLSMD) to address the fault of traditional methods at low SNR values. The noise signal is represented as a Toeplitz matrix, where the low-range portion represents clean speech and the sparse portion represents noise. The experimental results explain the distinction of the proposed method in preserving speech quality and reducing residual noise compared to traditional methods, especially under high-noise conditions.

2. Methods based on time domain

In these methods, the operations are performed directly on the speech signal. Therefore, the information is observed and extracted from the waveform time samples. There are several methods were proposed to perform SEA in time domain [40]. Some of these well-known methods are listed below:

A. Optimal filtering

Generally, to find the estimated signal from the noisy signal that is degraded by additive noise, a filtering process is applied that performs the suppression operation. It attempts to remove noise from speech in a way that results in the least possible error or distortion [41]. There are several significant filters working in time domain. These filters are mitigating the noise level that is picked up by the microphone signal. For example, the filter called the Minimum Variance Distortionless Response (MVDR) was presented by Capon [42] in this field. It is a linear filter that minimizes the output variance at its output, and simultaneously preserves the distortionless response to a specific input vector [43,44]. Some modification was performed at this filter such as Orthogonal Decomposition Based Minimum Variance Distortionless Response

(ODMVDR) [43]. There are other filters working in time domain such as the Linearly Constrained Minimum Variance (LCMV) [45] and Harmonic Decomposition Linearly Constrained Minimum Variance (HDLCMV) [46]. Each filter has its own limitations [43] that are dependent on the nature of the signal and noise type.

The time domain Wiener filter (WF), which achieves minimum mean square error without spectral transformation, is another example of this domain. The reconstruction of the clean signal can be gained by making a convolution operation between the WF gain and the observed signal [4]. The WF name is based on the mathematician Norbert Wiener; this scientist formulated WF problems and solved them [47]. Another types of optimum filters can be found in [48].

B. Comb filtering

Comb filtering is one of the earliest and simplest methods used to enhance noisy signals in time domain [49]. Its principle is based on the nature of the speech signal, which is a quasi-periodic. Finite Impulse Response (FIR) is used as a basic approach within it, where its coefficients are changed with the fundamental components of the noisy waveform. The input is the noisy signal. So, a decision to determine the voiced from unvoiced signal is made initially. A scaling factor is applied for unvoiced speech (also for silent periods) to suppress the noise, while for voiced signals, the FIR filter function is applied. The final output is the weighted sum of successive points separated by a constant pitch period of the desired signal. Consequently, the sequential periods of voiced signal are added. On the other hand, the unvoiced sound and noise are not. The output of this filter is given [50] by the Equation (5):

$$x(n) = \sum_{k=-L}^L a_k y(n - N_k) \quad (5)$$

where $y(n)$ is the noisy signal, $x(n)$ is the processed signal, and N_k represents the point of comb coefficients. In practice, this filter is not very effective because the ability to remove the noise depends on increasing the coefficient number, but this has a bad impact when the speech signal characteristics are quickly changing. Besides, when fundamental frequency varies, a comb filter impulse response becomes less desirable. Therefore, this filter is considered not competitive currently and other methods are found, such as adaptive filter [49].

C. ANN and HMM Filtering

Artificial neural networks (ANNs) and the Hidden Markov Model (HMM) have hidden states, and they are nonlinear modeling techniques which have a good performance in the speech enhancement (SE) process. HMM was applied firstly in the field of SE in [51] to model speech and noise statistics. The general structure is demonstrated in Figure 3. HMM-based methods model speech and noise statistics using composite probabilistic structures and typically estimate clean speech in the MMSE sense through weighted Wiener filtering.

Different aspects based on HMM models can be found in [52,53]. Although these approaches demonstrate good performance under matched noise conditions, they often suffer from speech distortion and residual noise, particularly in voiced segments, due to inaccurate state estimation in non-stationary environments.

On the other hand, ANN-based speech enhancement methods learn a direct mapping from noisy to clean speech representations through supervised training. The enhanced speech signal is obtained when NN is trained correctly; however, the training process is very time consuming. Besides, the training

is implemented for a specific noise type and SNR level, so it cannot be used for another type of noise. This limitation arises from the noise-dependent nature of training, which restricts their generalization capability. Consequently, ANN-based approaches require retraining or prior knowledge of noise characteristics, making them less suitable for real-time speech enhancement systems. For instance, NN in [54] was trained only for white noise and specific SNR. The integration of ANN into cochlear implant (CI) was also performed [55]. This system decomposes the degradation signal, and then extracts a set of features for feeding the NN to make the estimation. However, a minor decrease in performance was noticed and the work is implemented for CI vocoder simulation not for CI users. Basically, statistical learning-based SEAs achieve a good performance when they are working only in same noise condition as the training set [56].

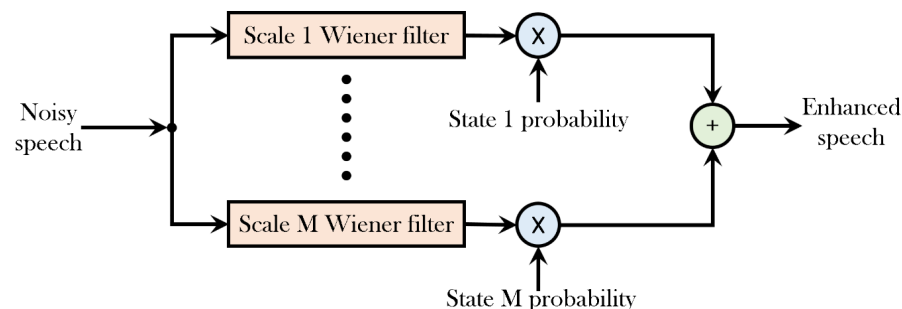


Figure 3. Schematic diagram of HMM-based MMSE estimator.

2.2.2. Machine Learning Speech Enhancement Techniques

Machine learning is generally defined as the ability of a system to improve its performance on a given task based on experience acquired from data [57]. In the situation of single-channel speech enhancement, machine learning techniques are used to learn the statistical properties of speech and noise from data in order to improve speech quality using only a single audio signal. Machine learning-based speech enhancement methods can indeed be categorized into traditional machine learning and deep learning approaches, primarily differentiated by their model complexity, data requirements, and performance in complex noise environments. This categorization is presented in detail as follows.

1. Traditional Machine Learning Approach

Traditional machine learning approaches to single-channel speech enhancement refer to methods that depend shallow machine learning models, where the feature extraction and speech enhancement phases are usually separated, relying on specific mathematical or statistical assumptions for modeling speech and noise, and with a limited number of parameters compared to deep learning methods. Depending on the implicit modeling strategy, most well-known machine learning approaches that are used for speech enhancement are: SVM-based, GMM-based [58], and NMF-based methods, which are discussed below.

A. Support vector machine (SVM)

SVM is a supervised machine learning algorithm used for classification and regression tasks. It attempts to find the best boundary, known as the hyperplane, that separates different classes in the data. It is useful when you want to perform binary classification that has been widely applied in various signal processing and speech-related applications. It is a famous machine learning technology that has grown in popularity due to its ability to address challenging classification problems in a wide range of applications. SVM models are often used in text

classification problems because their capacities to handle enormous quantities of features. It is a successful technique for data classification derived from statistical learning theory. The main idea of SVM in speech enhancement is to implicitly map the input features into a higher dimensional feature space by using a kernel function, where it is easier to classify with linear decision surfaces. Note that the SVM is also can be extended to multi-class classification problems using strategies such as one-vs-one [59] or one-vs-all. Therefore, SVMs have been verified to be successful classifiers on several classical speech recognition problems. In the SEA field, different works have used SVM as an important tool. For example, Yuto Kinoshita et al. [60] depended on the SVM algorithm to discriminate between the natural voice of the passengers and the sound of the advertisement produced from the speakers, using two spectral features: spectral centroid and spectral roll-off. The SVM creates a resolution boundary that separates the two categories in a binary manner, allowing the system to keep only the advertisement and remove unwanted sounds. The results showed that the classifier realized a high accuracy of 96%, demonstrating the efficiency of the SVM in separating useful speech in a noisy environment. Another method proposed by Ying-YeeKong et al. [61] used the SVM algorithm to enhance speech signals with high accuracy based on spectral features such as MFCC, gammatone, and spectral moments, within a very short time window of 8. After training the SVM on clean and noise-contaminated samples, the model classifies each segment into its correct type, enabling improved speech quality within hearing aids through dedicated processing for each sound category. The success of this classification process with such accuracy makes the SVM a crucial step in improving speech clarity for the hearing impaired.

B. Gaussian mixture model (GMM)

The GMM model supposes that any signal (such as noise or speech) does not follow a single, simple distribution, but rather consists of several Gaussian distributions, each representing a different part of the signal. Each Gaussian represents a specific spectral pattern in the noise or speech. By combining several Gaussian distributions, the model can accurately describe complex spectral patterns. A GMM for a mixture signal can be composed of K Gaussians [62], and is given by Equation (6):

$$\Gamma_m(f) = \sum_{k=L}^K C_{mk} G_{mk}(f; \mu_{mk}, \Sigma_{mk}) \quad (6)$$

where $\Gamma_m(\cdot)$ denotes to the probability density function (PDF) of the mixture signal when both noise and speech signals exist, C_{mk} is the contribution of the GMM component in the mixture recording, and G_{mk} is the Gaussian mixture of the mixture recording for the k^{th} GMM component and f^{th} frequency. Also, μ_{mk} and Σ_{mk} denote the mean and covariance matrix of the mixture for the k^{th} GMM component and f^{th} frequency, respectively.

Wageesha Manamperi et al. [62] presented a new single-channel speech enhancement approach using a GMM depending on multi-stage parametric Wiener filtering method at low SNR conditions. The proposed method simplifies dynamic noise adaptation and improves the speech quality while retaining the speech's intelligibility in a non-stationary noise environment with a simple multi-stage parametric Wiener filter. On the other hand, Soundarya M. et al. [63] presented GMM as the core model to improve speech accuracy by representing the correct distribution of voice characteristics in a flexible, probabilistic

way. The GMM algorithm models the difference of the audio signal such as pitch, speed, and accent by combining several Gaussian distributions, allowing speech sounds to be distinguished accurately even in the presence of noise. Basically, the use of GMM contributes to improving the quality of the input signal to the ASR system and makes the speech process more accurate and reliable compared to traditional methods.

C. Non-Negative Matrix Factorization (NMF)

NMF is a dimensionality reduction technique that contains a non-negative data matrix (a matrix with non-negative entries) as a product of two non-negative matrices of lower rank than the initial data matrix. This approach relies only on the amplitude information, while completely discarding the phases of the short-time Fourier transforms (STFTs) [64]. The concept is simply that any signal (or matrix) can be represented as a sum of simple, non-negative components. In the case of sound, the signal does not contain negative energy or intensity values; therefore, this type of analysis is realistic and coherent with the nature of the audio signal. NMF algorithms are suggested by Lee and Seung [65]. Given a non-negative matrix, find a non-negative matrix factor and that given in Equation (7):

$$V \approx WH \quad (7)$$

NMF can be applied to the statistical analysis of multivariable data in the following manner. Given a set of multivariable n -dimensional data vectors, the vectors are placed in the columns of an $n \times m$ matrix V where m is the number of examples in the dataset. This matrix is then approximately factorized into an $n \times r$ matrix W and an $r \times m$ matrix H . Usually, r is chosen to be smaller than n or m , so that W and H are smaller than the original matrix V . This results in a compressed version of the original data matrix doing the decomposition by minimizing the reconstruction error between the observation matrix and the model while limiting the matrices to be non-negative as inputs. This approach has been used in numerous unsupervised learning tasks and also in the analysis of music signals, where the non-negativity constraints alone are adequate for the separation of sound sources [65]. This approach has been used in numerous unsupervised learning tasks and also in the analysis of music signals, where the non-negativity constraints alone are adequate for the separation of sound sources [66].

2. Deep Learning Approaches

Recent advances in deep learning methods have provided significant support for progress in the SE research field. With the complexity of acoustic environments increasing and the increase in non-stationary noise, traditional methods have become less generalizable, and require precise feature design and manual parameter adjustment. This has led to the appearance of deep learning techniques, based on multi-layered neural networks, that are able to automatically learn speech representations from data without the need for manual feature extraction [21].

Advances in computing power include the availability of graphics processing units (GPUs), which enable the efficient training of large-scale neural networks through parallel computation. Additionally, the accessibility of large datasets has sped up the acceptance of deep learning in speech enhancement. Consequently, deep models such as DNN, CNN, GAN, and VAE have proven to attain a better performance compared to traditional methods, particularly in complex, noisy environments that are invisible during training, making deep learning the main trend in modern speech enhancement research [67]. Deep learning approaches can be classified into three classes according

to the learning strategy adopted during training, supervised, unsupervised and semi-supervised [21], which will be discussed below.

- A. **Supervised Single-Channel Speech Enhancement Algorithms (SSC-SEA)**
 Supervised Single-Channel Speech Enhancement Algorithms (SSC-SEA) build separate models for speech and noise signals using labeled data and then bring together these models during the enhancement process. The parameters of the models are gained via training based on the signal samples (speech and noise) and the interfacing model is defined by combining the individual models. Because of this, supervised methods depend on the prior supervision and classification of the speech and noise signals [6,22,58]. Supervised methods overtake unsupervised ones under matched conditions and adequate paired data. They include a wide variety of types, such as DNN, GMM, and SVM.

Deep Neural Network (DNN)

DNN is an artificial neural network (ANN) containing multiple hidden layers between the input and output layers. DNNs can build complex models for nonlinear processing, their architecture is artificial, and the input data is represented as a multi-layered structure. The additional layers assist with the configuration of the features from the lower layers and certainly can model complex input data containing few units as compared to shallow networks; they need a large dataset. In the case of small datasets, the underlying DNNs are unlikely to perform better than other competing methods. The appearance of DNNs has signaled a turning point in this field. By learning complex nonlinear matching between noisy and clean speech from huge datasets, DNN-based models explained maximum capabilities in preserving speech quality and intelligibility. This evolution from statistical modeling to data-driven deep learning approaches has enabled the identification of modern speech enhancement techniques and powerful performance in challenging acoustic environments [10]. Some of the work in recent years has been reviewed as follows: Binh Thien Nguyen et al. in [68] proposed new algorithm to improve enhanced speech by directly estimating phase using DNN, rather than depending on a clean amplitude spectrum as traditional methods do. The system is designed using a real-time convolutional recurrent network (CRN), where the model learns the phase–amplitude relationships from the contaminated signal only. The results showed that the proposed method outperformed traditional phase reconstruction algorithms in terms of the speech quality measure (PESQ) and intelligibility measure (STOI) across various signal-to-noise ratios. On the other hand, the residual noise generated is analyzed by DNN-based SEA by Yuxuan Ke et al. [69]. It has been detected that this noise is time inconsistent, auditorily annoying, and cannot be efficiently reduced using traditional post-processing methods. Three post-processing strategies were proposed based on estimating the noise power spectrum density using the MMSE criterion, while at the same time improving speech probability estimation. Both objective and subjective results proved the efficiency of the proposed methods in reducing residual noise and improving the perceptual quality of speech.

Generative Adversarial Network (GAN)

Generative adversarial networks (GANs) have indeed received significant attention and demonstrated promising results in the field of speech enhancement. GANs consist of a generator network and a discriminator network. In other words, competitive generative networks are generative models that train two DNN models simultaneously: a generator that captures the distribution of the

training data, and a discriminator that estimates the probability that the sample came from the training data and not from the generator [70]. GAN training is generally built on convolutional layers or fully connected layers. Speech enhancement based on GAN training was first introduced by Pascual et al. [71]. The generator network learns to map features of the noisy speech into the clean speech, while the discriminator network acts as a binary classifier that evaluates whether the samples come from the clean speech (real) or the enhanced speech (fake). In general, the two networks are trained in an adversarial manner [67]. Sanberk Serb et al. [72] proposed a two-stage system for real-time speech enhancement that combines a predictive filter (DeepFilterNet2) for noise removal and a GAN network to regenerate lost audio details. The experiments show that the system improves over the first stage predictive model in terms of speech quality. Shrishti Saha Shetu et al. [73] proposed DisCoGAN, which is a time–frequency domain generative adversarial network conditioned by the hidden features of a discriminative model pre-trained for speech enhancement in low SNR scenarios. DisCoGAN outperforms other methods in adverse SNR conditions and can also generalize well in high SNR scenarios.

Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) are widely used in speech enhancement because they can capably extract local patterns in time–frequency representations such as spectral diagrams. CNNs depend on convolutional kernels and weight sharing [74], which reduces computational complexity and increases the model’s robustness against non-stationary noise. Gautam S. Bhat et al. [75] proposed a speech enhancement algorithm based on CNNs and a multi-objective learning approach, where clean speech spectrum estimation is enhanced using additional features such as Mel-Frequency Spectral Coefficients (MFSCs). The proposed method possesses high computational efficiency, allowing real-time implementation on smartphones with low latency. Objective and subjective tests have proved its effectiveness and practical applicability in noisy environments and hearing-impaired support applications. Ashutosh Pandey¹ and DeLiang Wang [76] presented the temporal convolutional neural network (TCNN) mode, which depends on an encoder–decoder architecture with a time convolution module to report time dependencies using causal and extended convolutions. The encoder extracts a low-dimensional representation of the distorted signal, while the decoder employs this representation to reconstruct the time domain-enhanced speech. This pure convolutional architecture enables real-time improvement with low latency and fewer coefficients compared to recurrent models.

- B. **Unsupervised Single-Channel Speech Enhancement Algorithms (USSC-SEA)**
Unsupervised speech enhancement methods have been lately developed to improve the performance of models in SE field. At this point, the models do not use matching clean–noisy data for training. Instead, they use either not-matching noisy–clean data (i.e., the clean and noisy samples do not correspond), or clean data only, or noisy data only. In other words, enabling speech enhancement without the need for paired clean and noisy data. USSC-SEA can be further divided into noise agnostic (NA) and noise dependent (ND) methods. NA methods use only clean speech signals for training. In contrast, ND methods use noise or noisy speech training samples to learn some noise characteristics [77].

Variational Autoencoder (VAE)-based speech enhancement is the clearest and most widely known example of unsupervised deep learning for single-channel speech enhancement [78]. This algorithm depends on a deep variational dynamic autoencoder (DVAE) model pre-trained on pure speech signals to learn the dynamic distribution of speech over time. During the improvement process, this model is used as speech pre-information to distinguish speech from noise in a single-channel signal without the need for noise training data. Noise is modeled adaptively using non-negative matrix analysis (NMF), and speech is separated from noise by a variation expectation–maximization (VEM) algorithm that alternately updates speech and noise estimates. Though the DVAE-based approach is trained in an unsupervised and noise-agnostic manner using only clean speech signals, it becomes noise dependent during the speech enhancement phase by clearly modeling noise using non-negative matrix factorization analysis (NMF). This framework enables the combination of unsupervised speech structure learning and temporal modeling, resulting in higher performance, especially in cases of noise not visible during training.

C. Semi-supervised Single-Channel Speech Enhancement Algorithms

Semi-supervised learning is a combination of two classical learning algorithms: supervised and unsupervised. This approach uses labeled and unlabeled data to improve the quality of speech signals from a single channel, combining aspects of supervised and unsupervised machine learning [79]. The goal is to overcome the limitations of fully supervised algorithms by using large amounts of unlabeled data to create robust models that better generalize to invisible noise, in line with real-life situations. Traditional algorithms, and especially non-stationary noise reduction methods, do not adequately suppress transient interference that distorts speech signal. Therefore, a semi-supervised single-channel noise suppression has been appeared to reduce noise with minimal audio distortion [80].

Adaptation learning algorithms are a main example of semi-supervised speech enhancement; they involve three stages of training. In the first stage, a speech enhancement model is trained using a limited amount of manually labeled target data. Then, the initially trained model is used to enhance unlabeled data and generate estimated labels for clean speech. In the final stage, a high-quality subset of this enhanced data is selected to retrain the model, thus improving its performance and adaptability to the target testing environment [81]. Zihao Cui et al. [79] provided an estimator to measure speech purity. For clean speech utterances, a supervised ML is used to train a DL speech enhancement model. While, for noised utterances, the model is updated based on unsupervised manner. The training loss is a combination of the supervised loss and unsupervised loss that provides a semi-supervised SEA. Other works have used a method of semi-supervised learning, Yoshiaki Bando et al. [82] proposed a semi-supervised method for speech enhancement called Variational Autoencoder (VAE-NMF), where speech is shown using a pre-trained deep generative model of VAE on clean speech signals, while noise is modeled using NMF during processing. The speech model outperforms traditional supervised methods, especially in cases of invisible noise. It uses a probabilistic priority to estimate clean speech from the noise signal by sampling the subsequent distribution, while the noise model is adapted to unknown environments.

Although deep learning models vary in their architecture (such as CNN and DNN), their ability to generalize and perform efficiently with noise not present during training

depends less on the complexity of the architecture itself and more on how the model learns speech representations from the data. Models trained on different datasets and under noise conditions tend to learn consistent speech features in the presence of noise, thus improving their robustness in unusual environments. Furthermore, using techniques, such as regularization to prevent overfitting, semi-supervised or unsupervised learning and self-supervised learning, helps the model learn deeper and more general representations, thereby improving its ability to generalize.

For this reason, we notice that some deep learning models perform better than others when there is a difference between training and test noise, even if the architectural differences between them are minor.

2.3. Multi-Channel Speech Enhancement Algorithms

In multi-channel SEAs, audio signals are captured from different angles and distances. Two or more microphones are used to take advantage of spatial cues, such as timing, phase, and amplitude differences, to separate the target speech from background noise. Compared to single-channel approaches, multi-channel methods offer improved noise suppression and spatial selectivity. Multi-channel SE is an emerging field of research that assists in the improvement of different systems based on speech signals for various real-life applications. They are now popular among many listening devices (e.g., hearing aids or mobile phones) and allow us to use spatial diversity in addition to spectro-temporal diversity. Multi-channel algorithms appear to be crucial for current and future Assisted Listening Devices (ALDs). In contrast to single microphone signal enhancement algorithms, which have not been shown to provide high improve in speech intelligibility, multi-microphone signal enhancement algorithms are capable of increasing speech intelligibility, especially when the sound sources have different spatial characteristics [83]. The enhancement of speech signal is used by automatic speech recognition as a pre-step in different speech areas such as autonomous machines, AI, voice recognition, and smart homes. There is great interest in this field; therefore, many previous articles have appeared for multi-channel SE. The existing surveys inadequately include the general types of this approach, but it can be classified according to [84] into four major techniques including Blind Source Separation (BSS), beamforming, Multi-Channel Non-Negative Matrix Factorization (NMF), and deep learning-based multi-channel methods that are discussed as follows. The main difference between single-channel and multi-channel methods is shown in Figure 4.

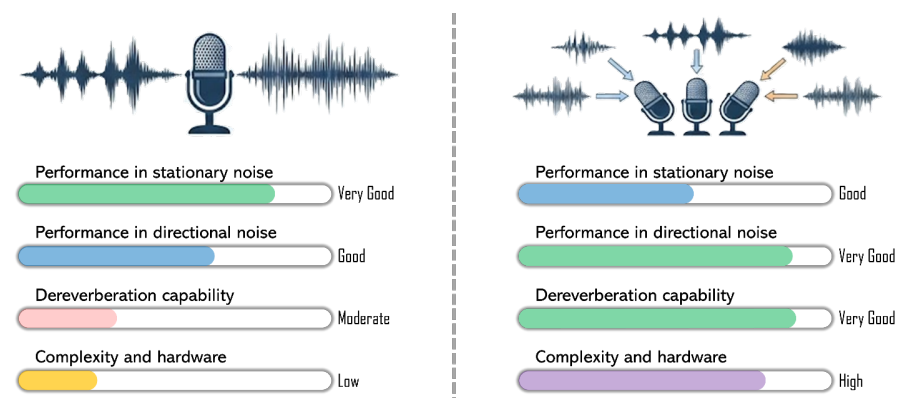


Figure 4. The main difference between single channel and multi-channel.

2.3.1. Blind Source Separation (BSS)

Blind Source Separation (BSS) techniques perform noise reduction by extracting individual source signals from a mixture without needing detailed previous knowledge about the sources, the main objective being to restore the original signals by using the

statistical properties of the mixed signals. Key techniques within BSS include Independent Component Analysis (ICA) and Time–Frequency (T-F) masking. This model for speech enhancement is built by assuming the reverberations generated by later reflections as additional and irrelevant noise components. Later, a blind signal extraction approach is designed to extract the direct sound and early reflected sounds, realizing reduced reverberation and noising [85].

2.3.2. Beamforming

Beamforming technology directs a virtual beam towards the target speaker while placing “dead zones” in the direction of noise sources. Basically, transformer arrays combined with signal processing algorithms are called beamformers when they transmit or receive energy in a focused direction. Beamforming algorithms can be divided into three main categories: data independence, statistically optimized, and adaptive [86]. Classical methods are mainly focused on using beamforming to bring together multiple signals after filtering for further noise reduction. The beamforming methods design a linear filter per frequency to improve or keep the signal from the target direction, while mitigating the interferences from other directions [25]. Beamforming achieves directional signal transmission from several sensors using signal processing while suppressing interference from other directions. In the concept of SE, it can be divided into two basic types: fixed and adaptive beamforming. where fixed beamforming uses a stationary noise source and the direction of speech. And adaptive beamforming includes the directivity of the speech signal differing with the change in audio environments [84].

2.3.3. Multi-Channel Non-Negative Matrix Factorization (NMF)

The Multi-Channel Non-negative Matrix Factoring (MNMF) algorithm is designed to handle multi-channel audio signals and is considered an extension of the traditional NMF algorithm. It uses the spectral and spatial information available when using multiple microphones. Unlike single-channel NMF, which depends only on spectral data, MNMF integrates spatial information such as phase differences between microphones, allowing for the differentiation of audio sources based on their spatial positions [87]. In [87] the model enhances the traditional NMF algorithm to operate in a multi-channel environment by incorporating spectral and spatial information into a unified probabilistic model, using semi-defined positive Hermitian matrices to show the spatial characteristics that are used to identify the phase differences between microphones. The model parameters are estimated via iterative updates, and each audio source is reconstructed based on its spectral and spatial contributions, achieving the efficient separation of multi-channel sources.

2.3.4. Deep Learning-Based Multi-Channel Methods

Deep learning-based multi-channel methods join the power of different neural networks (for example DNNs, CNNs, RNNs, and Transformers) with spatial processing techniques from classical signal processing. The main advantage is that these methods can jointly learn both spatial and spectral features directly from signals while traditional approaches cannot. In other words, instead of hand-designing mathematical filters, the neural network learns how to extract speech and reduce noise automatically, using thousands of noisy and clean training examples. The main challenges of these methods are that they require large datasets and have high computational cost for training and real-time processing. Additionally, they may generalize badly to unseen noise types if not trained correctly [88,89].

To provide a clearer and more intuitive comparison, Table 1 summarizes the main types of speech enhancement methods in terms of their performance trade-offs, computational requirements, and suitability for real-time applications. This comparison is planned to help readers quickly understand the practical strengths and limitations of each approach.

Table 1. Comparison of speech enhancement techniques.

Method Type	Performance Trade-Offs	Computational Complexity (Low, Medium, and High)	Real-Time Applicability	Strengths	Limitations
Spectral Subtraction [27] (Single Channel)	Speech quality may be affected if noise estimation is inaccurate.	Low	Available	Simple, fast, and no training required; effective with stationary noise.	Musical noise artifacts and poor performance in non-stationary noise.
STSA-MMSE [29] (Single Channel)	Improves speech quality with reduced musical noise.	Medium	Available	Suitable for stationary and quasi-stationary noise.	Limited performance in highly non-stationary noise; relies on statistical assumptions.
Wiener Filtering (WF) [34] (Single Channel)	Balances noise reduction while maintaining speech clarity.	Medium	Available	Clear and understandable statistical framework.	Depends on accurate SNR estimation; ineffective in highly non-stationary noise and complex environments.
Signal Subspace Algorithms (SSA) [38] (Single Channel)	Good speech–noise separation when subspaces are well estimated.	Medium–High	Limited	Reduces distortion; suitable for signals with clear structure.	High computational cost and poor robustness to non-stationary noise.
Traditional Machine Learning (SVM [59], GMM [62], NMF [64]) (Single Channel)	Better than spectral methods but weaker than deep learning in complex scenarios; highly dependent on feature quality.	Medium	Available	Works well with limited data; effective for noise classification and structured noise.	Requires handcrafted features; weaker performance in complex and unseen noise compared to deep learning.
Deep Learning Approaches (Supervised [6], Semi-Supervised [79], Unsupervised [77]) (Single Channel)	Best performance in audio quality and intelligibility, but highly dependent on training strategy and data diversity.	High	Limited	Strong performance in non-stationary noise; no manual feature design.	Requires large datasets; limited interpretability; and deployment challenges in real-time and low-resource systems.
Blind Source Separation [85] (Multi-Channel)	Highly effective when multiple microphones are available and sources are spatially independent.	Medium–High	Limited	Exploits spatial information; effective for multi-source scenarios.	Sensitive to reverberation; limited robustness in complex environments.
Beamforming [86] (Multi-Channel)	Strong performance when the speech source direction is known or estimable.	Medium	Available	Exploits spatial information; improves SNR; suitable as a pre-processing stage.	Limited performance in highly reverberant or spatially overlapping noise.

The comparison shows that traditional methods are good for real-time applications due to their low computational cost, but they often fail in complex and quickly changing noise conditions. In contrast, deep learning approaches provide significantly better enhancement performance, although this comes with deployment challenges and higher computational demands. Traditional machine learning and multi-channel methods fall between these extremes, providing useful performance gains under specific conditions while introducing their own practical constraints.

The effectiveness of speech enhancement algorithms varies significantly across different noise conditions. Classical methods perform reasonably well under stationary noise but degrade rapidly in non-stationary environments. Deep learning approaches demonstrate stronger robustness to complex noise; however, their performance heavily depends on the diversity of training data and noise mismatch.

3. Noise: Definition and Classification

Noise refers to undesired acoustic interference that contaminates speech signals during recording or transmission [5] and remains a primary factor limiting the performance of speech enhancement algorithms. Its characteristics, including temporal variability and spectral distribution, directly influence noise estimation accuracy and overall enhancement effectiveness. In general, most SEA require two major procedures: the estimation of the noise power spectrum and the estimation of the speech spectrum [90]. The problem occurs when interfering noise signals strongly degrade the desired speech signals. Generally, people are surrounded by noise wherever they go, for example, cars, streets, restaurants, offices, and shopping malls, which lead to the corruption of a person's speech. Additive noise signals come in different shapes and forms and are sometimes called background noise, a more commonly used term in research and in practice, which has led to more effective results with existing algorithms [4,22].

Different noise conditions impose distinct challenges on speech enhancement algorithms. Classical statistical methods such as Wiener filtering and STSA-MMSE are effective under stationary noise assumptions but exhibit significant performance degradation in highly non-stationary environments such as babble and traffic noise. In contrast, deep learning-based models demonstrate superior robustness in complex noise conditions due to their ability to learn nonlinear spectral mappings, although at the cost of increased computational complexity and limited generalization to unseen noise types. These distortions together degrade both speech intelligibility and quality, especially when the signal-to-noise ratio (SNR) is low. Besides, many speech processing tasks such as automatic speech recognition (ASR) and speaker identification (SID) become more difficult under these preserve noisy conditions [91]. Basically, as the distance between the source and the receiver of speech increases, speech signals become increasingly corrupted by surrounding noises, interferences, and echoes, and exposed to distortions introduced by communication media. This condition will lead to decreased speech intelligibility and reduces the effectiveness of communication quality. Overall, the impact of noise on speech enhancement performance is strongly dependent on its temporal behavior and acoustic structure. Algorithms optimized for stationary noise may fail under realistic, non-stationary conditions, while learning-based models offer improved adaptability at the cost of increased complexity and potential generalization issues. These trade-offs highlight the importance of selecting enhancement strategies that are aligned with the target noise environment and application constraints, particularly in real-time and resource-limited systems.

The general types of noise that affect speech signals are shown in Figure 5.

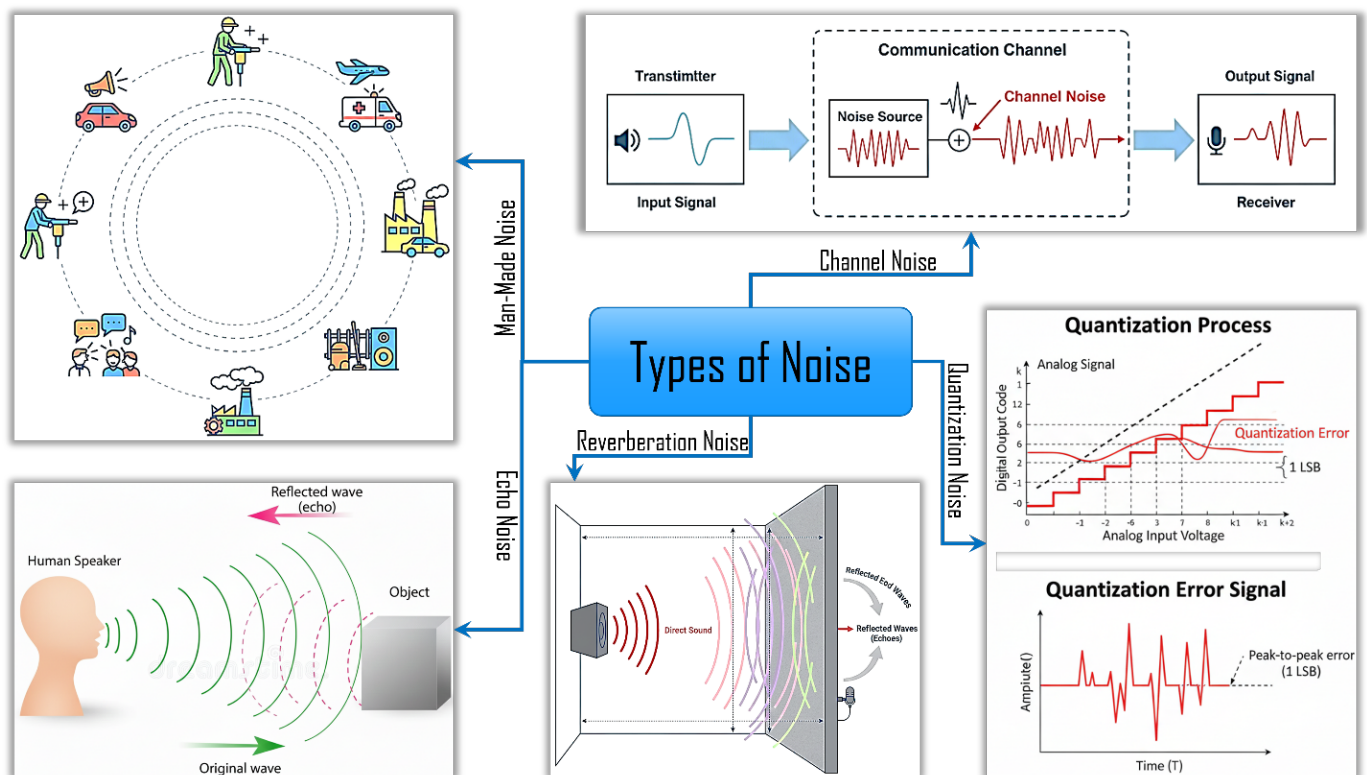


Figure 5. The basic types of noise in our environment.

Speech enhancement has numerous objectives, including increasing intelligibility, improving quality, solving the noise pollution problem, increasing speech recognition accuracy, and reducing listener fatigue. However, its key objective is to regain clean speech signals from noisy signals degraded by additive background noise. Recently, SEAs were considered substantial issues that have significant importance in our daily lives, because they are applied in different key applications.

4. Different Types of Datasets That Are Used in SEAs

In the field of speech enhancement, datasets refer to sets of human's speech recordings (either clean or corrupted by noise) used to train, test, and evaluate the algorithms that are designed to improve speech quality or/and intelligibility. They model the base of the deepest learning and traditional filtering algorithms, allowing fair comparisons of research results. Reliable datasets are essential to evaluate the performance of SEA, since they provide results with scale, robustness, and confidence. They are a fundamental tool in speech analytics that provide the data from which analysts extract significant information, and they can be used to conduct and identify trends in SEA. Speech under clean and noisy conditions is taken from various well-known databases to build a strong foundation of research.

The goal of releasing clean speech and noise datasets is to provide researchers with broad and representative datasets to train or evaluate their speech enhancement models. Earlier, the Microsoft Scalable Noisy Speech Dataset (MSS-NSD) was released with a focus on expansion. It is a context for noisy speech generation that takes in directories pointing to clean speech and noise as inputs. Other parameters that can be dominated in the dataset are the desired number of noisy speech clips, length of the clip, and SNR levels [92].

When researching speech enhancement, datasets can be divided into numerous main categories based on their purpose and how noise is present in them. Firstly EARS [93] and TIMIT [94], which are clean speech datasets that do not contain noise and high-quality speech recordings, and include multiple speakers and accents. Secondly, there are NOIZEUS [4] and

DEMAND [95], which are noisy speech datasets (additive noise) that contain speech signals mixed with background or environmental noise at different SNR level. Thirdly, CHiME-4/6 [96,97] and the DNS Challenge Dataset [98] are real-world noisy speech datasets that are used for training and testing models on realistic non-stationary noise. Fourth, Audio Set [99], which is weakly labeled, is used with unsupervised and semi-supervised learning algorithms. Fifth, the DDS Dataset [100] and AURORA-4 [101] are device-degraded speech database that contain distortions from microphones, smartphones, etc. Sixth, the REVERB Challenge Dataset [102] contains reverberant or convolutive datasets that are used for studying dereverberation. Lastly, WHAMR! [103], is synthetically generated, paired, single-channel dataset which is a mixed-purpose dataset (hybrid) that integrates clean, noisy, and reverberant conditions with multiple channels and labels. Table 2 summarizes the basic datasets that can be used in speech enhancement and their characteristics.

Table 2. Summary of Commonly Used Speech and Audio Datasets.

Dataset Name	Type of Data	Description	Reference	Purpose
EARS	Clean Speech	Contains approximately 100 h of clean speech recordings at a sampling rate of 48 kHz from 107 speakers with diverse backgrounds.	[93]	Speech enhancement and dereverberation
TIMIT	Clean Speech	Includes 630 speakers across eight different American English accents. Each speaker records ten sentences using a high-quality microphone in a quiet room. The sampling rate is 16 kHz.	[94,104]	Speech recognition
NOIZEUS	Clean and Noisy Speech	Consists of 30 spoken sentences from six speakers (three male and three female). Clean speech is corrupted with realistic noises such as babble, car, train, street, airport, and restaurant at SNR levels of 0, 5, 10, and 15 dB. The sampling rate is 8 kHz.	[4,105]	Evaluation of single-channel speech enhancement algorithms
DEMAND	Real Environmental Noise	A database of 16-channel environmental noise recordings captured from real (non-simulated) environments using multiple microphones. The sampling rate is 48 kHz.	[95,106]	Testing algorithms using real-world noise in diverse environments
CHiME-4/CHiME-6	Real-world Noisy Speech	CHiME-4 includes real and simulated multi-channel recordings based on WSJ0 sentences. CHiME-6 was recorded in real homes using multiple microphone arrays in different rooms.	[96,97,107]	CHiME-4: Evaluation of speech enhancement and multi-microphone ASR systems. CHiME-6: Speech recognition in everyday home environments
DNS Challenge Dataset	Real-world Noise with Simulated Noisy Speech	Contains thousands of hours of clean speech and real environmental noises, including office, street, car, restaurant, factory, storm, and household noises.	[98]	Training and evaluation of real-time deep noise suppression models
AudioSet	Weakly labeled Noisy Audio	A large-scale dataset containing approximately 5000 h of 10 s audio clips collected from YouTube, covering over 500 sound categories.	[99,108]	Training and evaluation of audio event classification systems
DDS Dataset	Noisy–Clean Speech	Includes studio-quality clean speech and corresponding distorted versions, totaling 1944 h of recordings collected under 27 realistic conditions using three microphones and nine acoustic environments.	[100]	Training and evaluation of speech enhancement models and adaptation for ASR systems

Table 2. *Cont.*

Dataset Name	Type of Data	Description	Reference	Purpose
AURORA-4	Noisy Speech	Contains 16 kHz speech data derived from the WSJ0 corpus with additive noises at SNRs between 10 and 20 dB. The training set includes 7138 utterances from 83 speakers.	[101,109]	Evaluation of speech enhancement and speech recognition systems
REVERB Challenge Dataset	Reverberant Noisy Speech	A database of speech affected by reverberation and noise, including real recordings and simulated far-field microphone data at 16 kHz.	[102,110]	Evaluation of robust speech recognition algorithms
WHAMR!	Clean, Noisy, and Reverberant Speech	Includes clean speech, two-speaker mixtures, real environmental noise, reverberation, and full mixtures combining speech, noise, and reverberation with random gains.	[103]	Evaluation of speech enhancement, echo cancellation, and speaker separation algorithms

5. Performance Enhancement Strategies for Quality and Intelligibility Assessment Measure for SEA

Speech quality concerns speech clarity, the nature of distortion, and the amount of background noise, while speech intelligibility concerns the percentage of words that can be obviously understood [23]. Good quality does not always guarantee high intelligibility, as they are independent of each other. Therefore, most speech enhancement algorithms improve quality at the count of reducing intelligibility or vice versa [23]. The speedy increase in the usage of speech processing algorithms in multi-media and telecommunications applications raises the need for speech quality evaluation. The speech quality assessment can be performed using subjective listening tests or using objective quality measures. Subjective evaluation includes comparisons of original and processed speech signals by a group of listeners who determine the rate of the quality of speech according to a predefined scale. Objective evaluation, on the other hand, includes a mathematical comparison between the original signal and processed speech signals. Quality is one of many characteristics of the speech signal and intelligibility is a different characteristic; the two are not equivalent. Therefore, different assessment methods are used to evaluate the quality and intelligibility of processed speech [111].

Among various quality and intelligibility speech enhancement measurements, the most important are as follows: the weighted spectral slope (WSS), the Coherence Speech Intelligibility Index (CSII), frequency-weighted segmental SNR (FWSNRseg), the Short-Time Objective Intelligibility measure (STOI), Perceptual Evaluation of Speech Quality (PESQ), and LPC-based objective measures including the log-likelihood ratio (LLR), segmental SNR (segSNR) and the composite measures. These measures are considered the most famous and they have been studied broadly in recent years and have enjoyed wide acceptance [112,113]. The most significant measurements are listed as follows.

5.1. Perceptual Evaluation of Speech Quality (PESQ)

The PESQ is a popular metric that it aims to measure speech quality after passing through the network and codec distortions. In PESQ, the original and degraded signals are assigned to an internal representation using a perceptual model. The difference in this representation is used by a cognitive model to predict the speech quality of the degraded signal. This perceived listening quality is expressed in terms of Mean Opinion Score (MOS), an average quality score over a large set of subjects [114]. The PESQ score is mapped to a MOS-like scale in the range of 0.5 to 4.5. It is an international standard for estimating MOS from both clean and degraded signals. The PESQ was officially standardized by the International Telecommunication Union–Telecommunication Standardization Sector

(ITU-T) as standard measure able to predict subjective quality with good correlation in a very wide range of conditions, including coding distortions, errors, noise, filtering, and delay and variable delay [115]. The structure of the PESQ measurement [114] is shown in Figure 6.

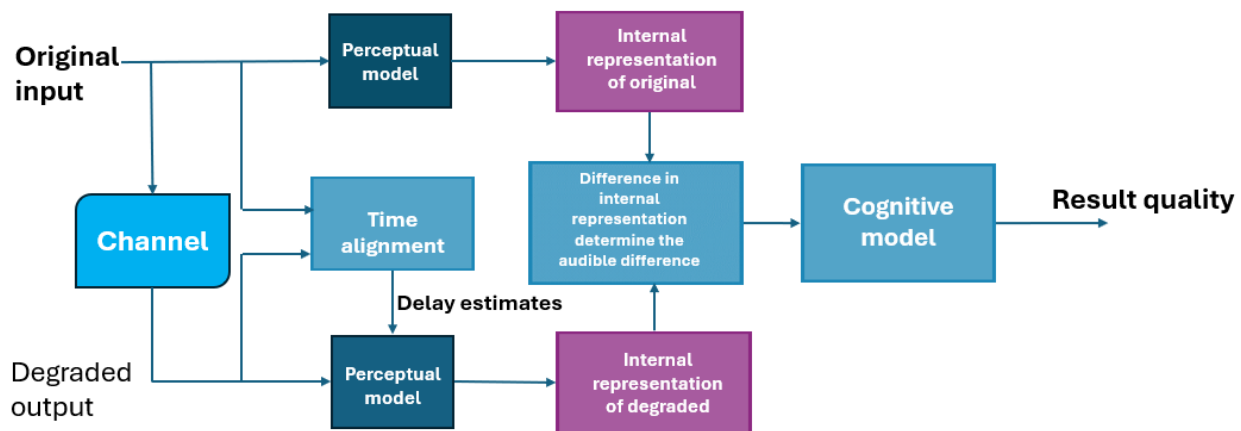


Figure 6. Basic diagram of PESQ.

5.2. Frequency-Weighted Segmental SNR Measures (FWSSNR)

FWSSNR is an objective metric used to assess the quality of enhanced speech in speech enhancement algorithms. This metric computes the segmental signal-to-noise ratio (SNR) over short segments of the signal, but adds an important step by assigning different weights to the acoustic frequencies depending on their auditory importance, which means given higher weights are given to frequencies that affect speech intelligibility. Thus, FWSSNR offers a more accurate estimate of speech quality than traditional metrics such as the overall signal-to-noise ratio. For example, it is used to evaluate the performance of algorithms such as the Wiener filter or deep neural network-based enhancement, measuring the degree to which the quality improves after noise removal in different frequency ranges [112].

The frequency-weighted segmental SNR (FWSSNR) was computed using the following Equation (8):

$$FWSSNR = \frac{10}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^K W(j, m) \log_{10} \left(\frac{|X(j, m)|^2}{|X(j, m) - \hat{X}(j, m)|^2} \right)}{\sum_{j=1}^K W(j, m)} \quad (8)$$

where $W(j, m)$ is the weight placed on the j^{th} frequency band, K is the number of bands, M is the total number of frames in the signal, $|X(j, m)|$ is the filter bank magnitude of the clean signal in the j^{th} frequency band at the m^{th} frame, and $|\hat{X}(j, m)|$ is the filter bank magnitude of the enhanced signal in the same band. This measure is the same of SNRseg, but with more averaging over frequency bands. These bands are proportional to the critical bands of the human ear [116].

Mingchi Hou and Ina Kodrasi [117] examine how the acoustic properties of clean speech affect the ease or difficulty of improving it using modern SE algorithms. The methodology involves extracting acoustic features such as format and pitch and linking them with improved quality metrics such as fwSegSNR and PESQ to detect the most influential features. The results show that sentences with a high and consistent format are improved more efficiently, emphasizing the importance of internal speech properties in evaluating and developing speech improvement models.

5.3. The Coherence Speech Intelligibility Index Measure (CSII)

Speech intelligibility is the main concern in hearing aids and many communication systems. The Speech Intelligibility Index Measure (SII) has been advanced and validated for additive noise and for bandwidth reduction, but the output of hearing aids is affected by distortion also by noise. The coherence between the input and the output of the device provides a tool for calculating the collective effects of noise and distortion in the processed output. Hence, it is better to modify the SII calculation procedure to use the signal-to-distortion ratio (SDR) computed from the coherence instead of the signal-to-noise ratio (SNR), and thus obtain the distortion along with the noise [118]. The powerful step of this measure is dividing the speech signal into three regions: low, mid, and high level. Then, for each region, the coherence SII (CSII) is calculated individually from the signal segments. Subsequently, the intelligibility is estimated of the coherence SII from a weighted combination of the three values as follows in Equation (9):

$$c = -3.47 + 1.84CSII_{Low} + 9.99CSII_{Mid} + 0.0CSII_{High} \quad (9)$$

Then, the weighted sum is transformed using a logistic function to obtain the correct predicted proportion sentences as follows in Equation (10):

$$I_3 = \frac{1}{1 + e^{-c}} \quad (10)$$

where I_3 denotes the intelligibility predicted by the speech intelligibility model. The three level CSII is an accurate procedure for estimating the intelligibility of the distortion conditions and noise. It is effective for normal and hearing-impaired listeners [118]. This measure is one of the important metrics used in the intelligibility measures [4].

Previous studies have shown that the CSII measure is active in detector improvements in speech intelligibility for hearing aid applications, and provides reliable valuation across low, mid, and high speech-level regions.

5.4. The Short-Time Objective Intelligibility Measure (STOI)

(STOI) creates high correlations with the intelligibility of noisy and time–frequency-weighted noisy speech [23]. It depends on the correlation between temporal envelopes of the clean and degraded speech in short-time (382) segments. This is unlike other measures, which usually consider the complete signal at once, or use a very short analysis length (20–30) [119]. Additionally, it calculates the average of the correlations across specific bands and blocks and then uses the average correlation to predict the intelligibility scores, and produces a final value that usually lies between zero and one, where a higher value indicates more intelligible speech. The clipping operation is efficient mainly in regions that contain only noise and thus it is used to reduce the importance or impact of those regions on SI.

Mohamed Medani et al. [120] used the STOI scale to measure how much speech intelligibility improves after signal enhancement, especially when combining enhanced and original features. By the STOI results, the study shows that the proposed model significantly improves intelligibility compared to traditional methods, emphasizing the ability to recover important speech details even in environments with strong noise.

5.5. Weight Spectral Slope (WSS)

The WSS distance measure calculates the weighted difference between the spectral slopes in each frequency band. The spectral slope is the difference between adjacent spectral magnitudes in decibels. The WSS measure evaluated is defined in Equation (11):

$$d_{wss} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^K W(j, m) (S_c(j, m) - S_p(j, m))^2}{\sum_{j=1}^K W(j, m)} \quad (11)$$

where $W(j, m)$ are the weights computed as pair, K is the number of bands, M is the total number of frames in signal, and $S_c(j, m)$ and $S_p(j, m)$ are the spectral slopes for j^{th} frequency band at m frame of the clean and processed speech signals, respectively. In our implementation, the number of bands was set to $K = 25$ [112].

Seorim Hwang et al. [121] used the Weighted Spectral Slope (WSS) metric to evaluate the model's ability to retain the spectral structure during speech enhancement, especially the format positions and the spectral nature of the speech. The results show that the dual-path model reduces the WSS value compared to simpler versions, indicating it is better at retaining the speech's spectral structure and reducing distortion, which interprets enhanced audio quality. Thus, WSS is a valuable indicator in this study for evaluating speech quality after noise removal.

5.6. The Composite Measures

Composite measures are a set of objective quality measures used to evaluate speech quality and the overall effectiveness of speech enhancement algorithms [23]. These measures were obtained by combining basic objective measures to constitute a new measure. Additionally, they are necessary when we cannot expect conventional objective measures (e.g., LLR) to relate highly speech/noise distortions with overall quality. They use both multiple linear regression analysis and MARS analysis to estimate three different composite measures: a composite measure for speech distortion situation (SIG), a composite measure for background noise intrusiveness (BAK), and a composite measure for overall speech quality (OVRL) [112]. Substantively, the BAK, SIG and OVL results are defined in the composite method based on the combination of four objective quality measures (LLR, WSS, SNRseg and PESQ) [122]. While SIG is used for signal distortion and is formed by linearly combining the LLR, PESQ, and WSS measures, BAK is used for noise distortion and is formed by combining SNRseg, PESQ, and WSS measures. And OVL is used for the overall quality (OVRL) formed by linearly combining the PESQ, LLR, and WSS measures. These three measures can separately measure SD (speech distortion), NR, and the overall quality of enhanced speech [123] as shown in Equations (12)–(14):

$$SIG = 3.093 - 1.029 LLR + 0.603 PESQ - 0.009 WSS, \quad (12)$$

$$BAK = 1.634 + 0.478 PESQ - 0.007 WSS + 0.063 SNR_{seg}, \quad (13)$$

$$OVL = 1.594 + 0.805 PESQ - 0.512 LLR - 0.007 WSS. \quad (14)$$

The idea of composite measures is that different objective measures take different characteristics of the reconstructed signal and, therefore, combining them in a linear or nonlinear manner can potentially yield significant gains in correlation.

Tomer Rosenbaum et al. [124] used composite measures, CSIG, CBAK, and COVL metrics, to evaluate the speech quality after a noise reduction by measuring the amount of remaining distortion, the background noise level, and the overall quality after the enhancement. The study depended on comparing these metrics before and after optimization to determine the effectiveness of the model, providing a comprehensive view in addition to the echo removal metrics. This paper emphasized that the composite measures offer an additional, complementary view to the other measurements, especially in cases of mixed noise and reverberation.

5.7. Scale-Invariant Signal-to-Distortion Ratio (SI-SDR)

The SI-SDR is suggested as a more powerful alternative to the traditional SDR metric as it introduces an ideal scaling factor to the target signal, which assures an orthogonal projection of the estimated signal onto the target signal. The SI-SDR function has also been widely used as a loss function for training models. The use of SI-SDR is limited to evaluating phonetic articulation with noisy references. Basically, SI-SDR is limited by the reference's signal-to-noise ratio (SNR) when the group accurately estimates noise-free phonetic articulation. The SI-SDR [125] is defined in Equation (15):

$$SI-SDR(s, \hat{s}) = 10 \log_{10} \left(\frac{\|\alpha s\|^2}{\|\alpha s - \hat{s}\|^2} \right) \quad (15)$$

where α is the optimal scaling factor, s is the reference signal, and $\hat{s}$ is the estimated signal.

Hao Zhang et al. [126] used the SI-SDR metric to evaluate a system's ability to reduce howling noise and improve speech quality without signal distortion. SI-SDR measures show how well the model retrieves the original signal compared to the degraded one. The results show that the presented hybrid model realizes the highest SI-SDR among all methods, indicating that the model is better able to remove noise and retain the signal shape.

5.8. Log-likelihood Ratio (LLR)

The LLR scale measures the degree of spectral distortion that occurs in speech after it has been enhanced. It depends on comparing the Linear Predictive Coding (LPC) coefficients between the original and enhanced speech. When the LLR is low, the enhanced signal is very close to the original, but, if it is high, the enhancement algorithm has caused distortion in the spectral shape [112].

Seorim Hwang et al. [121] used the LLR to measure the amount of spectral distortion between the enhanced and original signals, precisely in the form of spectral slope and energy within the signal. A lower LLR value means that the model better retains the natural acoustic structure, so researchers depended on it to detect cases of loss of fine details. The proposed model has shown a significant improvement in the LLR compared to other models, reflecting its better ability to reconstruct the spectrum without distortion based on this measurement.

Although this review does not provide direct quantitative comparisons, it highlights widely adopted benchmark datasets such as DEMAND, CHiME, and DNS, along with commonly used evaluation metrics including PESQ, STOI, and SI-SDR. These uniform experimental settings form the basis for fair comparison across studies and allow readers to understand reported performance trends in the literature.

6. Key Applications for Speech Enhancement Technologies

Speech enhancement technologies are widely applied in various fields of speech processing as a pre-processing step which aims to improve signal quality and intelligibility by reducing background noise and distortion. In many speech applications such as mobile phones, hand-free communications, hearing aids, speech recognition, audio processing in virtual reality, stereo sound systems, etc., speech enhancement plays an important role in their performance. Over the past years, researchers have been working on speech enhancement to improve signal quality and intelligibility in noisy and reverberant environments. Here, the quality of the speech signal indicates the clarity of the speech signal, whereas the intelligibility indicates the understandability of the words by the listener. SEA are selected depending on application at hand, the type of interference signal, and the various noisy and reverberant cases [127].

A significant issue in speech processing applications is how determined the active speech periods are within a given audio signal. Speech can be marked by a discontinuous signal because information is carried only when someone is talking. The regions where voice information is present are referred to as ‘voice-active’ segments and the stops between talking are called ‘voice-inactive’ or ‘silence’ segments. The decision to determine what class an audio segment belongs is based on an observation vector. Performance trade-offs are made by increasing the detection rate of active speech while reducing the false detection rate of inactive segments. Nevertheless, producing an accurate indication of the presence of speech, or its absence, is commonly difficult, especially when the speech signal is corrupted by highly unstationary background noise or unwanted interference [128]. Speech enhancement applications are growing in tandem with security, automation, medical, and industrial applications, requiring significant attention and development in this field, as can be seen in Figure 7.

Some of the most important of these applications are included in the following sections:

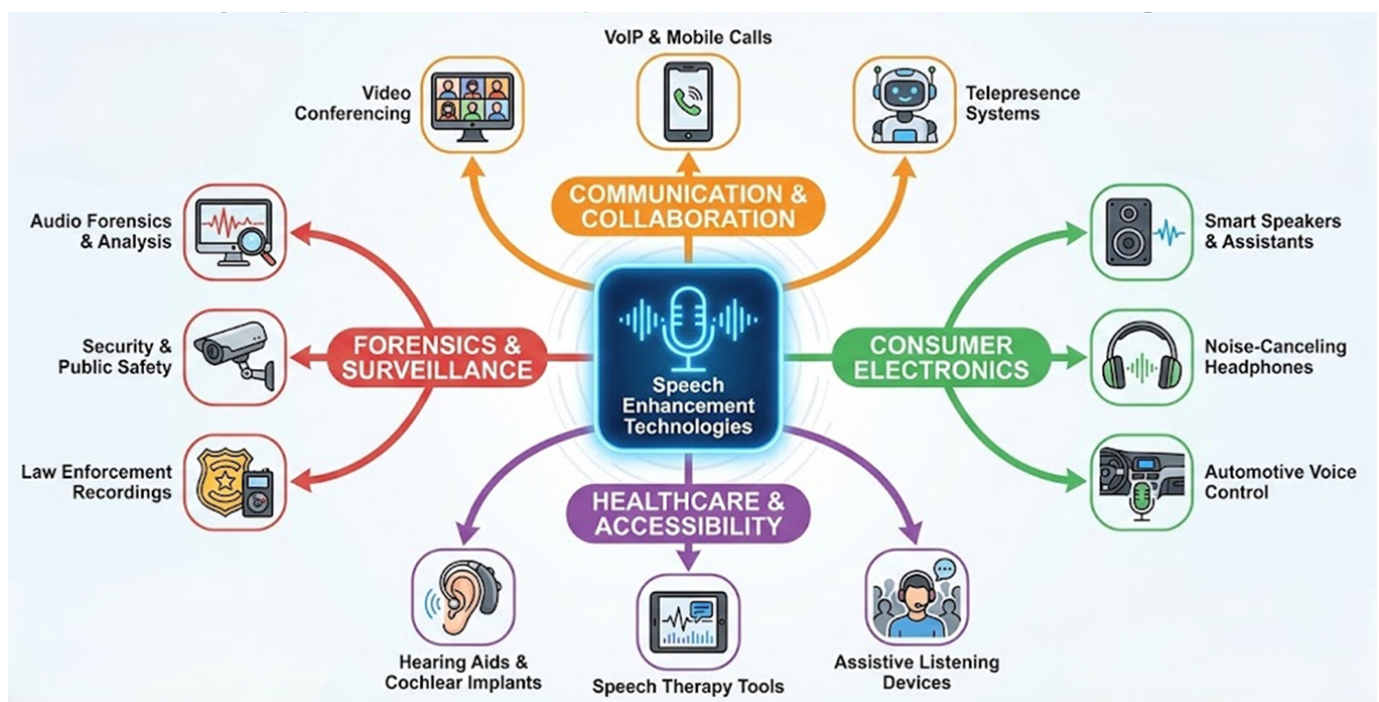


Figure 7. The major applications of speech enhancement technologies.

6.1. Smart Hearing Aids Application

The Smart Hearing Aid is a sophisticated digital hearing aid that goes beyond simply amplifying sound. It integrates modern technologies such as AI, ML, and Bluetooth connectivity to provide a highly personalized, adaptive, and connected good hearing experience. Smart hearing aids have the ability to detect important noise such as fire alarms, car horns, etc., and make them audible. Therefore, it can be used to avoid disastrous incidents that might have happened to the hearing impaired if these noises are not heard. Hearing aids manage to convert a deep neural network-based speech enhancement approach to be aware of the desired noise types and pass it through the network while removing undesired noise. So, the power of deep learning-based SEAs is used to learn the complex nonlinear function that maps from the noisy to clean speech, and then to separate the desired signal from the undesired one [129].

6.2. Hands-Free Communication Systems

Hands-free communication allows communication without any physical contact with the loudspeaker or the microphone. The signal quality in these systems is often unsatisfactory due to echo. Therefore, echo cancellation plays an important role in improving speech properties [130], where echo cancellation is considered one of speech enhancement types. In recent years, the demand for hands-free communication in vehicles or teleconferencing systems has led to a significant increase in research and development for this type of device. The success of these systems depends on the quality of the processed speech that is gained, which is polluted by different types of noise and interferences. Thus, the signals obtained by the microphones of the system before being transmitted through the communication channel are usually enhanced [131].

6.3. Speech Recognition Technologies

Speech Recognition (is also known as automatic speech recognition (ASR) or computer speech recognition) is the process of converting human speech signals to a sequence of words (written format). ASR has been used in computers to enable them to follow human voice commands and understand human languages. Today, these systems find widespread application in jobs that require a human-machine interface, such as automatic call processing in telephone networks, query-based information systems that provide updated travel information, stock price quotations, weather reports, data entry, and voice dictation. Additionally, they aid in providing access to information, and have applications in travel, banking, commands, avionics, automobile portals, speech transcription for people with disabilities (blind people), supermarkets, etc. [132].

Speech enhancement optimizes the quality of ASR. Basically, for automatic speaker identification (SID), there are four basics SEA that are used before the processing stage to reduce background noise: spectral subtraction (SS), statistical model based (MMSE), subspace, (PKLT) and Wiener filtering (SS, MMSE, WIN, and PKLT). To optimize the single-channel speaker identification system, the spectral subtraction method can be implemented, while for the multi-channel speaker identification system, the enhancement has been achieved by using adaptive noise cancellation and delay-and-sum speech beamforming methods [133].

6.4. Speech-Based Biometric System

Biometric speech recognition is an effective security measure and one of the most important applications of speech enhancement technology. The biometric system is used to communicate with the authorized individual and the network [133]. Biometric person authentication is the task of verifying a person's identity using human characteristics or attributes to limit access to a proposed service. This system is one of the applications of biometric person authentication systems. Usually, fingerprints or a face image/video are used as the biometric for such applications. Person authentication using speech biometric is generally termed as speaker verification (SV). New developments in SV research induced the utilization of speech as a biometric for use in practical person authentication systems, such as attendance systems, with logical reliability. Based on the limitation on the text content of the speech data, SV systems can be categorized into two types: text dependent and text independent. In a text-dependent SV system, the text is fixed or identified to the system, while in a text-independent SV system, the user is able to say anything [134].

In summary, speech signaling is of paramount importance and plays a pivotal role in various signal processing applications, and it has recently undergone tremendous development based on deep learning. Speech signaling is closely related to computational linguistics, human-machine interaction, psycholinguistics, across various applied fields,

and natural language processing. Figure 8 is a detailed classification of speech processing fields (adopted from [135]).

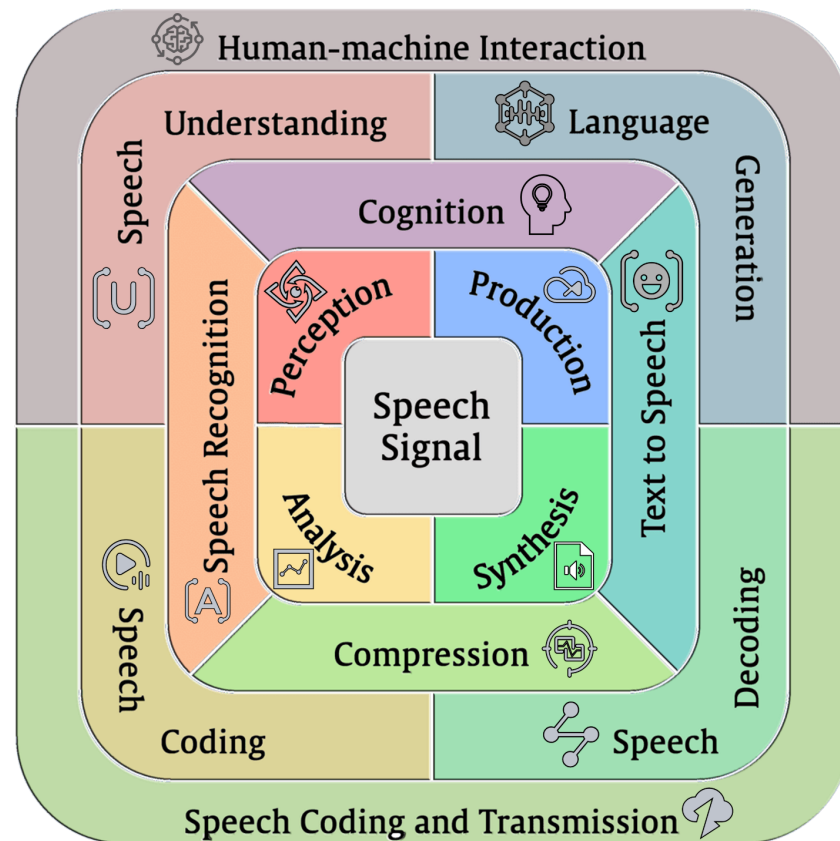


Figure 8. A detailed classification of speech processing fields.

7. Limitations and Challenges of SEA

Despite advances in speech enhancement algorithms, results in real-world environments should be more effective at reducing different types of noise without damaging speech components. This is mainly because the noise signal and speech signal sometimes differ significantly from what the algorithms were designed to handle. Therefore, several key limitations and challenges are summarized in this review, as shown in the following sub-sections, to highlight the points that should be considered when designing SEA. The main challenges are outlined below:

7.1. Generalization to Invisible Noise Types

In terms of the concept of deep learning-based SEA, deep learning architectures such as (DNN, CNN, and UNET) are considered the main part of the enhancement process. Within the framework of this type of enhancement, it was observed that the noise reduction function is highly affected by the types of noise the model was trained on. When the model encounters new or unusual sounds (e.g., the sound of an instrument not in the data), the performance drops significantly. Although large training sets increase generalization, a huge number of noise types may still be undefined, since in real-world scenarios there can be almost countless types of noise [136]. So, one of the most important limitations of the speech enhancement process is the limited training data or its dependence on identical environments, as this means the model's ability becomes weak in terms of generalizing when dealing with types of noise that are not visible during training, causing a degradation in performance in real-world environments.

7.2. Trade-Off Between Noise Reduction and Speech Distortion

Each algorithm that tries to remove noise may cause distortion in the speech signal components. If the filter is powerful, the noise decreases, but the quality of the speech decreases simultaneously. Therefore, a significant challenge is to make a balance between reducing noise and preserving speech quality so that increasing the noise reduction strength result in noise attenuation as a positive effect, but signal distortion as a negative effect. The perfect situation of total noise removal without distortion can never be realized in practice (unless there is a former knowledge available on the original speech and noise signals). Therefore, the major challenge in designing SEA is to find the best trade-off between achieving a significant reduction in noise while simultaneously preserving the speech signal without distortion in its components [137].

7.3. Phase Estimation Problem

Many SEA algorithms, such as spectral subtraction or the Wiener filter, concentrate on magnitude and neglect phase. But phase has an important impact on speech quality, especially at higher frequencies. Modern algorithms such as UNET-, DCCRN- and GAN-based models try to address phase, but they increase computational complexity. The phase information guarantees consistency across time frames and between adjacent frequency bins in the time–frequency domain. Precisely keeping it during enhancement helps to maintain the temporal characteristics and harmonic structures of the original speech. For the TF-domain phase-aware speech enhancement methods, some focused on enhancing the short-time complex spectrum, which implied recovering both magnitude and phase within the complex domain [138]. However, reusing phase from noisy speech has clear limitations, especially under very noisy conditions, or, in other words, when the signal-to-noise ratio (SNR) is low [139].

7.4. Computational Complexity and Real-Time Constraints

DNN and transformer-based algorithms require a high processing power (GPU/TPU) for their implementation. This sometimes makes them unfit for systems with limited resources such as phones, hearing aids, or embedded systems. The challenge here is to develop minor and fast algorithms (lightweight models) without immolate performance. Lately, algorithms based on deep neural networks have led to a great speech enhancement performance improvement in terms of speech quality and intelligibility both for offline as well as online processing. Nevertheless, obtaining a resource-efficient and low-complexity system is still a challenge issue [140].

In the context of speech enhancement, there are some other significant limitations that affect the performance the enhancement process, such as improving both the properties (quality and intelligibility) of cleaned speech simultaneously, and the correct selection of speech and noise priors, such as Gaussian, Laplacian, and, in general, super-Gaussian, in the estimators that affect the performance of estimation processing. And, for high SNR conditions, the present SEA provides good enhancement and good results over low SNR conditions, which necessities developing an improved SEA in low SNR levels.

Therefore, advanced studies in the field of SEA should take into account the points mentioned above. Furthermore, these systems must be cost effective, fast, reliable, and capable of handling various types of background noise with greater accuracy and efficiency to improve performance across diverse speech signaling applications where, generally, there are trade-offs among main limitations which are complexity, latency, and performance in most SEA.

7.5. Reverberation and Echo

Reverberation is an audio phenomenon that happens when sound reflects from surrounding surfaces (such as the walls, ceiling, and floor) in an enclosed place, reaching after the original sound with a short delay. In simpler terms, the sound does not arrive just once, but as a direct sound plus weak, delayed reverberations. This causes overlapping audio clips which reduce the speech intelligibility and quality and make speech enhancement more difficult than normal noise [141]. Dereverberation algorithms are methods that aim to remove reverberation from recordings, and are a separate type of speech enhancement.

In a real situation, however, sometimes the recordings of speech are contaminated with background noise and reverberation at the same time. There have been some studies that have tried to remove both noise and reverberation from the speech signal, either by performing noise suppression and dereverberation in sequence or by suggesting an integrated approach like [142,143]. Despite important progress in speech enhancement, many open challenges remain, such as, firstly, data bias, which is considered to be a critical issue as many datasets are limited to specific noise types, which negatively affects model generalization in real-world scenarios. Secondly, model robustness is another concern, especially under noise mismatch, reverberation, and rapidly changing acoustic environments. Thirdly, the deployment of advanced deep learning models on low-resource or embedded systems remains challenging due to constraints on memory, computational power, and energy consumption. Addressing these issues is essential for the practical adoption of speech enhancement systems in real-world applications.

In general, the main limitations of enhancement process can be summarized in Figure 9.



Figure 9. The general SEA challenges and limitations.

8. Conclusions

This research presents a comprehensive review of speech enhancement algorithms, with a primary focus on categorizing algorithms depending on the number of channels and the adopted signal processing frameworks. Traditional methods, machine learning approaches, and deep learning-based techniques were comprehensively reviewed, which are used in speech enhancement processes, with a description of the features of each category and their area of use, in addition to their practical applications. In addition, different types of noise and their characteristics, along with different types of SEA datasets, are presented. Then, the quality and eligibility metrics were defined and discussed. This review also addressed the major challenges encountered by SEAs, such as limited training data, poor generalization to types of invisible noise, and the adverse effects of reverberation and echo; another critical challenge discussed is the difficult balance between reducing noise and maintaining speech quality. Recent research trends indicate an increasing reliance on advanced deep learning models, and the integration of noise removal with echo removal within a unified framework, along with the need to develop more efficient and generalizable models capable of operating in complex real-world environments. Future speech enhancement systems should emphasize on improving robustness to unseen and highly dynamic noise conditions through different training data. In addition, reducing computational complexity while preserving high enhancement performance is essential for real-time and embedded applications. Hybrid approaches that combine signal processing knowledge with data-driven learning are promising directions. Also, adaptive and noise-aware frameworks that dynamically adjust processing strategies based on acoustic conditions can enhance generalization. Finally, standardized benchmarking and fair evaluation practices remain crucial for guiding meaningful progress in the field of speech enhancement.

Author Contributions: Conceptualization, B.M.M.; methodology, N.T.A. and B.M.M.; formal analysis, N.T.A.; investigation, N.T.A. and B.M.M.; resources, N.T.A. and B.M.M.; writing—original draft preparation, N.T.A. and B.M.M.; writing—review and editing, B.M.M. and N.T.A.; and visualization, B.M.M. and N.T.A. 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: Data is contained within the article.

Acknowledgments: The authors would like to thank University of Baghdad for their general support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
CNN	Convolutional Neural Network
CRN	Convolutional Recurrent Network
CSII	Coherence Speech Intelligibility Index
DCCRN	Deep Complex Convolutional Recurrent Network
DCHT	Deep Complex Hybrid Transform
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DKTT	Discrete Krawtchouk–Tchebichef Transform

DNN	Deep Neural Network
DSP	Digital Signal Processing
DVAE	Deep Variational Dynamic Autoencoder
DWT	Discrete Wavelet Transform
EMD	Empirical Mean Decomposition
FFT	Fast Fourier Transform
FTB	Frequency Transformation Block
FWSSNR	Frequency-Weighted Segmental SNR
GAN	Generative Adversarial Network
GEVD	Generalized Eigenvalue Decomposition
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HDLCMV	Harmonic Decomposition Linearly Constrained Minimum Variance
HMM	Hidden Markov Model
ICA	Independent Component Analysis
JLSMD	Joint Low-Range Scattered Decomposition
LCMV	Linearly Constrained Minimum Variance
LLR	Log-Likelihood Ratio
LSTM	Long Short-Term Memory
MAP	Maximum A Posteriori
MFSC	Mel-Frequency Spectral Coefficient
MMSE	Minimum Mean Square Error
MMSE-LSA	Minimum Mean Square Error Log-Spectral Amplitude
STSA-MMSE	Short-Time Spectral Amplitude Minimum Mean Square Error Estimator
MNMF	Multi-Channel Non-Negative Matrix Factorization
MOS	Mean Opinion Score
MVDR	Minimum Variance Distortionless Response
NMF	Non-Negative Matrix Factorization
ODMVDR	Orthogonal Decomposition-Based Minimum Variance Distortionless Response
PDF	Probability Density Function
PESQ	Perceptual Evaluation of Speech Quality
SC-SE	Single-Channel Speech Enhancement Algorithm
SEA	Speech Enhancement Algorithm
SID	Speaker Identification
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio
SNR	Signal-to-Noise Ratio
SS	Spectral Subtraction
SSA	Signal Subspace Algorithm
SSC-SEA	Supervised Single-Channel Speech Enhancement Algorithm
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility
SV	Speaker Verification
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TCNN	Temporal Convolutional Neural Network
UNET	U-shaped Network
USSC-SEA	Unsupervised Single-Channel Speech Enhancement Algorithm
VAE	Variational Autoencoder
VEM	Variational Expectation Maximization
WF	Wiener Filtering
WSST	Wavelet Synchro-Squeezing Transform

References

- Oruh, J.; Olaniyi, M. Deep learning approach for automatic speech recognition in the presence of noise. *South Fla. J. Dev.* **2024**, *5*, e4099.
- Naik, D.C.; Murthy, A.S.; Nuthakki, R. A literature survey on single channel speech enhancement techniques. *Int. J. Sci. Technol. Res.* **2020**, *9*, 5082–5091.
- Hussein, H.A.; Hameed, S.M.; Mahmmod, B.M.; Abdhussain, S.H.; Hussain, A.J. Dual stages of speech enhancement algorithm based on super gaussian speech models. *J. Eng.* **2023**, *29*, 1–13.
- Loizou, P.C. *Speech Enhancement: Theory and Practice*; CRC Press: Boca Raton, FL, USA, 2007.
- Nasir, R.J.; Abdumohsin, H.A. Noise Reduction Techniques for Enhancing Speech. *Iraqi J. Sci.* **2024**, *65*, 5798–5818.
- Saleem, N.; Khattak, M.I.; Verdú, E. On improvement of speech intelligibility and quality: A survey of unsupervised single channel speech enhancement algorithms. *Int. J. Interact. Multimed. Artif. Intell.* **2020**, *6*, 78–89.
- Alam, M.J.; O'Shaughnessy, D. Perceptual improvement of Wiener filtering employing a post-filter. *Digit. Signal Process.* **2011**, *21*, 54–65.
- Mahmmod, B.M.; Abdhussain, S.H.; Al-Haddad, S.A.R.; Jassim, W.A. Low-distortion MMSE speech enhancement estimator based on Laplacian prior. *IEEE Access* **2017**, *5*, 9866–9881.
- Ephraim, Y.; Van Trees, H.L. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 251–266.
- 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.
- Diehl, P.U.; Zilly, H.; Sattler, F.; Singer, Y.; Kepp, K.; Berry, M.; Hasemann, H.; Zippel, M.; Kaya, M.; Meyer-Rachner, P.; et al. Deep learning-based denoising streamed from mobile phones improves speech-in-noise understanding for hearing aid users. *Front. Med Eng.* **2023**, *1*, 1281904.
- Wang, Z.Q.; Wichern, G.; Watanabe, S.; Le Roux, J. STFT-domain neural speech enhancement with very low algorithmic latency. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *31*, 397–410.
- Muhammed Shifas, P.V.; Adiga, N.; Tsiaras, V.; Stylianou, Y. A non-causal FFTNet architecture for speech enhancement. In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), Graz, Austria, 15–19 September 2019.
- Martin, R.; Breithaupt, C. Speech enhancement in the DFT domain using Laplacian speech priors. In Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC), Kyoto, Japan, 8–11 September 2003; pp. 87–90.
- Geng, C.; Wang, L. End-to-end speech enhancement based on discrete cosine transform. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 27–29 June 2020; pp. 379–383.
- Lee, S.; Wang, S.S.; Tsao, Y.; Hung, J. Speech enhancement based on reducing the detail portion of speech spectrograms in modulation domain via discrete wavelet transform. In Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei City, Taiwan, 26–29 November 2018; pp. 16–20.
- Matsubara, T.; Kakutani, A.; Iwai, K.; Kurokawa, T. Improvement of wavelet-synchrosqueezing transform with time shifted angular frequency. *J. Adv. Simul. Sci. Eng.* **2023**, *10*, 53–63.
- Mahmmod, B.M. A Speech Enhancement Framework Using Discrete Krawtchouk-Tchebichef Transform. Ph.D. Thesis, Universiti Putra Malaysia, Seri Kembangan, Malaysia, 2018.
- Koduri, S.K.; T., K.K. Hybrid Transform Based Speech Band Width Enhancement Using Data Hiding. *Trait. Du Signal* **2022**, *39*, 969.
- Li, J.; Li, J.; Wang, P.; Zhang, Y. DCHT: Deep complex hybrid transformer for speech enhancement. In Proceedings of the 2023 Third International Conference on Digital Data Processing (DDP), Luton, UK, 27–29 November 2023; pp. 117–122.
- Mehrish, A.; Majumder, N.; Bharadwaj, R.; Mihalcea, R.; Poria, S. A review of deep learning techniques for speech processing. *Inf. Fusion* **2023**, *99*, 101869.
- Saleem, N.; Khattak, M.I. A review of supervised learning algorithms for single channel speech enhancement. *Int. J. Speech Technol.* **2019**, *22*, 1051–1075.
- Yousif, S.T.; Mahmmod, B.M. Speech Enhancement Algorithms: A Systematic Literature Review. *Algorithms* **2025**, *18*, 272.
- Weninger, F.; Erdogan, H.; Watanabe, S.; Vincent, E.; Le Roux, J.; Hershey, J.R.; Schuller, B. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In Proceedings of the International Conference on Latent Variable Analysis and Signal Separation, Liberec, Czech Republic, 25–28 August 2015; pp. 91–99.
- Wang, Z.Q.; Wang, D. All-Neural Multi-Channel Speech Enhancement. In Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech), Hyderabad, India, 2–6 September 2018; pp. 3234–3238.
- Cohen, I.; Huang, Y.; Chen, J.; Benesty, J.; Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009.

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; pp. 200–203.
28. Gupta, M.; Singh, R.K.; Singh, S. Analysis of optimized spectral subtraction method for single channel speech enhancement. *Wirel. Pers. Commun.* **2023**, *128*, 2203–2215.
29. Ephraim, Y.; Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust.* **2003**, *32*, 1109–1121.
30. Cohen, I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 466–475.
31. Bahrami, M.; Faraji, N. Minimum mean square error estimator for speech enhancement in additive noise assuming Weibull speech priors and speech presence uncertainty. *Int. J. Speech Technol.* **2021**, *24*, 97–108.
32. Lotter, T.; Vary, P. Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. *EURASIP J. Adv. Signal Process.* **2005**, *2005*, 354850.
33. Ranjbaryan, R.; Abutalebi, H.R. Multiframe maximum a posteriori estimators for single-microphone speech enhancement. *IET Signal Process.* **2021**, *15*, 467–481.
34. Lim, J.S.; Oppenheim, A.V. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE* **2005**, *67*, 1586–1604.
35. Dwivedi, S.; Khunteta, A. Performance Comparison among Different Wiener Filter Algorithms for Speech Enhancement. *Int. J. Microsystems IoT* **2024**, *2*, 505–514.
36. Garg, A. Speech enhancement using long short term memory with trained speech features and adaptive Wiener filter. *Multimed. Tools Appl.* **2023**, *82*, 3647–3675.
37. Jaiswal, R.K.; Yeduri, S.R.; Cenkeramaddi, L.R. Single-channel speech enhancement using implicit Wiener filter for high-quality speech communication. *Int. J. Speech Technol.* **2022**, *25*, 745–758.
38. Sun, C.; Mu, J. An eigenvalue filtering based subspace approach for speech enhancement. *Noise Control Eng. J.* **2015**, *63*, 36–48.
39. Chengli, S.U.N.; Jianxiao, X.I.E.; Yan, L. A signal subspace speech enhancement approach based on joint low-rank and sparse matrix decomposition. *Arch. Acoust.* **2016**, *41*, 245–254.
40. Benesty, J.; Chen, J.; Huang, Y. *Speech Enhancement in the Karhunen-Loève Expansion Domain*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2011.
41. 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.
42. Capon, J. Investigation of long-period noise at the large aperture seismic array. *J. Geophys. Res.* **1969**, *74*, 3182–3194.
43. Benesty, J.; Chen, J.; Habets, E.A.P. *Speech Enhancement in the STFT Domain*; Springer Science & Business Media: Heidelberg, Germany, 2011.
44. Pados, D.A.; Karystinos, G.N. An iterative algorithm for the computation of the MVDR filter. *IEEE Trans. Signal Process.* **2001**, *49*, 290–300.
45. Frost, O.L. An Algorithm for Linearly Constrained Adaptive Array Processing. *Proc. IEEE* **1972**, *60*, 926–935.
46. Christensen, M.G.; Jakobsson, A. Optimal filter designs for separating and enhancing periodic signals. *IEEE Trans. Signal Process.* **2010**, *58*, 5969–5983.
47. Wiener, F.M. The diffraction of sound by rigid disks and rigid square plates. *J. Acoust. Soc. Am.* **1949**, *21*, 334–347.
48. Hadei, S.; Lotfizad, M. A family of adaptive filter algorithms in noise cancellation for speech enhancement. *arXiv* **2011**, arXiv:1106.0846.
49. Frazier, R.; Samsam, S.; Braid, L.; Oppenheim, A. Enhancement of speech by adaptive filtering. In Proceedings of the ICASSP'76 IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA, 12–14 April 1976; pp. 251–253.
50. Lim, J.; Oppenheim, A. All-pole modeling of degraded speech. *IEEE Trans. Acoust.* **1978**, *26*, 197–210.
51. Ephraim, Y. A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Trans. Signal Process.* **2002**, *40*, 725–735.
52. Xiang, Y.; Shi, L.; Højvang, J.L.; Rasmussen, M.H.; Christensen, M.G. A novel NMF-HMM speech enhancement algorithm based on Poisson mixture model. In Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 721–725.
53. Xiang, Y.; Shi, L.; Højvang, J.L.; Rasmussen, M.H.; Christensen, M.G. A speech enhancement algorithm based on a non-negative hidden Markov model and Kullback-Leibler divergence. *EURASIP J. Audio Speech Music Process.* **2022**, *2022*, 22.
54. Scalart, P.; Filho, J.V. 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; pp. 629–632.
55. Bolner, F.; Goehring, T.; Monaghan, J.; Van Dijk, B.; Wouters, J.; Bleeck, S. Speech enhancement based on neural networks applied to cochlear implant coding strategies. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 6520–6524.

56. Xia, B.; Bao, C. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Commun.* **2014**, *60*, 13–29.
57. Mitchell, T.M. Does machine learning really work? *AI Mag.* **1997**, *18*, 11–19.
58. Wang, D.; Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1702–1725.
59. Yuan, W.; Xia, B. A speech enhancement approach based on noise classification. *Appl. Acoust.* **2015**, *96*, 11–19.
60. Kinoshita, Y.; Hirakawa, R.; Kawano, H.; Nakashi, K.; Nakatoh, Y. Speech enhancement system using SVM for train announcement. In Proceedings of the 2021 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 10–12 January 2021; pp. 1–3.
61. Kong, Y.Y.; Mullangi, A.; Kokkinakis, K. Classification of fricative consonants for speech enhancement in hearing devices. *PLoS ONE* **2014**, *9*, e95001.
62. Manamperi, W.; Samarasinghe, P.N.; Abhayapala, T.D.; Zhang, J. GMM based multi-stage Wiener filtering for low SNR speech enhancement. In Proceedings of the 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), Bamberg, Germany, 5–8 September 2022; pp. 1–5.
63. Soundarya, M.; Anusuya, S.; Narayanan, L.K. Enhancing Automatic Speech Recognition Accuracy Using a Gaussian Mixture Model (GMM). In Proceedings of the 3rd International Conference on Optimization Techniques in the Field of Engineering (ICOFE-2024), Sambalpur, India, 22–24 January 2024.
64. Subramani, K.; Smaragdis, P.; Higuchi, T.; Souden, M. Rethinking Non-Negative Matrix Factorization with Implicit Neural Representations. *arXiv* **2024**, arXiv:2404.04439.
65. Lee, D.; Seung, H.S. Algorithms for non-negative matrix factorization. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 1 January 2000; Volume 13.
66. Virtanen, T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1066–1074.
67. Yuliani, A.R.; Amri, M.F.; Suryawati, E.; Ramdan, A.; Pardede, H.F. Speech enhancement using deep learning methods: A review. *J. Elektron. Dan Telekomun.* **2021**, *21*, 19–26.
68. Nguyen, B.T.; Wakabayashi, Y.; Geng, Y.; Iwai, K.; Nishiura, T. DNN-based phase estimation for online speech enhancement. *Acoust. Sci. Technol.* **2025**, *46*, 186–190.
69. Ke, Y.; Li, A.; Zheng, C.; Peng, R.; Li, X. Low-complexity artificial noise suppression methods for deep learning-based speech enhancement algorithms. *EURASIP J. Audio Speech Music Process.* **2021**, *2021*, 17.
70. Wu, J.; Hua, Y.; Yang, S.; Qin, H.; Qin, H. Speech enhancement using generative adversarial network by distilling knowledge from statistical method. *Appl. Sci.* **2019**, *9*, 3396.
71. Pascual, S.; Bonafonte, A.; Serra, J. SEGAN: Speech enhancement generative adversarial network. *arXiv* **2017**, arXiv:1703.09452.
72. Serbest, S.; Stojkovic, T.; Cernak, M.; Harper, A. DeepFilterGAN: A Full-band Real-time Speech Enhancement System with GAN-based Stochastic Regeneration. *arXiv* **2025**, arXiv:2505.23515.
73. Shetu, S.S.; Habets, E.A.P.; Brendel, A. GAN-based speech enhancement for low SNR using latent feature conditioning. In Proceedings of the ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1–5.
74. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377.
75. Bhat, G.S.; Shankar, N.; Reddy, C.K.A.; Panahi, I.M.S. A real-time convolutional neural network based speech enhancement for hearing impaired listeners using smartphone. *IEEE Access* **2019**, *7*, 78421–78433.
76. Pandey, A.; Wang, D. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6875–6879.
77. Lin, X.; Leglaive, S.; Girin, L.; Alameda-Pineda, X. Unsupervised speech enhancement with deep dynamical generative speech and noise models. *arXiv* **2023**, arXiv:2306.07820.
78. Bie, X.; Leglaive, S.; Alameda-Pineda, X.; Girin, L. Unsupervised speech enhancement using dynamical variational autoencoders. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2993–3007.
79. Cui, Z.; Zhang, S.; Chen, Y.; Gao, Y.; Deng, C.; Feng, J. Semi-Supervised Speech Enhancement Based On Speech Purity. In Proceedings of the ICASSP 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Rhodes, Greece, 4–10 June 2023; pp. 1–5.
80. Ullah, R.; Islam, M.S.; Ye, Z.; Asif, M. Semi-supervised transient noise suppression using OMLSA and SNMF algorithms. *Appl. Acoust.* **2020**, *170*, 107533.

81. Purushotham, U.; Chethan, K.; Manasa, S.; Meghana, U. Speech enhancement using semi-supervised learning. In Proceedings of the 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, UK, 17–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 381–385.
82. Bando, Y.; Mimura, M.; Itoyama, K.; Yoshii, K.; Kawahara, T. Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 716–720.
83. Doclo, S.; Kellermann, W.; Makino, S.; Nordholm, S.E. Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. *IEEE Signal Process. Mag.* **2015**, *32*, 18–30.
84. Zaland, Z.; Mustafa, M.B.; Kiah, M.L.M.; Ting, H.N.; Yusoo, M.A.M.; Don, Z.M.; Muthaiyah, S. Multichannel speech enhancement for automatic speech recognition: A literature review. *PeerJ Comput. Sci.* **2025**, *11*, e2772.
85. Xie, Y.; Zou, T.; Sun, W.; Xie, S. Blind extraction-based multichannel speech enhancement in noisy and reverberation environments. *IEEE Sens. Lett.* **2025**, *9*, 7001104.
86. Kajala, M. A Multi-Microphone Beamforming Algorithm with Adjustable Filter Characteristics. Ph.D. Thesis, Tampere University, Tampere, Finland, 2021.
87. Sawada, H.; Kameoka, H.; Araki, S.; Ueda, N. New formulations and efficient algorithms for multichannel NMF. In Proceedings of the 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 16–19 October 2011; pp. 153–156.
88. Pandey, A.; Wang, D. On cross-corpus generalization of deep learning based speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2489–2499.
89. Nossier, S.A.; Wall, J.; Moniri, M.; Glackin, C.; Cannings, N. An experimental analysis of deep learning architectures for supervised speech enhancement. *Electronics* **2020**, *10*, 17.
90. Cohen, I.; Berdugo, B. Speech enhancement for non-stationary noise environments. *Signal Process.* **2001**, *81*, 2403–2418.
91. Zhao, Y.; Wang, D. Noisy-Reverberant Speech Enhancement Using DenseUNet with Time-Frequency Attention. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech), Shanghai, China, 25–29 October 2020; pp. 3261–3265.
92. Reddy, C.K.A.; Beyrami, E.; Pool, J.; Cutler, R.; Srinivasan, S.; Gehrke, J. A scalable noisy speech dataset and online subjective test framework. *arXiv* **2019**, arXiv:1909.08050.
93. Richter, J.; Wu, Y.; Krenn, S.; Welker, S.; Lay, B.; Watanabe, S.; Richard, A.; Gerkmann, T. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. *arXiv* **2024**, arXiv:2406.06185.
94. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L.; Zue, V. TIMIT *Acoustic-Phonetic Continuous Speech Corpus*; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.
95. Botinhao, C.V.; Wang, X.; Takaki, S.; Yamagishi, J. Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech), San Francisco, CA, USA, 8–12 September 2016; pp. 352–356.
96. Barker, J.P.; Marxer, R.; Vincent, E.; Watanabe, S. The CHiME challenges: Robust speech recognition in everyday environments. In *New Era for Robust Speech Recognition: Exploiting Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 327–344.
97. Watanabe, S.; Mandel, M.; Barker, J.; Vincent, E.; Arora, A.; Chang, X.; Khudanpur, S.; Manohar, V.; Povey, D.; Raj, D.; et al. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv* **2020**, arXiv:2004.09249.
98. Reddy, C.K.A.; Gopal, V.; Cutler, R.; Beyrami, E.; Cheng, R.; Dubey, H.; Matushevych, S.; Aichner, R.; Aazami, A.; Braun, S.; et al. The Interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. *arXiv* **2020**, arXiv:2005.13981.
99. Kong, Q.; Liu, H.; Du, X.; Chen, L.; Xia, R.; Wang, Y. Speech enhancement with weakly labelled data from AudioSet. *arXiv* **2021**, arXiv:2102.09971.
100. Li, H.; Yamagishi, J. DDS: A new device-degraded speech dataset for speech enhancement. *arXiv* **2021**, arXiv:2109.07931.
101. Mitra, V.; Wang, W.; Franco, H.; Lei, Y.; Bartels, C.; Graciarena, M. Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions. In Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech), Singapore, 14–18 September 2014; pp. 895–899.
102. Kinoshita, K.; Delcroix, M.; Yoshioka, T.; Nakatani, T.; Habets, E.; Haeb-Umbach, R.; Leutnant, V.; Sehr, A.; Kellermann, W.; Maas, R.; et al. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013; pp. 1–4.
103. Maciejewski, M.; Wichern, G.; McQuinn, E.; Le Roux, J. WHAMR!: Noisy and reverberant single-channel speech separation. In Proceedings of the ICASSP 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 696–700.

104. Mahmmod, B.M.; Baker, T.; Al-Obeidat, F.; Abdulhussain, S.H.; Jassim, W.A. Speech enhancement algorithm based on super-Gaussian modeling and orthogonal polynomials. *IEEE Access* **2019**, *7*, 103485–103504.
105. Hu, Y.; Loizou, P.C. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Commun.* **2007**, *49*, 588–601.
106. Thiemann, J.; Ito, N.; Vincent, E. The Diverse Environments Multi-Channel Acoustic Noise Database (DEMAND): A Database of Multichannel Environmental Noise Recordings. In *Proceedings of Meetings on Acoustics*; Acoustical Society of America: Melville, NY, USA, 2013. Available online: <https://cir.nii.ac.jp/crid/1881709542901868672> (accessed on 5 January 2026).
107. Vincent, E.; Watanabe, S.; Nugraha, A.A.; Barker, J.; Marxer, R. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Comput. Speech Lang.* **2017**, *46*, 535–557.
108. Gemmeke, J.F.; Ellis, D.P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 5–9 March 2017; pp. 776–780.
109. Qian, Y.; Bi, M.; Tan, T.; Yu, K. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2263–2276.
110. Ribas, D.; Llombart, J.; Miguel, A.; Vicente, L. Deep speech enhancement for reverberated and noisy signals using wide residual networks. *arXiv* **2019**, arXiv:1901.00660.
111. Loizou, P.C. Speech quality assessment. In *Multimedia Analysis, Processing and Communications*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 623–654.
112. Hu, Y.; Loizou, P.C. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *16*, 229–238.
113. 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)*, Khorfakkan, United Arab Emirates, 6–8 November 2024; pp. 31–36.
114. Beerends, J.G.; Hekstra, A.P.; Rix, A.W.; Hollier, M.P. Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part II-psychoacoustic model. *J. Audio Eng. Soc.* **2002**, *50*, 765–778.
115. de Oliveira, D.; Welker, S.; Richter, J.; Gerkmann, T. The PESQetarian: On the Relevance of Goodhart’s Law for Speech Enhancement. *arXiv* **2024**, arXiv:2406.03460.
116. Grundlehner, B.; Lecocq, J.; Balan, R.; Rosca, J. Performance assessment method for speech enhancement systems. In *Proceedings of the 1st Annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2005)*, Antwerp, Belgium, 19–20 April 2005; pp. 1–4.
117. Hou, M.; Kodrasi, I. Influence of Clean Speech Characteristics on Speech Enhancement Performance. *arXiv* **2025**, arXiv:2509.18885.
118. Kates, J.M.; Arehart, K.H. Coherence and the speech intelligibility index. *J. Acoust. Soc. Am.* **2005**, *117*, 2224–2237.
119. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136.
120. Medani, M.; Saleem, N.; Fkih, F.; Alohal, M.A.; Elmannai, H.; Bourouis, S. End-to-end feature fusion for jointly optimized speech enhancement and automatic speech recognition. *Sci. Rep.* **2025**, *15*, 22892.
121. Hwang, S.; Park, S.W.; Park, Y. Design of a Dual-Path Speech Enhancement Model. *Appl. Sci.* **2025**, *15*, 6358.
122. Jassim, W.A.; Paramesran, R.; Zilany, M.S.A. Enhancing noisy speech signals using orthogonal moments. *IET Signal Process.* **2014**, *8*, 891–905.
123. Ding, H.; Lee, T.; Soon, Y.; Yeo, C.K.; Dai, P.; Dan, G. Objective measures for quality assessment of noise-suppressed speech. *Speech Commun.* **2015**, *71*, 62–73.
124. Rosenbaum, T.; Winebrand, E.; Cohen, O.; Cohen, I. Deep-learning framework for efficient real-time speech enhancement and dereverberation. *Sensors* **2025**, *25*, 630.
125. Jepsen, S.D.; Christensen, M.G.; Jensen, J.R. A study of the scale invariant signal to distortion ratio in speech separation with noisy references. *arXiv* **2025**, arXiv:2508.14623.
126. Zhang, H.; Yu, M.; Wu, Y.; Yu, T.; Yu, D. Hybrid AHS: A hybrid of Kalman filter and deep learning for acoustic howling suppression. *arXiv* **2023**, arXiv:2305.02583.
127. Priyanka, S.S. A review on adaptive beamforming techniques for speech enhancement. In *Proceedings of the 2017 Innovations in Power and Advanced Computing Technologies*, Vellore, India, 21–22 April 2017; pp. 1–6.
128. Sakhnov, K.; Verteletskaya, E.; Simak, B. Dynamical energy-based speech/silence detector for speech enhancement applications. In *Proceedings of the World Congress on Engineering*, London, UK, 1–3 July 2009; Volume 2.
129. Nossier, S.A.; Rizk, M.R.M.; Moussa, N.D.; El Shehaby, S. Enhanced smart hearing aid using deep neural networks. *Alex. Eng. J.* **2019**, *58*, 539–550.
130. Grace, N.V.A.; Sumithra, M.G. Speech enhancement in hands free communication. In *Proceedings of the 4th WSEAS International Conference on Electronic, Signal Processing and Control*, Rio de Janeiro, Brazil, 25–27 April 2007; pp. 1–6.

131. Alvarez, D.A. Speech Enhancement Algorithms for Audiological Applications. Ph.D. Thesis, Universidad de Alcalá, Alcalá de Henares, Spain, 2013.
132. Kalamani, M.; Valarmathy, S.; Poonkuzhali, C.; JN, C. Feature selection algorithms for automatic speech recognition. In Proceedings of the 2014 International Conference on Computer Communication and Informatics, Udaipur Rajasthan, 14–16 November 2014; India, IEEE: Piscataway, NJ, USA, 2014; pp. 1–7.
133. 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.
134. Dey, S.; Barman, S.; Bhukya, R.K.; Das, R.K.; Haris, B.C.; Prasanna, S.M.; Sinha, R. Speech biometric based attendance system. In Proceedings of the 2014 Twentieth National Conference on Communications (NCC), Kanpur, India, 28 February–2 March 2014; pp. 1–6.
135. Natarajan, S.; Al-Haddad, S.A.R.; Ahmad, F.A.; Kamil, R.; Hassan, M.K.; Azrad, S.; Macleans, J.F.; Abdulhussain, S.H.; Mahmmod, B.M.; Saparkhojayev, N.; et al. Deep neural networks for speech enhancement and speech recognition: A systematic review. *Ain Shams Eng. J.* **2025**, *16*, 103405.
136. Rehr, R.; Gerkmann, T. Normalized features for improving the generalization of DNN based speech enhancement. *arXiv* **2017**, arXiv:1709.02175.
137. Brons, I.; Dreschler, W.A.; Houben, R. Detection threshold for sound distortion resulting from noise reduction in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* **2014**, *136*, 1375–1384.
138. Lu, Y.X.; Ai, Y.; Ling, Z.H. Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement. *Neural Netw.* **2025**, *189*, 107562.
139. Choi, H.S.; Kim, J.H.; Huh, J.; Kim, A.; Ha, J.W.; Lee, K. Phase-aware speech enhancement with deep complex U-Net. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
140. Sinha, R.; Rollwage, C.; Doclo, S. Low-complexity Real-time Single-channel Speech Enhancement Based on Skip-GRUs. In Proceedings of the Speech Communication, 15th ITG Conference, Aachen, Germany, 20–22 September 2023; pp. 181–185.
141. Valentini-Botinhao, C.; Yamagishi, J. Speech enhancement of noisy and reverberant speech for text-to-speech. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1420–1433.
142. Li, A.; Liu, W.; Luo, X.; Yu, G.; Zheng, C.; Li, X. A simultaneous denoising and dereverberation framework with target decoupling. *arXiv* **2021**, arXiv:2106.12743.
143. Song, S.; Cheng, L.; Luan, S.; Yao, D.; Li, J.; Yan, Y. An integrated multi-channel approach for joint noise reduction and dereverberation. *Appl. Acoust.* **2021**, *171*, 107526.

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.