Proceedings of Machine Learning Research vol 278:1–10, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

Speech Enhancement for Headphones Based on Spectrum Subtraction and DMA

Jiaqi Yu 1612906844@QQ.COM
Harbin Engineering University, Harbin 150006, China
**Corresponding author*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

For headphone speech enhancement, traditional mono noise reduction methods, such as Wiener filter, have limited noise reduction effects and large computational costs, the directivity of the differential microphone array (DMA) is used to optimize the voice pickup and improve the signal-to-noise ratio of the input signal. Experimental simulation results show that the proposed method improves the output signal-to-noise ratio by 6dB under different output signal-to-noise ratios, and still has a 2dB speech enhancement effect under the condition of low signal-to-noise ratio of -5dB, and has strong anti-interference ability. The effect is stable in the background of broadband noise and low-frequency noise, which can better optimize the user's experience during call.

Keywords: Speech Enhancement, Spectrum Subtraction, Voice Activity Detection, Differential Microphone Array.

1. Introduction

TWS earphones have been favored by more users because of their convenience, but according to the survey, the noise reduction performance of the existing TWS earphones, which is the main concern of the consumer, cannot meet the market demand (Hu, 2022).

The problem of speech enhancement in earphone communication constitutes a specialized branch of microphone speech enhancement, which can be categorized into monophonic noise reduction and multichannel noise reduction approaches. In the domain of monophonic noise reduction, several well-established algorithms have been extensively studied, including spectral subtraction (Boll, 1979), Wiener filtering (Mourad et al., 2012), and the minimum mean-square error (MMSE) estimator based on short-term spectral amplitude (Ephraim and Malah, 1985). Notably, Tu and Li (2024) proposed a microphone array speech enhancement algorithm grounded in spectral amplitude estimation, implementing dual enhancement strategies through MMSE estimation and maximum a posteriori (MAP) estimation of spectral amplitudes. Yang (2024) integrated spectral subtraction with neural networks to advance speech enhancement research. Wang et al. (2013) incorporated spectral subtraction with VAD for speech denoising. However, these methods rely heavily on accurate noise signal estimation, and their employed noise detection mechanisms exhibit limited effectiveness. Traditional monaural noise reduction methods (e.g., Wiener filtering, MMSE) require accurate estimation of power spectral density or statistical parameters of signal and noise. In low input SNR scenarios, such estimates become unreliable, leading to significant performance degradation. Wiener filtering also exhibits limited robustness against nonlinear distortions (e.g., impulsive noise). Spatial filtering enhances input SNR through multi-channel processing, while voice activity detection (VAD) improves noise tracking accuracy, offering viable enhancements to conventional approaches.

Microphone arrays can be categorized into additive arrays and differential arrays. Additive arrays exhibit significant frequency-dependent beam pattern variations, resulting in potential directional deviations during low-frequency signal processing that may compromise system performance. In contrast, differential microphone arrays (DMA) employ compact microphone spacing to achieve ultra-directional characteristics with frequency-independent beam patterns (Yan et al., 2022). This intrinsic property grants differential arrays a distinct advantage in processing wideband fluctuating signals, as their spatial selectivity remains stable across frequency spectra. The algorithmic implementation of differential arrays thus demonstrates superior robustness for voice-related applications requiring precise spatial filtering under broadband conditions. Building upon these foundational studies, this paper proposes a speech enhancement algorithm based on a differential array and spectral subtraction. The DMA enhances signal-to-noise ratio (SNR) via directional pickup, while voice activity detection (VAD) refines noise estimation efficacy, the proposed method achieves enhanced performance compared to the approach documented in (Wang et al., 2013). Figure 1 shows the algorithm flowchart.

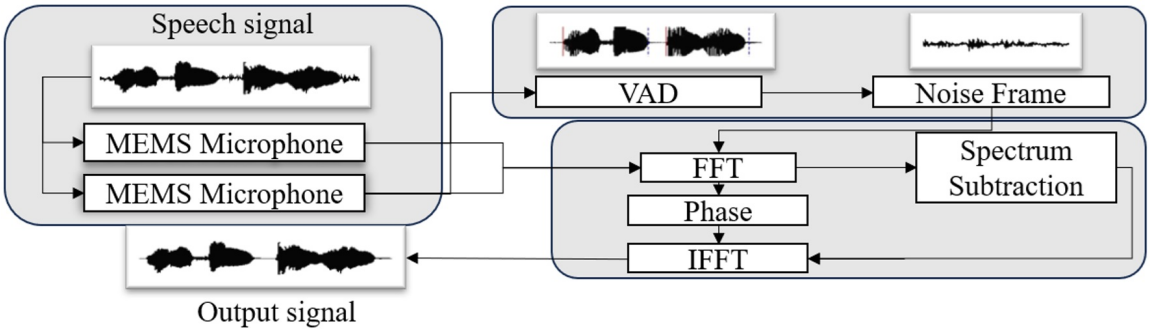

Figure 1: Algorithm flowchart.

2. Spectral subtraction and improvements

2.1. Spectral subtraction

Spectral subtraction, a mainstream monophonic noise reduction algorithm, demonstrates strong applicability in TWS earphones due to its low computational complexity and effective performance. This short-time spectral estimation method requires pre-emphasis, framing, and windowing (Guo and Chen, 2021) operations. The algorithm operates by subtracting the estimated noise power spectrum (typically derived from noise estimation) from the mixed signal's power spectrum to obtain the enhanced speech spectrum. The implementation framework comprises three fundamental steps:

First, a mixed speech signal containing additive noise $z(i)$ is established as (Wang, 2023):

$$z(i) = s(i) + n(i) \tag{1}$$

The clean speech signal $s(i)$ and estimated noise component $n(i)$ are derived from the mixed signal. As a short-time spectral estimation algorithm, spectral subtraction requires preprocessing steps including framing, windowing, and discrete Fourier transform (DFT) to decompose the signal into spectral magnitude and phase components:

$$Z_\omega(\omega) = S_\omega(\omega) + N_\omega(\omega) \tag{2}$$

By squarering both sides of the equation at the same time, the power spectrum of different signals can be obtained.

$$|Z_\omega(\omega)|^2 = |S_\omega(\omega)|^2 + |N_\omega(\omega)|^2 + S_\omega(\omega)N_\omega(\omega)^* + N_\omega(\omega)S_\omega(\omega)^* \tag{3}$$

$|Z_\omega(\omega)|^2$, $|S_\omega(\omega)|^2$, $|N_\omega(\omega)|^2$ is the power spectrum of $z(i)$, $s(i)$ and $n(i)$, $S_\omega(\omega)N_\omega(\omega)^*$ and $N_\omega(\omega)S_\omega(\omega)^*$ equals 0 because the additive noise signal does not change with the change of the speech signal. The power spectrum is a non-negative value, so when the noise power spectrum is larger than the mixed voice power spectrum, the value of the pure speech power spectrum is set to 0.

$$\begin{cases} |S_\omega(\omega)|^2 = |Z_\omega(\omega)|^2 + |N_\omega(\omega)|^2, |Z_\omega(\omega)|^2 \geq |N_\omega(\omega)|^2 \\ |S_\omega(\omega)|^2 = 0, |Z_\omega(\omega)|^2 < |N_\omega(\omega)|^2 \end{cases} \tag{4}$$

The proposed algorithm achieves monophonic speech enhancement through precise noise estimation, a critical technical component detailed in prior sections. A voice activity detection (VAD) algorithm is subsequently introduced to isolate noise-only segments from speech-containing components within mixed signals. This dual-stage approach ensures robust noise suppression while preserving speech integrity, particularly effective in dynamic acoustic environments.

2.2. Spectral Entropy-Based VAD

VAD algorithms are broadly categorized into time-domain and frequency-domain approaches. While time-domain methods (e.g., short-term energy (Moattar and Homayounpour, 2009) and zero-crossing rate (Li, 2022)) exhibit lower computational complexity, frequency-domain approaches generally achieve superior performance (Adiga and Bhandarkar, 2016). Among frequency-domain methods, spectral entropy-based VAD demonstrates enhanced accuracy by leveraging spectral characteristics.

Implementation Workflow:

1. Signal Segmentation: Input signals are partitioned into overlapping frames through framing and windowing.

2. Feature Extraction: For each frame, the normalized spectral probability density function is computed via Fourier transform:

For each frame, the normalized spectral probability density function is computed via Fourier transform (Huang et al., 2023):

$$p_m = \frac{p_m(k)}{\sum_{k=0}^{N} p_m(k)}, k = 0, 1, \ldots, N \tag{5}$$

Where N is the number of the frequency point, $P_m(k)$ is the power in each frequency point, so spectral entropy for the $m-th$ frame is derived as:

$$H_m = -\sum_{k=0}^{N} p_m \log p_m \tag{6}$$

3. Threshold Decision: A noise threshold $\mathrm{TH} = \mathrm{k} \cdot \mathrm{H}_{\mathrm{ave}}$ is established using the average spectral entropy over M noise-only frames.

$$H_{\mathrm{ave}} = \frac{H_m}{\sum_{m=1}^{M} H_m} \tag{7}$$

Voice activity is detected by comparing frame-wise entropy Hm with TH:

$H_m \geq$ TH: Indicates speech onset.

$H_m <$ TH: Marks speech offset if preceded by detected speech; otherwise, confirms noise dominance.

This method dynamically discriminates speech segments from noise through entropy-driven threshold adaptation.

2.3. VAD anti-misjudgment mechanism

This study proposes an enhanced VAD algorithm to address frequent false speech-onset detection caused by impulsive noise and dynamic ambient noise during user mobility in practice. The improved methodology incorporates a multi-frame temporal validation mechanism: Upon preliminary speech-onset tagging, the algorithm sequentially analyzes A subsequent consecutive frames. Provisional markers are revoked and reclassified as impulsive noise if threshold criteria persistently remain unsatisfied within this temporal window, where A is adaptively tuned based on environmental characteristics.

To reconcile real-time processing constraints with enhanced noise estimation robustness, a dynamic noise statistics update strategy is implemented: When noise-dominant frames consecutively exceed a predefined duration threshold M frames, where $M > A$, the system autonomously refreshes noise estimation. This dual-mechanism architecture achieves optimal balance between detection accuracy and computational efficiency. Figure 2 illustrates the steps for implementation.

Spectral Entropy-Based Voice Activity Detection and Spectral Subtraction Noise Reduction

Algorithm 1 Spectral Entropy-Based VAD with Spectral Subtraction Denoising

Input: Speech signal $x[n]$
Output: Segregated speech/noise frames, denoised signal
Compute average spectral entropy $\mathrm{H_{avg}}$ from first M frames as threshold
for each frame i starting from $M+1$ **do**
    Calculate spectral entropy H_i of current frame
    if $\mathrm{H}_i > \mathrm{H_{avg}}$ **then**
        Proceed to validate subsequent A frames
        if any frame among A has $\mathrm{H} < \mathrm{H_{avg}}$ **then**
            Label as noise frame
        else
            Label as speech frame
        end if
    else
        Label as noise frame
    end if
end for
if detected noise frames $\geq M$ **then**
    Apply spectral subtraction: compute average noise power spectrum
    Subtract noise spectrum from speech signal to output denoised signal
end if

Figure 2: Implementation of adaptive spectral subtraction denoising algorithm with VAD.

3. Spatial Filtering Algorithm Based on Differential Microphone Arrays

Considering the advantageous operational principles of DMA, this study employs a directivity-based filtering algorithm utilizing a first-order DMA for spatial speech enhancement. The methodology aims to regulate signal acquisition through directional selectivity, employing a dual omnidirectional microphone array to emulate MEMS microphone configurations typical in earphone designs (schematic illustrated below).

Adopting a far-field model (Chen, 2022), the system implements a linear array with inter-microphone spacing for spatial filtering. Let the frequency-domain representations $X(\omega)$. A pair of omnidirectional microphones are linearly arranged symmetrically about the origin with an inter-element spacing distance d. The workflow is shown in Figure 3. To ensure the array maintains directional characteristics across the broad frequency range of 300-8000 Hz, the spacing parameter d must satisfy the spatial sampling criterion $\lambda_{\min}/2$, where $\lambda_{\min}$ corresponds to the wavelength at the maximum frequency of 8 kHz. This constraint prevents the occurrence of spatial aliasing lobes (grating lobes) in the beam pattern. Considering both the dimensional requirements of typical TWS earphone designs and the far-field assumption validity conditions, the inter-microphone distance is empirically set to 1 cm.

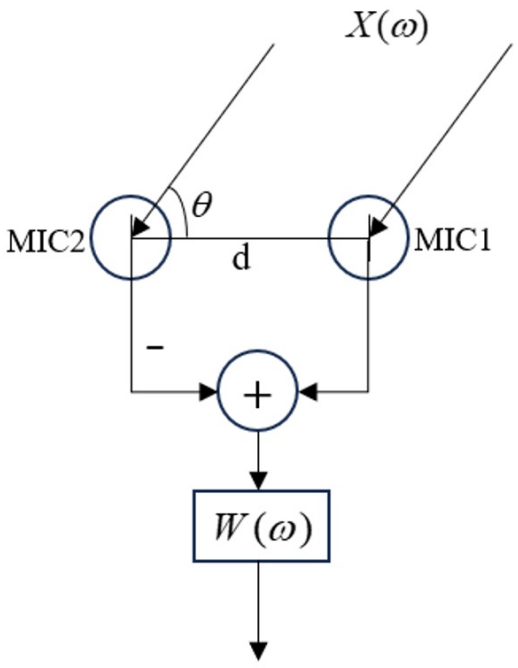

Figure 3: Spatial filtering workflow.

When the incident angle θ is 0, the acoustic wavefront propagates parallel to the axis of the microphone array. Under this condition, the sound wave arrives at MIC 1 earlier than MIC 2 due to the spatial separation between the two microphones. The resulting path difference Δd induces a phase shift $S(\omega, \theta)$ between the signals received by the two microphones, which can be derived

from the relationship:

$$S(\omega,\theta) = \begin{bmatrix} 1 \\ e^{-j\omega\frac{d\cos\theta}{c}} \end{bmatrix} \tag{8}$$

The signal output is denoted as $S(\omega,\theta)X(j\omega)$, and the directional sensitivity of the system to different spatial orientations can be modulated by controlling the weighting function $W(\theta)$ at the output stage. The weighted output is formulated as:

$$D(\omega,\theta) = W(\omega)S(\omega,\theta)X(j\omega) \tag{9}$$

$W(\omega) = [W_1(\omega) \quad W_2(\omega)]$ In this context, $W_1(\omega)$ and $W_2(\omega)$ represent the weighting functions assigned to MIC 1 and MIC 2, respectively. To derive the expression for $W(\omega)$, two constraints must be satisfied. The constraints adopted in this study are as follows:

1. Maximum gain condition: The array achieves maximum sensitivity when the signal arrives parallel to the linear array axis ($\theta = 0^{\circ}$).

2.Minimum gain condition: The array attains minimum sensitivity (null steering) when the signal arrives perpendicular to the array axis ($\theta = 90^{\circ}$).

$$\begin{bmatrix} S(\omega,0) \\ S(\omega,90) \end{bmatrix} W(\omega) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{10}$$

$W(\omega)$ is formulated as:

$$W(\omega) = \frac{1}{1-e^{-j\omega d/c}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{11}$$

Under the condition that the ratio d/c is sufficiently small (where d denotes the inter-sensor spacing and c represents the speed of sound), the original formulation can be simplified to:

$$W(\omega) = \frac{1}{\frac{\omega d}{c}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{12}$$

The beam pattern is formulated as:

$$B(\omega,\theta) = \frac{j}{-\frac{d}{c}\omega}\left[1 - e^{-\frac{j\omega d}{c}(\cos\theta)}\right] = \cos\theta \tag{13}$$

The differential directivity pattern of the array in polar coordinates is illustrated as follows. The differential directivity pattern of the array in polar coordinates demonstrates superior acoustic signal acquisition performance along the horizontal axis (0° and 180° azimuthal angles), while exhibiting pronounced null characteristics at orthogonal orientations (90° and 270° azimuthal angles), with significantly attenuated acoustic responses in the vertical plane. Figure 4 shows the directivity pattern.

4. Experimental Validation

The verification framework was implemented on MATLAB with frame length=100, frame shift=50, lead-in silence=0.5s, M=10, and A=5 (noise segments shorter than lead-in duration were excluded from estimation). A single omnidirectional microphone and a dual-MEMS array were deployed

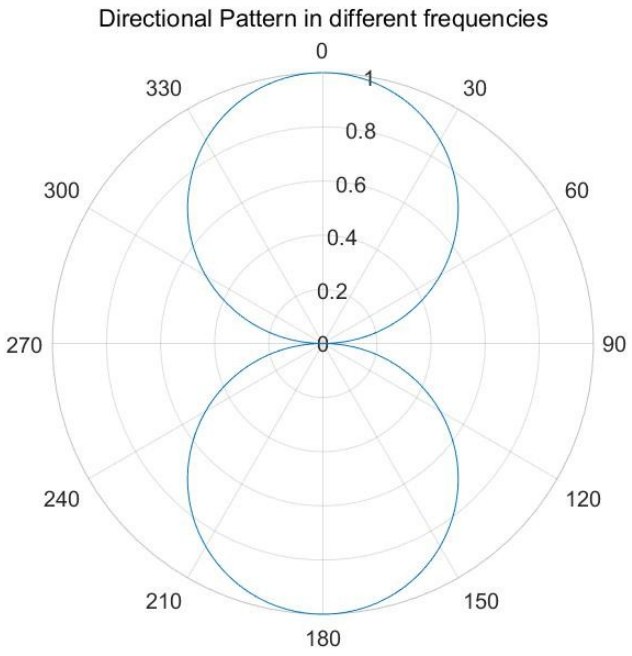

Figure 4: Directivity pattern of a dual-microphone array in polar coordinates.

at the origin, surrounded by 360 isotropic noise sources spaced at 1° intervals. Mixed signals (speech+noise) were emitted from ±45° azimuths, while pure noise sources occupied remaining angles, simulating a mouth-oriented beamforming scenario.

The speech signal was acquired from a cleanly articulated Mandarin utterance ("Tàiyáng cóng dōngfāng shēngqǐ") recorded in an anechoic chamber under controlled acoustic conditions, while the noise signal includes marketplace ambient noise as broadband noise and wind-induced turbulence noise as low-frequency noise.

To quantitatively evaluate the denoising performance, the proposed method was validated under controlled input SNR conditions. The contaminated mixture $y(i)$ was synthesized by linearly combining Speech signals $s(i)$ and noise segments $n(i)$ with varying input SNR levels, defined as:

$$y(i) = s(i) + an(i), SNR_{in} = 10\lg\left(\frac{\sum|s(i)|^2}{\sum|n(i)|^2}\right)(dB) \tag{14}$$

By comparing the enhanced signal $y(i)$ with the clean speech $s(i)$, the output SNR (SNRout-SNRout) is defined as

$$SNR_{\text{out}} = 10\lg\left(\frac{\sum|s(i)|^2}{\sum|y(i)-s(i)|^2}\right)(dB) \tag{15}$$

Input signals with signal-to-noise ratios (SNRs) of -5 dB, -3 dB, 0 dB, 3 dB, and 5 dB were processed, and the output SNR values and corresponding performance curves are tabulated and plotted in Figure 5.

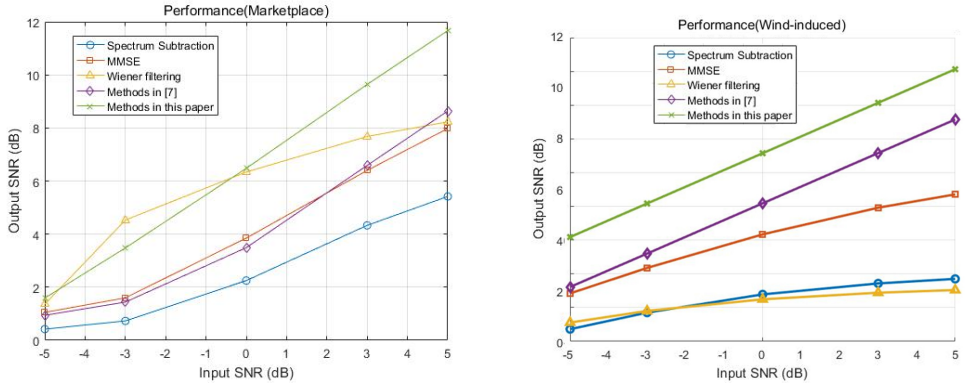

Figure 5: Output signal-to-noise ratio (SNR) curves for different denoising algorithms.

The output SNR values and corresponding performance curves are tabulated and plotted in Figure 6.

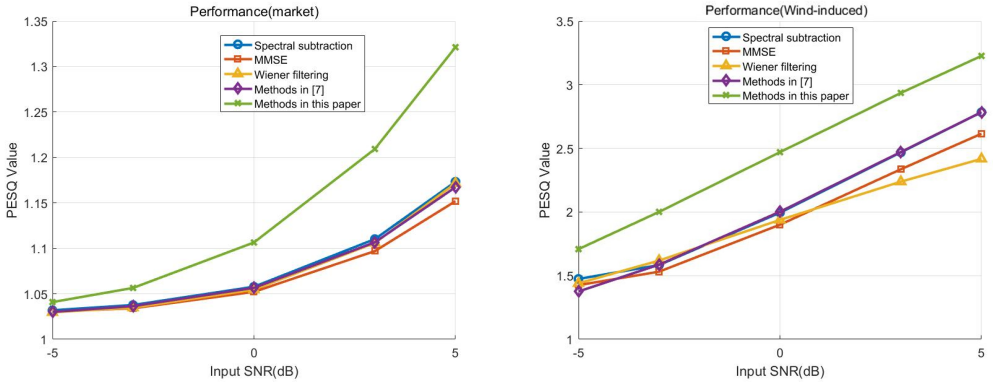

Figure 6: PESQ curves for different denoising algorithms.

5. Conclusion

This study addresses insufficient noise reduction in earphone communication by integrating spectral entropy-based VAD and spatial filtering with enhanced spectral subtraction. Experimental results demonstrate a maximum 6 dB improvement over classical spectral subtraction under broadband and low-frequency noise, and a 3 dB gain compared to other methods, and the experimental results based on the PESQ metric demonstrate that the proposed methodology consistently exhibits superior performance compared to other existing approaches, while maintaining low computational complexity makes it suitable for resource-constrained devices like TWS Bluetooth earphones. The proposed algorithm proves effective for call enhancement in high-noise environments. Future work may incorporate dynamic threshold adaptation to further optimize noise suppression performance.

References

M. Tejus Adiga and Rekha Bhandarkar. Improving single frequency filtering based voice activity detection (vad) using spectral subtraction based noise cancellation. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 18–23, 2016. doi: 10.1109/SCOPES.2016.7955823.

S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979. doi: 10.1109/TASSP.1979.1163209.

Yuan Chen. Research and implementation of speech enhancement algorithm of circular microphone array. Master's thesis, Nanjing University of Posts and Telecommunications, 2022.

Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2): 443–445, 1985. doi: 10.1109/TASSP.1985.1164550.

Lili Guo and Yonghong Chen. An improved spectral subtractive speech enhancement algorithm. *Communications Technology*, 54(6):1350–1355, 6 2021.

Liang Hu. Research on tws earphone user reviews based on text mining. Master's thesis, Zhongnan University of Economics and Law, Wuhan, 2022.

Baolin Huang, Zhi Liu, and Wei Lin. Voice activity detection algorithm in radiocommunication. *Audio Engineering*, 47(09):104–107, 2023. ISSN 1002-8684. doi: 10.16311/j.audioe.2023.09.031.

Jiajing Li. Research on remote broadcast signal monitoring based on short-time energy and short-time zero-crossing rate. *Audio Engineering*, 46(11):100–102, 2022. ISSN 1002-8684. doi: 10.16311/j.audioe.2022.11.026.

M. H. Moattar and M. M. Homayounpour. A simple but efficient real-time voice activity detection algorithm. In *2009 17th European Signal Processing Conference*, pages 2549–2553, 2009.

Talbi Mourad, Lotfi Salhi, Ben nasr Mohamed, and Adnane Cherif. Wiener filtering application in the bionic wavelet domain for speech enhancement. *International Journal of Advancements in Computing Technology*, 4:146–160, 02 2012. doi: 10.4156/ijact.vol4.issue2.19.

Jingxian Tu and Lianfen Li. Microphone array speech enhancement algorithm based on spectral amplitude estimation. *Journal of Hengyang Normal University*, 45(03):68–74, 2024. ISSN 1673-0313. doi: 10.13914/j.cnki.cn43-1453/z.2024.03.011.

Ping Wang, Jixiang Lu, Suihuai Yu, and Changde Lu. Improved spectral subtractive speech enhancement algorithm in cloud terminal voice interaction. *Computer Integrated Manufacturing Systems*, 19(07):1721–1725, 2013. ISSN 1006-5911. doi: 10.13196/j.cims.2013.07.283.wangp.026.

Xuan Wang. Research on the voice quality enhancement technology of underwater intercoms. Master's thesis, Harbin Engineering University, 2023.

Longfei Yan, Weilong Huang, W. Bastiaan Kleijn, and Thushara D. Abhayapala. Phase error analysis for first-order linear differential microphone arrays. In *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5, 2022. doi: 10.1109/IWAENC53105.2022.9914748.

Tao Yang. Research on speech enhancement technology based on machine learning. *Audio Engineering*, 48(3):39–41, 3 2024.