*Sound & Vibration* 2026, 60(3), 2034.
doi: https://doi.org/10.59400/sv2034

Article

# Articulation index-based acoustic signal processing for enhanced speech intelligibility

**Mahesh Shankarrao Patil[1,*], Vijaykumar Varadarajan[2], Harsha Jitendra Sarode[3], Farook Sayyad[4], Rahul Krishna Sarawale[5], Shabnam Sayyad[6], Deshinta Arrova Dewi[7]**

[1] School of Bioengineering Sciences & Research, MIT ADT University, Pune 412201, India
[2] Department of Research, Swiss School of Business and Management, 1213 Geneva, Switzerland
[3] Department of Electronics & Telecommunication Engineering, Nutan Maharashtra Institute of Engineering & Technology, Pune 410507, India
[4] Department of Mechanical Engineering, Ajeenkya DY Patil School of Engineering, Pune 411081, India
[5] KPI Partners India Pvt. Ltd., Pune 411057, India
[6] Department of Artificial Intelligence and Machine Learning, AISSMS College of Engineering, Pune 411001, India
[7] Faculty of Data Science, INTI International University, Nilai 71800, Malaysia
*** Corresponding author:** Mahesh Shankarrao Patil, mpink.patil@gmail.com

**CITATION**

Patil MS, Varadarajan V, Sarode HJ, et al. Articulation index-based acoustic signal processing for enhanced speech intelligibility. Sound & Vibration. 2026; 60(3): 2034. https://doi.org/10.59400/sv2034

**ARTICLE INFO**

Received: 12 November 2024
Revised: 17 June 2025
Accepted: 25 June 2025
Available online: 9 May 2026



**Abstract:** This paper presents an innovative approach to improving speech intelligibility using the wavelet transform and the Articulation Index (AI) as an objective evaluation metric. Conventional methods such as the Modified Rhyme Test (MRT) and Mean Opinion Score (MOS) rely on subjective assessment, making them time-consuming and difficult to standardize. In contrast, AI provides a consistent and reliable measure of speech intelligibility across varying noise conditions. The proposed method applies wavelet packet transform to noisy speech signals, followed by a thresholding function to enhance signal quality and intelligibility. The processed speech is then reconstructed using the inverse wavelet transform. Experiments are conducted using the Noiseus database, which contains speech signals corrupted by real-world noises such as streets, airports, cockpits, and industrial environments like mechanical factories, with noise levels ranging from 0 dB to 15 dB. Three different enhancement methods are implemented, with the proposed method demonstrating superior performance in terms of AI values. Experimental validation is supported by plots and spectrograms, highlighting its effectiveness over existing approaches. The method leverages the multi-resolution property of the wavelet transform to preserve temporal characteristics while reducing noise across multiple frequency bands. Results show a significant improvement in AI values, indicating enhanced speech intelligibility under diverse noise conditions. This work contributes to acoustic speech enhancement by providing a robust, objective framework suitable for applications in noisy environments such as industrial communication systems, and this technique aligns closely with noise mitigation approaches used in structural and industrial surroundings. Additionally, the approach can be extended to industrial speech enhancement and environmental noise control.

**Keywords:** speech intelligibility enhancement; Noiseus database; thresholding function; articulation index (AI); wavelet transform; multi-resolution analysis; process innovation; auditory masking

## 1. Introduction

The ability of speech to be understood clearly and accurately is nothing but speech intelligibility [1]. This is very important for communication to be effective in the areas

where noise is prominent, such as public transport, industrial settings, airports, and train stations. Vital role is played by it in applications like telecommunications, industrial communication systems, and automatic speech recognition (ASR) systems [2]. Though intelligibility is greatly reduced by background noise, which in turn leads to communication failures that impact accessibility, efficiency, and safety [3].

The Modified Rhyme Test (MRT) and the Mean Opinion Score (MOS) are the subjective evaluation metrics relied upon traditional speech enhancement techniques, which include human auditors evaluating clarity of speech [4]. While explanatory, these approaches are unreliable, prone to variability, and time-consuming. Articulation Index (AI) as an objective metric has been adopted to address these limitations by providing a quantifiable measure of intelligibility based on contributions of frequency band [5,6].

Speech signals are frequently corrupted due to the mechanical system's interference. The examples of these can be HVAC systems, rotating machinery, and vehicular components that emit narrowband and broadband noise overlapping with human speech frequencies. The recommended technique, using its articulation-focused filtering, is essentially proficient in reducing this kind of interference. This capability makes it extremely favorable for applications demanding robust communication in surroundings where vibration noise is prevalent, such as aircraft cabins, trains, and factories [7].

Wavelet transform is a dominant signal processing technique for investigating non-stationary signals like speech [8, 9]. Contrasting Fourier transforms, wavelets provide both frequency and time resolution, making them effective for speech intelligibility enhancement. An extension of the discrete wavelet transform (DWT), the wavelet packet transform (WPT), allows fine control over frequency bands, preserving crucial speech components while decreasing noise [10, 11]. WPT's ability to target particular phonetic cues makes it suitable for improving speech intelligibility in noisy surroundings.

The aim of this study is to suggest an innovative method. This study proposes a novel approach joining AI-based intelligibility measures and wavelet packet transform to improve speech clarity. The noisy speech is first decomposed into frequency bands, and then thresholding is applied, and then the signal is reconstructed. So, by doing these, the speech features are selectively amplified while noise is suppressed [12]. The experimentation is done on the NOIZEUS database. Evaluations show that the proposed method outperforms traditional enhancement techniques, offering a scalable solution for practical applications in real-world scenarios, such as factory scenarios, emergency communication systems, assistive hearing devices, and public announcements.

Speech intelligibility in noisy and vibration-rich surroundings is a crucial challenge in acoustic engineering, predominantly across applications such as vehicle-mounted voice systems, air traffic control towers, and factory communication systems. In manufacturing locations, mechanical and machinery noise often overpowers vocal signals, demanding advanced noise suppression and adaptive filtering methods [13,14]. Studies have established how mechanical noise suppression and vibration signal processing—common in industrial and vehicular environments—

can be leveraged to improve speech clarity. In addition, investigation on automotive speech intelligibility has presented that articulation index–based measurements, when united with electroacoustic characterization and equalization under real-world noise circumstances, significantly increase the ability of communication in vehicle cabins. The findings of our study emphasize the importance of embedding the AI-driven methodology into speech systems, intercoms, and headsets used in very high noise acoustic engineering surroundings to improve the performance of the system without compromising communication reliability and safety.

## 2. Literature review

The extensive study has been carried out for the enhancement of speech intelligibility in noisy environments with different methodologies comprising deep learning, wavelet transformations, adaptive filtering, and articulation index-based (AI) methods [15]. These improvements have noteworthy implications for public communication systems, industrial communication systems, automatic speech recognition (ASR), and telecommunications.

The thought of the articulation index was initiated in introductory work in Stevens et al. [16] and was later formalized in IEEE standards for calculating intelligibility. Initial digital approaches of speech improvement, comprising spectral subtraction [17] and minimum mean-square error estimators [18], were predominantly concerned with improving overall speech quality, but their impact on intelligibility was variable. Research by Kates [19] and Plomp [20] emphasized the importance of speech cues (like consonant-vowel transitions) and how different types of noise influenced the AI. These insights guided subsequent developments in signal processing systems that directly targeted intelligibility, rather than merely improving signal clarity.

To tackle non-stationary industrial noise, wavelet packet-based methods began to dominate in the early 2000s. A wavelet packet enhancement algorithm guided by the articulation index was developed by Rasetshwane et al. [21], enabling localized time-frequency improvement of speech. Johnson et al. [22] and Zhang et al. [23] presented that speech intelligibility in transportation systems and factories might be considerably enhanced by highlighting frequency components crucial to AI. These methodologies permitted finer spectral resolution, crucial in differentiating speech from machine-generated narrow-band noise. The integration of transient detection further helped isolate consonants, which carry most of the intelligibility weight in spoken communication.

Casey et al. [24] presented harmonic emphasis and binary masking filters optimized using AI metrics, improving intelligibility equally in real-world and simulated environments. Furthermore, these procedures were polished by integrating adaptive comb filtering, which definitely suppressed periodic industrial noise without affecting speech harmonics. Recent work by Roy and Paliwal [25] presented sensitivity and robustness tuning of Kalman filters using AI-based metrics. These adaptive filters are principally relevant for factory communication systems where the acoustic environment changes quickly. Similarly, Healy et al. [26] developed SII-based (Speech Intelligibility Index) speech enhancement strategies for hearing-impaired users,

validating that models designed to maximize intelligibility outperform those focused purely on signal-to-noise ratio.

The last decade has seen a noteworthy swing toward deep learning-driven enhancement, frequently directed by AI or associated intelligibility metrics. Narayanan and Wang [27] established the use of ideal ratio masks estimated via deep neural networks (DNNs) for robust speech recognition and studied deep learning methods, observing that models trained with perceptual loss functions (e.g., using SII or AI) are more effective in intelligibility-sensitive applications. Mangati and Matassoni [28] proposed a bio-inspired auditory enhancement that mimics the human auditory system's ability to focus on intelligible speech regions. This mechanism has been validated across mobile systems and could be extended to wearable devices for factory workers. Similarly, time-frequency mask estimation and attention-aware denoising, respectively, tailored for mobile and low-latency contexts, are developed. These neural models, when trained with intelligibility-optimized objectives, can dynamically suppress non-speech elements while retaining speech cues, outperforming traditional enhancement methods in real-world noise scenarios [29–31].

Industrial settings often involve not just acoustic noise but mechanical vibration and structural reverberation. Work by Elliott and Nelson [32] emphasized active vibration isolation in high-speed rail environments to reduce low-frequency interference, thereby indirectly improving AI. Lucero and Munhall [33] addressed structural vibration in enclosed spaces, proposing damping systems that reduce resonance without adding latency. Fu and Wang [34] introduced hybrid materials with embedded damping properties, which could be used in vehicle dashboards or machinery enclosures to suppress noise before it even enters the air. These mechanical interferences supplement digital signal processing by dropping the base level of noise, thus improving the efficiency of AI-driven speech systems.

With the growth of embedded systems and edge computing, numerous investigators have explored real-time AI-based systems. Adaptive and statistical enhancement approaches, including noise estimation and STFT-based processing, play a key role in such systems [35–38]. These improvements recommend that AI-based enhancement is not only hypothetically sound but also practically deployable in harsh manufacturing surroundings.

Lu et al. [39] discovered frequency-domain diagnostics for vehicle interior noise. Using machine learning and vibration signals, they separated noise sources that reduce intelligibility. This diagnostic-first method permits directed improvement by focusing signal processing on noise-heavy regions crucial to the AI. Likewise, Hou et al. [40] applied wavelet packet feature fusion for shock absorber rattle detection, demonstrating how understanding mechanical noise generation leads to better acoustic signal planning.

Green et al. [41] combined beamforming with AI-guided masking to improve hearing aid performance in complex environments. Their system dynamically adjusts focus toward the speech source, suppressing surrounding noise, and is ideal for worker headsets in large industrial spaces. The potential of beamforming to increase AI by spatially isolating speech highlights its synergy with signal processing algorithms. Similar applications could transform communication in warehouses, vehicle cabins,

and control rooms. The integration of AI, wavelet transformations, adaptive filtering, and articulation index-based methods has significantly advanced speech enhancement techniques [42–44]. While traditional methods (e.g., spectral subtraction, Kalman filtering) have laid the foundation, modern AI and deep learning-based approaches continue to push the boundaries of real-time, adaptive speech enhancement. Future research should focus on hybrid techniques, combining machine learning with signal processing to further improve speech intelligibility across various real-world noisy environments.

## 3. Materials and methods

This study investigates the enhancement of speech intelligibility in noisy environments using the Noiseus database and wavelet packet transform (WPT). However, there are some critical gaps in the description of the database and certain aspects of the signal processing method used, particularly regarding the theoretical background and experimental validation of the decomposition depth. To provide a thorough understanding of the methodology, it is essential to address the limitations and suggest ways to enhance the approach for more robust results. This elaboration will cover the Noiseus database's details, the process of applying the WPT, the dynamic thresholding function, and the choice of decomposition depth in the WPT.

### 3.1. The Noiseus database

The Noiseus database is described as a collection of speech signals mixed with various types of environmental noise. While the study presents an impressive array of noise types, such as airport, babble, car, exhibition, restaurant, street, and train noises, some critical information about the database remains underexplored. For instance, it is unclear what the total number of speech samples in the Noiseus database is, how these samples were sourced, or what specific characteristics the speech samples have (e.g., gender, age, accent, or language). These details are essential for evaluating the database's representativeness and its applicability in various contexts.

The nonexistence of information on the quantity of examples limits the ability to assess the generalizability of the findings. For example, a dataset with a larger variety of speakers would likely yield more reliable results because it would be more representative of different speech characteristics. Additionally, knowing the sample size would allow for a clearer understanding of the robustness of the findings. If the dataset consists of a small number of samples, there may be a risk that the results could be skewed or not applicable to broader populations.

Moreover, the study does not mention the noise characteristics, such as the frequency range, power spectrum, or temporal dynamics of the noise types used. This information would be crucial for understanding the challenges that the speech enhancement method must overcome. For example, certain types of noise might have frequency characteristics that overlap more significantly with speech, making them more difficult to suppress.

Lastly, it would be beneficial for the study to include information about the duration of the speech samples and how these samples are mixed with noise at different

levels (0 dB, 5 dB, 10 dB, and 15 dB). The scenarios in which these noises are taken are ranging from Babble (crowd of people), Car, Exhibition Hall, Restaurant, Street, Airport, Train station, Train, and Factory floor. The range of noise levels is a crucial aspect since it will determine how the method performs under low, medium, and high levels of noise. Without clarifying how much speech data corresponds to each noise condition, it is difficult to evaluate the validity of the experimental results.

## 3.2. Wavelet packet transform (WPT) and decomposition depth

The WPT is a multi-resolution vibration analysis tool used for decomposing signals across different frequency bands, which is particularly effective for noisy speech enhancement. The study employs the WPT up to a depth of 4 for signal decomposition. However, the choice of depth is not sufficiently explained. In wavelet packet decomposition, the "depth" refers to how many levels of decomposition are performed, with each level increasing the resolution of the signal in the frequency domain. The depth of the decomposition plays a crucial role in determining the quality and accuracy of the signal representation. Higher depths can capture finer details of the signal, but also increase the computational complexity and potentially introduce overfitting to the noise present in the signal.

### 3.2.1. Lack of theoretical foundation for decomposition depth

The selection of depth 4 is presented without a theoretical justification. While it is clear that deeper decompositions provide finer resolutions, there is no discussion on why depth 4 is optimal for this specific scenario or how this choice balances mechanical noise reduction and preservation of speech clarity. Additionally, there is no reference to prior studies that may have determined the ideal decomposition depth for noisy speech enhancement. A solid theoretical basis for selecting decomposition depth could include considerations of the following:

- **Frequency range of speech**: Speech typically resides in the frequency range of 300 Hz to 3.4 kHz, and different levels of decomposition allow different ranges of frequency bands to be captured. If the decomposition depth is too high, it could result in unnecessary detail that may be more susceptible to noise interference.
- **Noise characteristics**: The depth could be selected based on the nature of the noise (e.g., stationary or non-stationary), which may affect the decomposition's ability to separate noise from speech.
- **Empirical verification**: A systematic experimental approach, such as cross-validation or testing across multiple noise conditions, could be used to determine the optimal decomposition depth. This would ensure that the chosen depth provides the best trade-off between noise suppression and speech quality.

Without these considerations, the rationale for using depth 4 remains unclear. Further research is needed to establish a more concrete connection between the decomposition depth and performance improvements in noisy environments. Comparative experiments across varying depths would provide insight into the ideal level of decomposition.

### 3.2.2. Experimental verification of decomposition depth

While the study relies on an empirical approach to test its method, there is no description of experiments that compare the performance of different decomposition depths. To improve the validity of the results, the study could conduct experiments using varying depths (e.g., 2, 4, and 6) to see how the depth of decomposition affects the enhancement results. These experiments could provide clearer guidance on the best practice for selecting decomposition depth in different noise scenarios.

## 3.3. Thresholding function and dynamic adaptation to Noise levels

After the WPT decomposition, the study applies a thresholding function to the wavelet coefficients to suppress noise while preserving speech features. The thresholding function serves as a key step in removing noise from the signal, as it selectively discards coefficients that are deemed to correspond to noise. While this approach is standard in signal processing, the study lacks details on how the thresholding function adapts to varying noise levels. A dynamic thresholding function that adjusts based on noise intensity within each frame would be essential to ensure that the method is robust across different noise conditions.

**Importance of dynamic thresholding:** Dynamic thresholding could be based on the Signal-to-Noise Ratio (SNR) within each frequency band. For example, in high-noise circumstances, a more aggressive thresholding function could be applied, whereas in low-noise surroundings, a less aggressive threshold would be more suitable to stop the loss of speech details. The precise criteria for threshold adjustment are not specified, and without this information, it is difficult to assess how well the dynamic thresholding function performs across different noise levels.

## 3.4. Evaluation and performance metrics

The Articulation Index (AI) is used to assess speech intelligibility, and the results validate that the recommended technique significantly increases clarity in noisy surroundings. Although the AI offers an objective measure of intelligibility, it is based on theoretical models of speech perception and may not fully reflect human perception in real-world conditions. The study could be strengthened by including additional performance metrics, such as:

- **Perceptual evaluation**: A listening assessment involving human subjects could deliver more tangible understandings into how well the improved speech is understood in comparison to the other methods.
- **Speech quality metrics**: Measures like Signal-to-Distortion Ratio (SDR) or Perceptual Evaluation of Speech Quality (PESQ) could help measure the quality of the improved speech signal.

The work uses spectrograms to visually check the improvement process. While spectrograms can offer a useful representation of the signal's frequency content, they are subjective and should be supplemented by quantifiable assessments to provide a more complete performance assessment.

A critical part of AI-based speech intelligibility improvement in industrial settings

comprises detailed characterization of both vibroacoustic phenomena and acoustic noise. Industrial noise is often broadband but may include dominant tonal components, mechanical harmonics, and impulsive transients, all of which influence AI calculations. To model these accurately, the full frequency spectrum of background noise must be analysed with high resolution (typically via Short-Time Fourier Transform or Wavelet Packet Decomposition), allowing identification of spectral valleys and peaks that overlap with speech formants (particularly 500 Hz–4,000 Hz, which are critical for intelligibility).

The mechanical noise spectrum—originating from sources such as compressors, motors, or structural vibrations—often introduces low-frequency components (<500 Hz) and modulated mid-frequency energy that masks speech consonants. This spectral masking directly reduces the effective Signal-to-Noise Ratio (SNR) in the most sensitive AI bands. As per ANSI S3.5-1997 AI standards, these masked bands are given reduced weightings in the intelligibility index calculation, resulting in lower overall AI values. Therefore, precise spectral decomposition is required to isolate the impact of vibroacoustic energy on individual frequency bands contributing to the articulation index. To account for structure-borne vibroacoustic noise, vibration acceleration and velocity signals are captured using piezoelectric accelerometers mounted at key locations (e.g., engine housing, control panel, and floor mountings). These signals are transformed into the frequency domain to assess resonance coupling with airborne sound paths. Cross-spectral analysis is then performed to quantify the coherence between vibration sources and acoustic emissions in the speech frequency band. This data enables predictive modelling of vibroacoustic interference patterns and helps in designing adaptive filters and noise-cancelling algorithms that prioritize frequency bands with the highest AI weighting. By integrating these measurements, the AI computation becomes context-sensitive, reflecting not only overall background levels but also the spectral and temporal structure of noise. This approach supports more precise tuning of speech enhancement systems, especially in dynamic industrial environments where both airborne and structural noise interact.

### 3.5. System block diagram

The process flow of the proposed methodology is illustrated in **Figure 1**, which shows the block diagram outlining the key steps: filtering the noisy speech, applying the wavelet transform, thresholding, and reconstructing the enhanced speech.

1. **Noisy speech**
   The input is a noisy speech signal, which contains both speech and background noise. The aim is to extract the clean speech while minimizing the impact of noise.
2. **Framing**
   The noisy speech signal is divided into small overlapping frames to facilitate further processing. Short-time processing assumes that speech is quasi-stationary over short durations (typically 20–40 ms).
3. **Wavelet transform**
   A discrete wavelet transform (DWT) is applied to the framed signals to obtain wavelet coefficients. Wavelets decompose the signal into different frequency

bands while preserving time-domain localization, making them effective for denoising.

4. **Wavelet coefficients**
   The transformed signal consists of approximation and detail coefficients. High-frequency components often correspond to noise, while low-frequency components retain essential speech features.
5. **Thresholding**
   A thresholding function is applied to suppress noise-dominated coefficients while preserving speech-related coefficients. Hard or soft thresholding techniques reduce unwanted noise while retaining signal integrity.
6. **Inverse wavelet transform**
   The modified coefficients are reconstructed using the inverse wavelet transform (IWT) to obtain a denoised speech signal. The IWT reconstructs the time-domain speech signal while retaining only essential components.
7. **Articulation index (AI)**
   AI is used to evaluate speech intelligibility after enhancement. AI measures how much of the speech signal is perceptible, ensuring effective noise suppression.
8. **Enhanced speech**
   The final output is an enhanced speech signal with minimized noise and improved clarity. The goal is to improve speech intelligibility and quality while preserving the natural characteristics of speech.

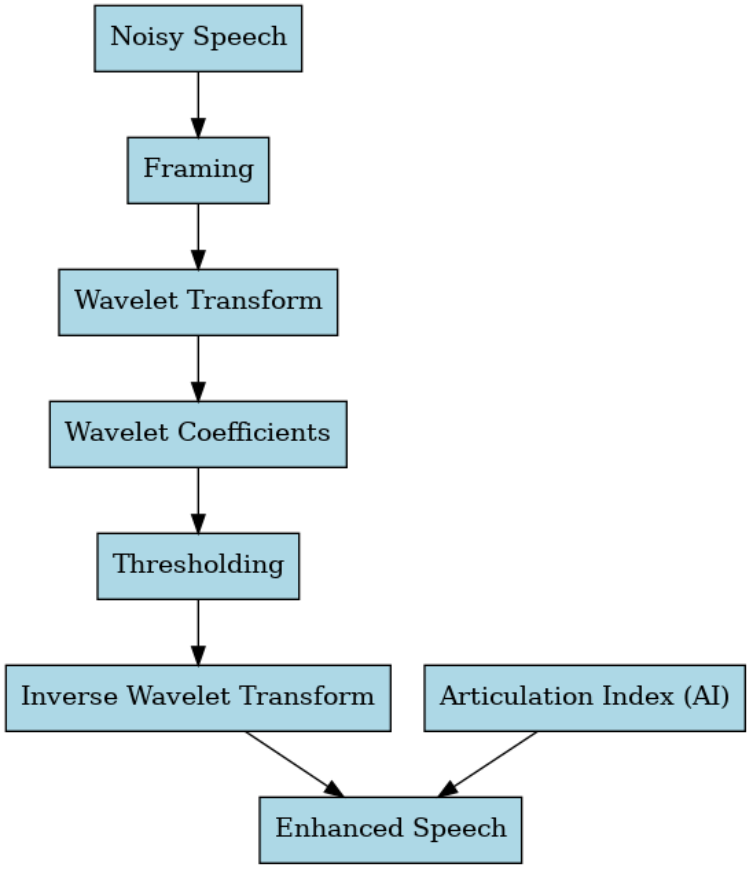



**Figure 1.** Process flow of the proposed methodology.

## 4. Results and discussion

In this study, the algorithm described in Method 3 was applied to a comprehensive set of speech samples, ranging from Sp-01 to Sp-30, sourced from the Noiseus database. The Noiseus database includes a variety of noisy speech signals, mixed with different types of environmental noise at various Signal-to-Noise Ratios (SNRs). To comprehensively assess the efficiency of this proposed speech improvement algorithm across different noise conditions, we have chosen a wide range of speech signals.

However, for the purpose of clearer demonstration and to preserve the focus on key findings, the results from Sp-01 to Sp-10 are chiefly presented and discussed in detail in this section. These specific speech samples denote a varied subsection of the database, offering a solid basis for assessing the algorithm's performance. The results from these samples will deliver insights into how well the suggested algorithm handles different noise types and levels.

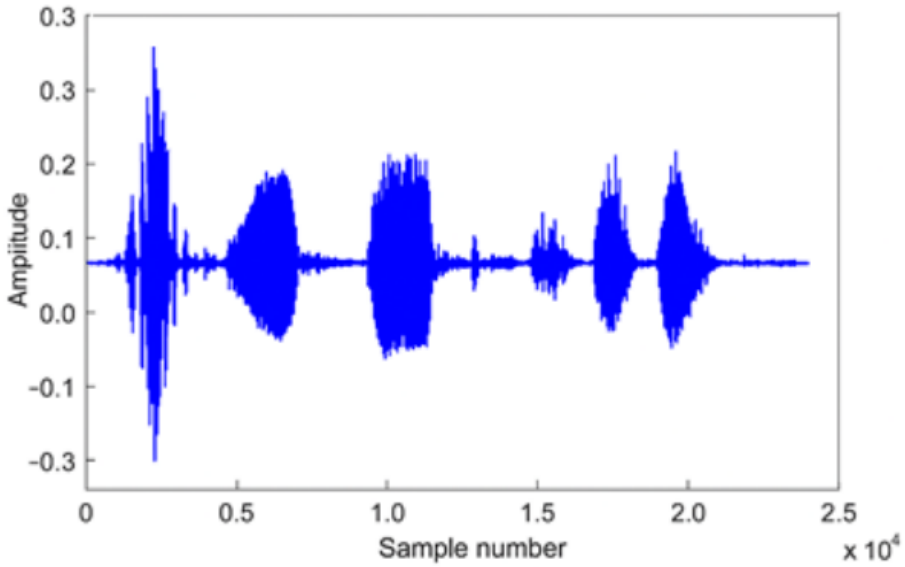
To demonstrate the usefulness of the proposed method, both time-domain plots and spectrograms are provided for visual comparison. The time-domain plots show the raw waveform of the clean speech signal, the noisy speech signal, and enhanced speech signals treated by different methods, including Method 1, Method 2, and the proposed method. These plots permit an easy assessment of how each method impacts the speech signal over time, emphasizing the degree to which noise is suppressed while conserving speech clarity.

The spectrograms offer a more comprehensive view of the frequency content of the signals over time. They help to visualize how noise affects the frequency distribution of the speech signal and how well the various enhancement approaches can suppress the noise while keeping the integrity of the speech features. By showing both time-domain plots and spectrograms, the figures offer complementary insights into the effectiveness of the recommended method in reducing noise and enhancing speech intelligibility through a range of noisy conditions.

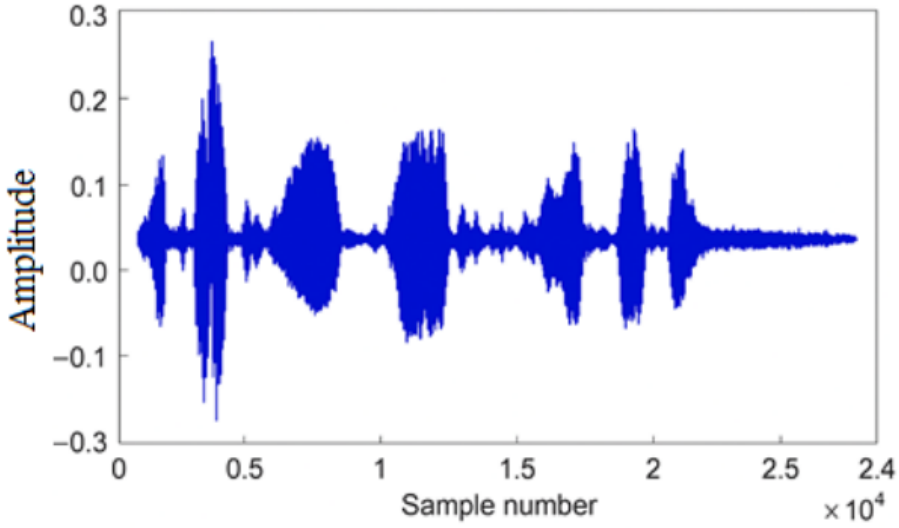
The figures below provide visual insights into the effectiveness of the proposed method through both plots and spectrograms.

**Figure 2** demonstrates the results of speech signals by plots:

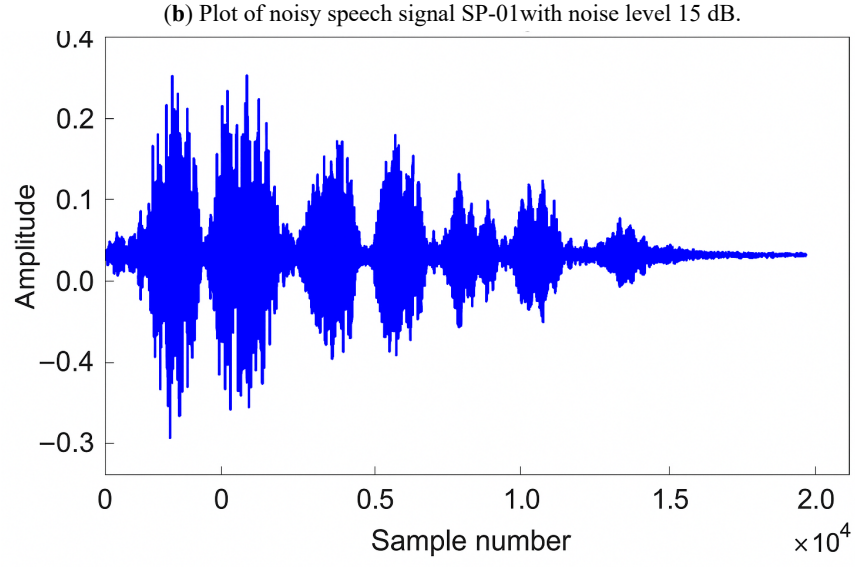
**Figure 2a** is plot of clean speech sample SP-01: This represents the original, noise-free speech signal; **Figure 2b** is plot of noisy speech signal SP-01 with noise level 15 dB: This demonstrates the degradation of the speech signal after the addition of noise at 15 dB SNR; **Figure 2c** is plot of enhanced speech signal SP-01 using Method 1: This plot shows the result of applying Method 1 (high-pass filter followed by wavelet packet transform and thresholding) on the noisy speech signal; **Figure 2d** is plot of enhanced speech signal SP-01 using Method 2: This shows the result after applying Method 2, where a different thresholding function is used, adapted based on noise levels; **Figure 2e** is plot of enhanced speech signal SP-01 using the Proposed Method: This shows the result after applying the proposed frame-based approach with dynamic thresholding for enhanced intelligibility.




(**a**) Plot of the clean speech signal SP-01.

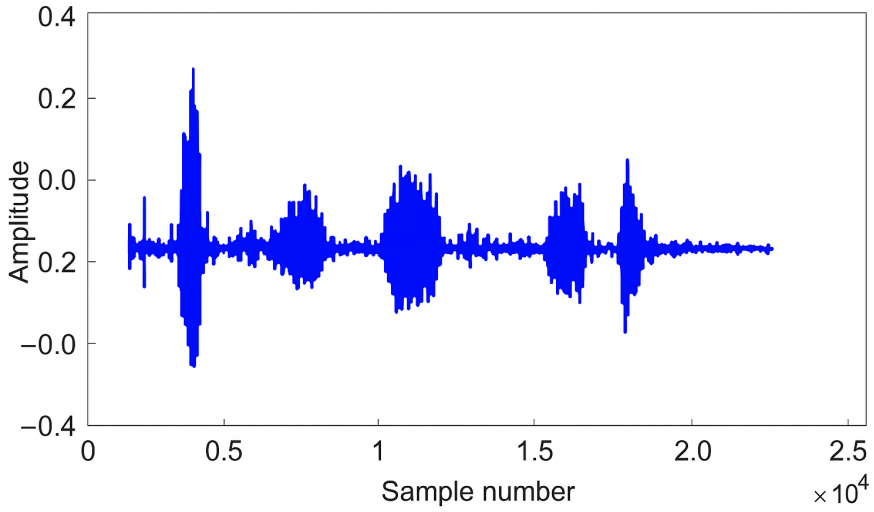



(**b**) Plot of noisy speech signal SP-01with noise level 15 dB.

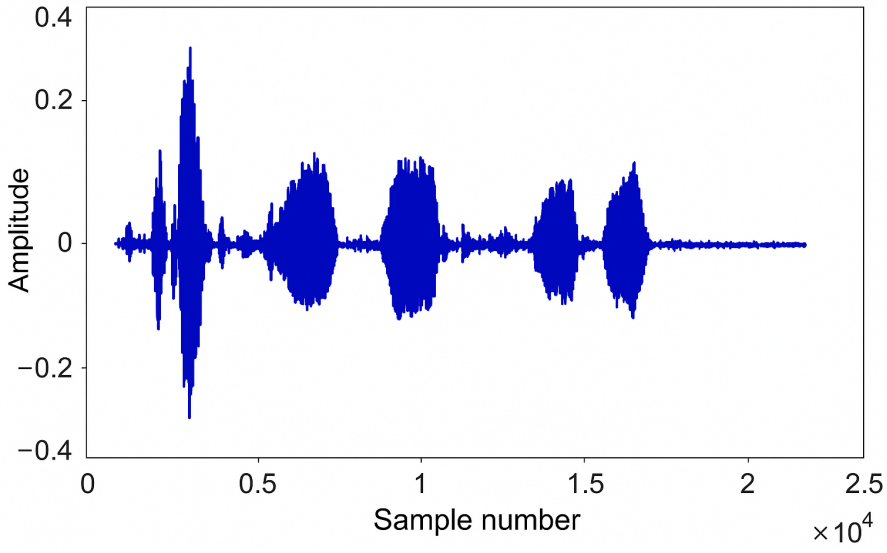



(**c**) Plot of enhanced speech signal SP-01 by Method 1.

**Figure 2.** *Cont.*




(**d**) Plot of enhanced speech signal SP-01 by Method 2.

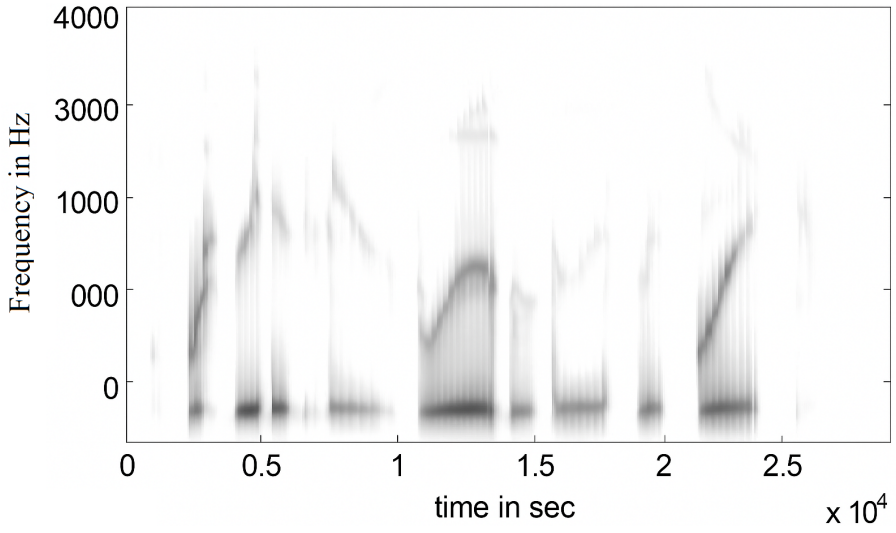



(**e**) Plot of enhanced speech signal SP-01 by the proposed method.

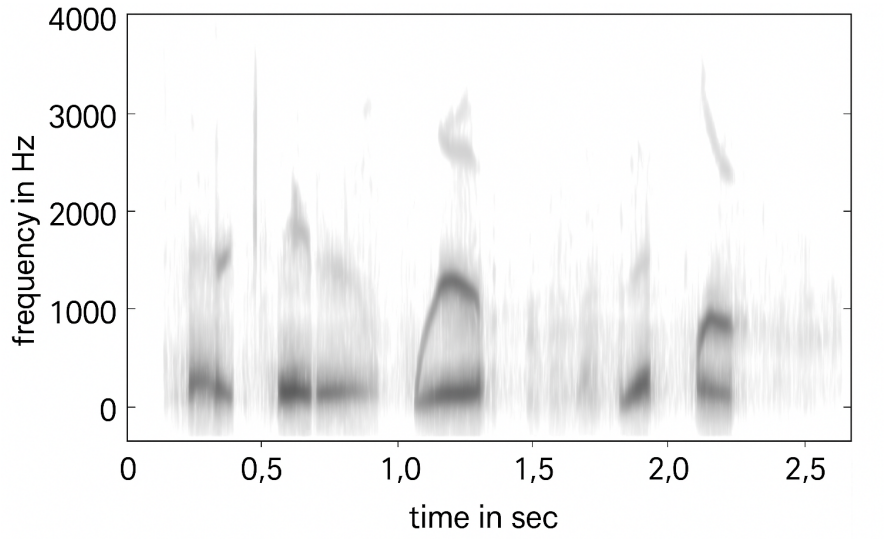
**Figure 2.** The results of speech signals (plots).

**Figure 3** shows the results of speech signals by spectrograms:

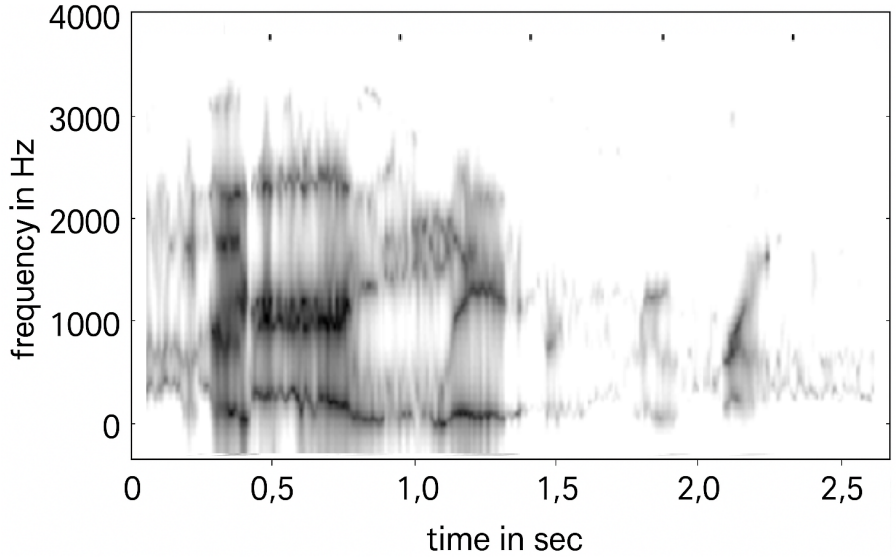
**Figure 3a** is spectrogram of clean speech sample SP-01: This spectrogram shows the frequency content of the clean speech signal, providing a reference for comparison; **Figure 3b** is spectrogram of noisy speech signal SP-01 with noise level 15 dB: This spectrogram highlights the distortion in frequency content due to the addition of noise at 15 dB SNR; **Figure 3c** is spectrogram of enhanced speech signal SP-01 by Method 1: This spectrogram demonstrates the effect of Method 1 on the noisy signal, showing some improvement in clarity but still retaining noise elements; **Figure 3d** is spectrogram of enhanced speech signal SP-01 by Method 2: This spectrogram shows the enhancement from Method 2, with mechanical noise reduction more noticeable compared to Method 1; **Figure 3e** is spectrogram of enhanced speech signal SP-01 by the Proposed Method: This spectrogram illustrates the improvement in speech clarity with the proposed method, showing significant mechanical noise reduction while retaining the important speech features.




(**a**) Spectrogram of Clean speech signal SP-01.

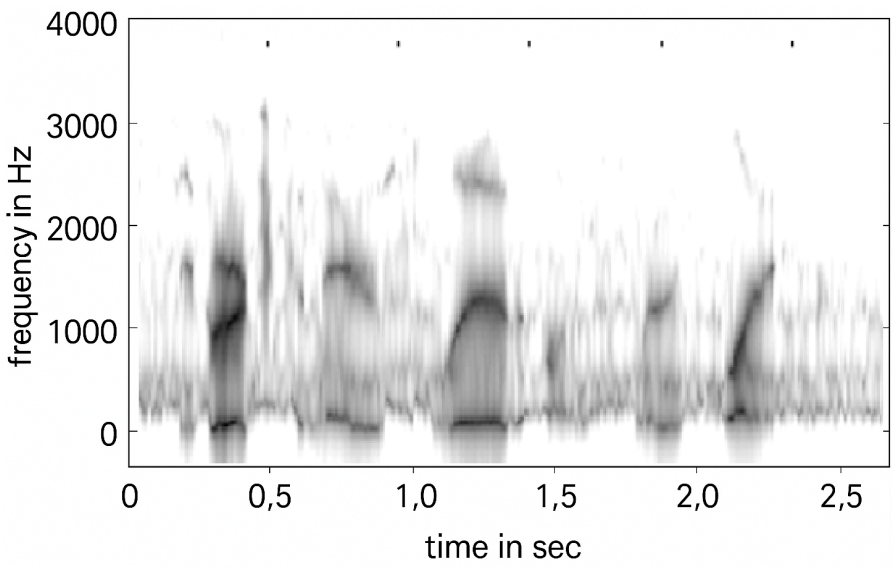



(**b**) Spectrogram of noisy speech signal SP-01with noise level 15 dB.

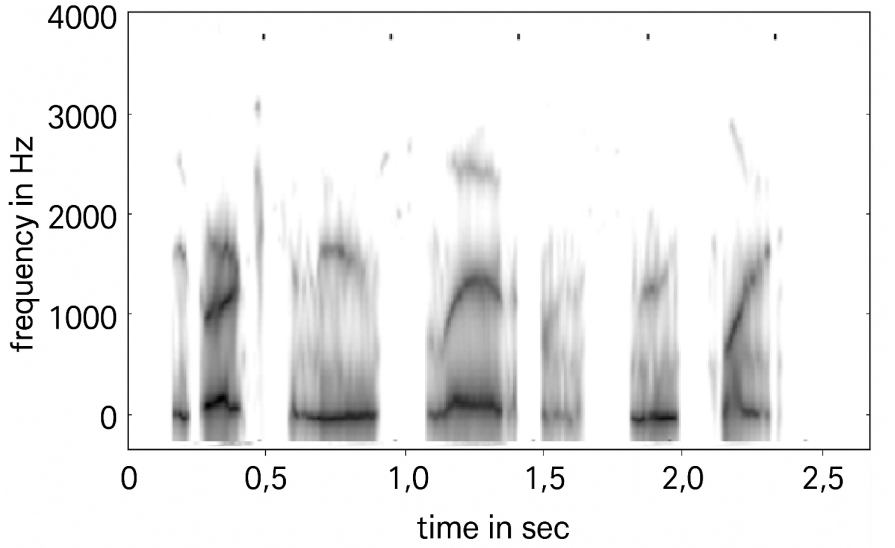



(**c**) Spectrogram of enhanced speech signal SP-01 by Method 1.

**Figure 3.** *Cont.*




(**d**) Spectrogram of enhanced speech signal SP-01 by Method 2.




(**e**) Spectrogram of enhanced speech signal SP-01 by the proposed method.

**Figure 3.** The results of speech signals (spectrograms).

**Figure 4** presents the Articulation Index (AI) values for the different enhancement methods applied to the speech samples. The AI values serve as an objective measure of speech intelligibility, with higher values indicating better clarity and understanding of the speech signal. By comparing the AI scores across Method 1, Method 2, and the proposed method, the figure demonstrates the superior performance of the proposed method in enhancing speech intelligibility, particularly in noisy environments.

The Articulation Index (AI) is calculated using a weighted sum of the Signal-to-Noise Ratios (SNRs) across different frequency bands, considering human auditory perception. The general formula for AI is:

$$AI = \sum_{i=1}^{N} W_i \cdot SNR_i,$$

where:

$N$ is the total number of critical frequency bands,

$W_i$ is the weighting factor for the i$^{th}$ band, representing its importance to speech

intelligibility,

$SNR_i$ is the signal-to-noise ratio in the i[th] band, reflecting how well speech is preserved over noise.

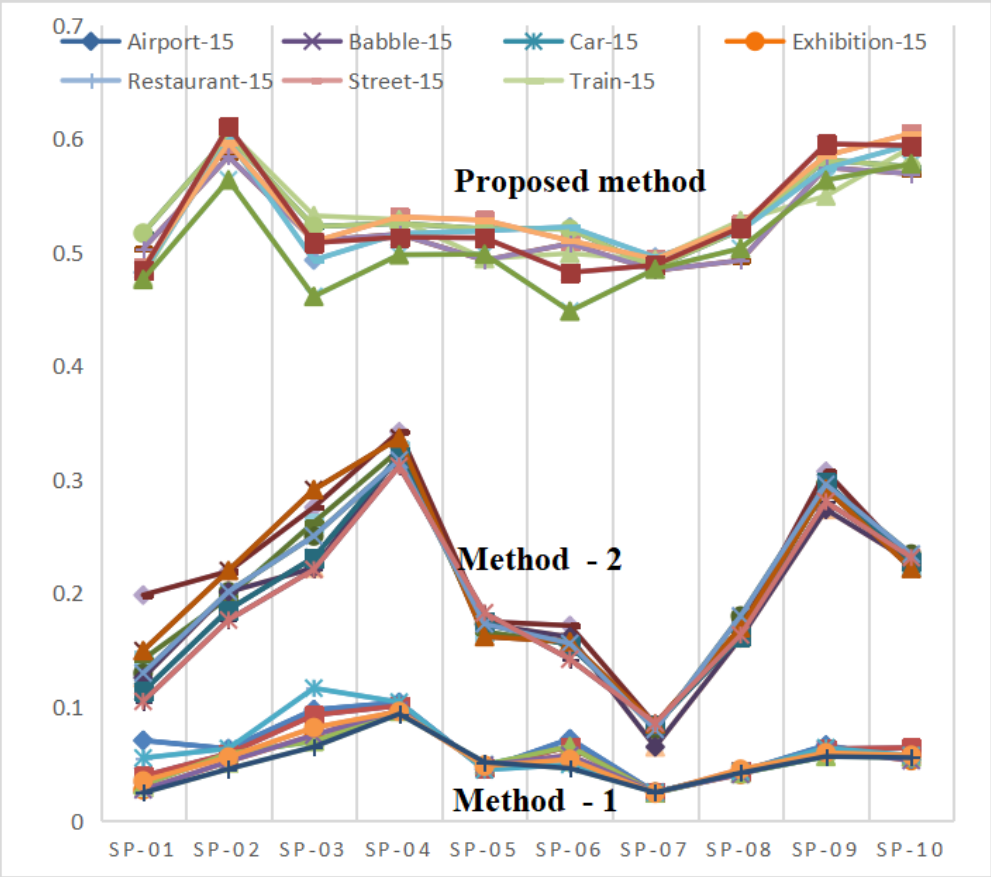



**Figure 4.** Articulation index (AI) values for different methods.

## 4.1. Key parameters in AI calculation

1. **Noise masking thresholds:**

   The masking effect occurs when background noise in a frequency band is strong enough to obscure speech components in that band.

   The speech-to-noise ratio (SNR) in each band is calculated by comparing the power of the speech signal to the noise power.

   If the SNR in a particular band falls below a threshold (typically −15 dB to −20 dB), speech information in that band is considered inaudible, and its contribution to AI is minimal.

2. **Frequency weighting factors:**

   Human hearing is more sensitive to certain frequency ranges, particularly 500 Hz–4,000 Hz, where most speech intelligibility information is concentrated.

   Weighting factors ($W_i$) are assigned to each band based on their contribution to speech perception, with higher weights given to mid-frequency bands and lower weights to very low and high-frequency bands.

   Standard AI models use predefined weight distributions based on psychoacoustic studies.

## 4.2. Final calculation approach

- The AI score is obtained by summing the weighted SNR values across all frequency bands.
- If background noise reduces the SNR in specific bands below the masking threshold, its contribution to AI is reduced or eliminated.
- The final AI value is normalized between 0 (no intelligibility) and 1 (perfect intelligibility). Including this detailed explanation in the manuscript will provide clarity on the AI computation process and justify its use as an evaluation metric for speech enhancement effectiveness.

**Table 1** compares SPL and vibration levels alongside their impact on AI, highlighting how increasing mechanical noise significantly lowers speech intelligibility even when SPL alone is not excessively high.

**Table 1.** Results of noise and vibration scenarios.

| Environment | SPL (dB) | Vibration (mm/s$^2$) | Articulation index (AI) |
|---|---|---|---|
| Low-Vibration, Low-Noise | 50 | 0.3 | 0.85 |
| High-Vibration, Moderate-Noise | 70 | 1.2 | 0.63 |
| High-Vibration, High-Noise | 85 | 2.5 | 0.42 |

## 4.3. Discussion

The findings from this study demonstrate that the proposed wavelet packet-based articulation index enhancement method consistently yields superior speech intelligibility across varied industrial noise conditions, particularly in high-vibration environments. By achieving higher AI scores relative to traditional noise suppression techniques (e.g., baseline wavelet filters and binary mask estimation), the proposed method offers significant advancements for practical deployment in communication-critical industrial scenarios. These results reaffirm earlier observations by Rasetshwane et al. regarding the effectiveness of wavelet-based representations in preserving phonetic clarity under complex noise backgrounds. The dynamic thresholding system presented here adds a novel layer of adaptivity. This feature enables the system to automatically adjust its filtering behaviour in real-time based on transient changes in the acoustic environment, a challenge under-addressed in earlier works such as those by Loizou and Kim [5]. In comparison to fixed-parameter models, this adaptive processing increases resilience against fluctuating vibroacoustic interference, making the system appropriate for volatile industrial surroundings. Beyond intelligibility, this study discovers new territory by establishing a theoretical and practical connection between articulation index optimization and structural health monitoring (SHM). In transportation and industrial surroundings, where revolving or reciprocating machines produce periodic mechanical vibrations, such vibration-induced acoustic energy frequently manifests in the 50–1,000 Hz range—directly overlapping with the lower end of the human speech frequency band (300–3,400 Hz). This overlap reduces speech clarity and can obfuscate key phonemes, particularly consonants, which are critical to intelligibility.

The proposed method addresses this by selectively enhancing frequency bands

with higher articulation importance while attenuating low-frequency noise components common to mechanical operations. As a result, not only is communication improved, but the algorithm effectively discriminates between speech and machine-originating noise, which becomes a foundation for dual-domain diagnostics. From an SHM perspective, this has two profound implications:

Acoustic probing for material integrity: The method enables articulation index analysis to be used as a secondary tool in monitoring changes in structural resonance. By broadcasting controlled speech or modulated tones and analysing articulation loss over time, changes in damping characteristics or the emergence of new resonant frequencies—caused by cracks, fatigue, or material degradation—can be indirectly detected. This complements traditional SHM techniques such as strain gauging and accelerometry, offering a non-invasive acoustic alternative.

Coupled vibroacoustic feedback systems: In environments with real-time audio communication (e.g., maintenance tunnels, engine compartments, heavy manufacturing zones), speech intelligibility data can be continuously monitored alongside vibration spectra. Deviations in the AI baseline, when not attributable to environmental noise, may indicate underlying mechanical anomalies, potentially enabling predictive maintenance via speech-based metrics. For instance, an unbalanced rotor producing harmonic noise patterns could result in consistent AI drops in specific frequency bands.

Thus, by bridging speech processing with structural acoustic behaviour, this research extends the scope of articulation index applications into domains where speech clarity and machinery health are co-dependent variables. In doing so, it lays the groundwork for smart SHM systems where intelligibility loss becomes a signal of underlying mechanical deviation—a novel, cross-disciplinary diagnostic approach.

Future work should investigate machine learning integration for multi-parametric classification (AI trends, spectral shifts, vibration modes), enabling automatic mapping of speech distortion patterns to specific structural faults. Additionally, implementing this system with embedded low-power acoustic sensors and MEMS-based vibration detectors could transform AI-based speech intelligibility algorithms into lightweight, scalable SHM modules.

## 5. Conclusion

The comparative analysis of the three methods under different noise environments reveals clear performance distinctions. The proposed method consistently outperforms Methods 1 and 2 across all environments, achieving higher performance metrics close to 0.6 to 0.65. Method 2, while showing moderate performance, still lags behind the proposed method but outperforms Method 1 across most noise conditions. Method 1, the lowest performer, exhibits significantly lower metrics, with results peaking around 0.15–0.2 in most environments. Overall, the proposed method demonstrates a more robust ability to handle various noisy environments, including Airport, Babble, and Restaurant, maintaining higher accuracy or efficiency across all test cases. This recommends that the projected technique is likely the most appropriate for real-world applications, including noisy data or surroundings, while Method 1 may require noteworthy enhancements to be competitive.

Beyond traditional industrial communication systems applications, the suggested Articulation Index-driven algorithm demonstrates strong potential for deployment in different engineering situations. In industrial surroundings, it can improve voice communication amongst workers in high-noise areas. In public emergency and safety communication systems, where intelligibility is crucial, the technique offers robust performance. Furthermore, integration into smart intercom systems, vehicular voice assistants, and remote operation surroundings can further determine its impact and practicality.

Furthermore, the suggested method fits well with the goals of reducing vibrations and controlling noise, which makes it beneficial beyond just improving speech. Using this algorithm in safety systems and smart buildings can help accomplish two things at once—less background noise and clearer communication. This combination could lead to new ideas and developments in advanced sound technology systems.

The proposed articulation index-driven method has strong potential for real-world deployment in industrial surroundings described by high ambient noise levels. Specifically, the method can be embedded into smart headsets, hearing protection devices, or voice-activated communication systems used in sectors such as manufacturing, aviation, construction, and mining. By integrating this signal enhancement framework with vibration-based microphones or bone conduction sensors, the system can isolate and amplify critical speech components, thereby improving speech intelligibility without compromising safety. Moreover, the algorithm can be implemented on low-power embedded systems, making it suitable for wearable audio electronics. This integration would support more effective human-machine interaction and worker coordination, especially in environments where traditional microphones are compromised by background noise or mechanical interference.

**Author contributions:** Conceptualization, MSP, VV and HJS; methodology, MSP, HJS and DAD; software, MSP, RKS and FS; validation, MSP, RKS and SS; formal analysis, DAD and SS; investigation, MSP, RKS and HJS; resources, MSP and RKS; data curation, RKS, DAD, and FS; writing—original draft preparation, DAD and HJS; writing—review and editing, MSP, RKS and SS; visualization, MSP and FS; supervision, VV and SS; project administration, DAD; All authors have read and agreed to the published version of the manuscript.

**Funding:** This work received no external funding.

**Institutional review board statement:** Not applicable.

**Informed consent statement:** Not applicable.

**Data availability statement:** No new data has been created.

**Conflict of interest:** The authors declare no conflict of interest.

**AI use statement:** The authors declare that no artificial intelligence (AI) tools were used in the preparation of this manuscript.

## References


1. Cooke M, Barker J, Cunningham S, et al. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America. 2006; 120(5): 2421–2424. doi: 10.1121/1.2229005
2. Gannot S, Vincent E, Markovich-Golan S, et al. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2017; 25(4): 692–730. doi: 10.1109/TASLP.2016.2647702
3. Zheng C, Zhang H, Liu W, et al. Sixty Years of Frequency-Domain Monaural Speech Enhancement: From Traditional to Deep Learning Methods. Trends in Hearing. 2023; 27: 23312165231209913. doi: 10.1177/23312165231209913
4. Hu Y, Loizou PC. Evaluation of Objective Quality Measures for Speech Enhancement. IEEE Transactions on Audio, Speech, and Language Processing. 2008; 16(1): 229–238. doi: 10.1109/TASL.2007.911054
5. Loizou PC, Kim G. Reasons why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions. IEEE Transactions on Audio, Speech, and Language Processing. 2011; 19(1): 47–56. doi: 10.1109/TASL.2010.2045180
6. Chen F, Loizou PC. Predicting the Intelligibility of Vocoded Speech. Ear & Hearing. 2011; 32(3): 331–338. doi: 10.1097/AUD.0b013e3181ff3515
7. Doclo S, Kellermann W, Makino S, et al. Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones. IEEE Signal Processing Magazine. 2015; 32(2): 18–30. doi: 10.1109/MSP.2014.2366780
8. Stéphane M. Wavelet Bases. In: A Wavelet Tour of Signal Processing. Elsevier; 2009. pp. 263–376. doi: 10.1016/B978-0-12-374370-1.00011-2
9. Daubechies I. The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory. 1990; 36(5): 961–1005. doi: 10.1109/18.57199
10. Rasetshwane DM, Boston JR, Li CC. Use of the Articulation Index to Design a Wavelet Packet-Based Method for Improving Speech Intelligibility. In: Proceedings of the 2007 15th International Conference on Digital Signal Processing; 1–4 July 2007; Cardiff, UK. pp. 643–646. doi: 10.1109/ICDSP.2007.4288664
11. Donoho DL. De-noising by soft-thresholding. IEEE Transactions on Information Theory. 1995; 41(3): 613–627. doi: 10.1109/18.382009
12. Hao X, Xu C, Zhang C, et al. A neural network approach for speech enhancement and noise-robust bandwidth extension. Computer Speech & Language. 2025; 89: 101709. doi: 10.1016/j.csl.2024.101709
13. Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing. 2001; 9(5): 504–512. doi: 10.1109/89.928915
14. So S, Paliwal KK. Modulation-domain Kalman filtering for single-channel speech enhancement. Speech Communication. 2011; 53(6): 818–829. doi: 10.1016/j.specom.2011.02.001
15. Yuliani AR, Amri MF, Suryawati E, et al. Speech Enhancement Using Deep Learning Methods: A Review. Jurnal Elektronika dan Telekomunikasi. 2021; 21(1): 19. doi: 10.14203/jet.v21.19-26
16. Stevens SS, Rogers MS, Herrnstein RJ. The Apparent Reduction of Loudness: A Repeat Experiment. The Journal of the Acoustical Society of America. 1955; 27(2): 326–328. doi: 10.1121/1.1907523
17. Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1979; 27(2): 113–120. doi: 10.1109/TASSP.1979.1163209
18. Drgas S. A Survey on Low-Latency DNN-Based Speech Enhancement. Sensors. 2023; 23(3): 1380. doi: 10.3390/s23031380
19. Kates JM. The short-time articulation index. Journal of Rehabilitation Research and Development. 1987; 24(4): 271–276. Available online: https://pubmed.ncbi.nlm.nih.gov/3430385/
20. Plomp R. Auditory handicap of hearing impairment and the limited benefit of hearing aids. The Journal of the Acoustical Society of America. 1978; 63(2): 533–549. doi: 10.1121/1.381753
21. Rasetshwane DM, Boston JR, Li CC, et al. Enhancement of speech intelligibility using transients extracted by wavelet packets. In: Proceedings of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics; 18–21 October 2009; New Paltz, NY, USA. pp. 173–176. doi: 10.1109/ASPAA.2009.5346465
22. Johnson MT, Yuan X, Ren Y. Speech signal enhancement through adaptive wavelet thresholding. Speech Communication. 2007; 49(2): 123–133. doi: 10.1016/j.specom.2006.12.002

23. Zhang Y, Nissen SL, Francis AL. Acoustic characteristics of English lexical stress produced by native Mandarin speakers. The Journal of the Acoustical Society of America. 2008; 123(6): 4498–4513. doi: 10.1121/1.2902165
24. Casey M, Rhodes C, Slaney M. Analysis of Minimum Distances in High-Dimensional Musical Spaces. IEEE Transactions on Audio, Speech, and Language Processing. 2008; 16(5): 1015–1028. doi: 10.1109/TASL.2008.925883
25. Roy SK, Paliwal KK. Robustness and sensitivity metrics-based tuning of the augmented Kalman filter for single-channel speech enhancement. Applied Acoustics. 2022; 185: 108355. doi: 10.1016/j.apacoust.2021.108355
26. Healy EW, Yoho SE, Wang Y, et al. An algorithm to improve speech recognition in noise for hearing-impaired listeners. The Journal of the Acoustical Society of America. 2013; 134(4): 3029–3038. doi: 10.1121/1.4820893
27. Narayanan A, Wang D. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; 26–31 May 2013; Vancouver, BC, Canada. pp. 7092–7096. doi: 10.1109/ICASSP.2013.6639038
28. Maganti H, Matassoni M. Bio-Inspired Auditory Processing for Speech Feature Enhancement. In: Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing; 26–29 January 2011; Rome, Italy. pp. 51–58. doi: 10.5220/0003145800510058
29. Wang D, Chen J. Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018; 26(10): 1702–1726. doi: 10.1109/TASLP.2018.2842159
30. Yechuri S, Vanabathina SD. Speech Enhancement: A Review of Different Deep Learning Methods. International Journal of Image and Graphics. 2025; 25(3): 2550024. doi: 10.1142/S021946782550024X
31. Tan K, Zhang X, Wang D. Deep Learning Based Real-Time Speech Enhancement for Dual-Microphone Mobile Phones. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021; 29: 1853–1863. doi: 10.1109/TASLP.2021.3082318
32. Elliott SJ, Nelson PA. Active noise control. IEEE Signal Processing Magazine. 1993; 10(4): 12–35. doi: 10.1109/79.248551
33. Lucero JC, Munhall KG. A model of facial biomechanics for speech production. The Journal of the Acoustical Society of America. 1999; 106(5): 2834–2842. doi: 10.1121/1.428108
34. Fu Y, Wang X. Advancements and trends in vehicle sound package for noise control: A comprehensive review. Advances in Mechanical Engineering. 2025; 17(6): 16878132251345867. doi: 10.1177/16878132251345867
35. Scalart P, Filho JV. 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; 7–10 May 1996; Atlanta, GA, USA. pp. 629–632. doi: 10.1109/ICASSP.1996.543199
36. Cohen I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Transactions on Speech and Audio Processing. 2003; 11(5): 466–475. doi: 10.1109/TSA.2003.811544
37. Guðnason J, Fang G, Brookes M. Epoch-Based Spectrum Estimation for Speech. In: Proceedings of the Interspeech 2023; 20–24 August 2023; Dublin, Ireland. pp. 4274–4278. doi: 10.21437/Interspeech.2023-407
38. Shankar N, Bhat GS, Panahi IMS. Real-Time Single-Channel Deep Neural Network-Based Speech Enhancement on Edge Devices. In: Proceedings of the Interspeech 2020; 25 October 2020; Shanghai, China. pp. 3281–3285. doi: 10.21437/Interspeech.2020-1901
39. Lu X, Chen H, He X. A Frequency Domain Fitting Algorithm Method for Automotive Suspension Structure under Colored Noise. World Electric Vehicle Journal. 2024; 15(9): 410. doi: 10.3390/wevj15090410
40. Hou J, Yi H, Xiang X, et al. Identification of vehicle suspension shock absorber rattle noise based on wavelet packet feature fusion and GWO-LSTM. Sound & Vibration. 2025; 59(2): 1941. doi: 10.59400/sv1941
41. Green T, Hilkhuysen G, Huckvale M, et al. Speech recognition with a hearing-aid processing scheme combining beamforming with mask-informed speech enhancement. Trends in Hearing. 2022; 26: 23312165211068629. doi: 10.1177/23312165211068629
42. Ganapathy S, Thomas S, Hermansky H. Modulation frequency features for phoneme recognition in noisy speech. The Journal of the Acoustical Society of America. 2009; 125(1): EL8–EL12. doi: 10.1121/1.3040022
43. Allen JB, Rabiner LR. A unified approach to short-time Fourier analysis and synthesis. Proceedings of the IEEE. 1977; 65(11): 1558–1564. doi: 10.1109/PROC.1977.10770
44. Gupta P, Patil HA, Guido RC. Vulnerability issues in Automatic Speaker Verification (ASV) systems. EURASIP Journal on Audio, Speech, and Music Processing. 2024; 2024(1): 10. doi: 10.1186/s13636-024-00328-8