Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 20, No. 3, 2026

Research on Intelligent Coding Optimization and Speech Denoising & Enhancement Methods for Audio Processing

Shuxuan Li*

School of Computer Science and Technology, Shandong University of Technology, Zibo, CO 255000, China
* **Corresponding author**: Shuxuan Li (Email: lisxand@163.com)

Abstract: To address the trade-off between audio quality and file size as well as the demand for enhancing noisy speech intelligibility, this paper proposes a full-process audio processing scheme integrating intelligent analysis, adaptive coding optimization, speech denoising and quantitative evaluation. First, 34 music and speech samples with diverse coding parameters are used to construct the Quality-Size Index (QSI), and combined with a random forest regression model, bit rate and compression algorithm are identified as the core factors affecting coding performance. Second, a random forest classifier based on audio time-frequency features achieves 100% accuracy in music-speech classification, enabling adaptive coding optimization via matching optimal parameters for different audio types. Finally, an optimized spectral subtraction-Wiener filtering cascade denoising algorithm is designed for non-reference noisy speech. Experimental results show that the adaptive coding scheme effectively balances audio quality and file size; the proposed denoising algorithm improves the SNR of two speech samples by 6.22 dB and 5.51 dB respectively, significantly reduces spectral flatness, eliminates musical noise and enhances speech intelligibility. This scheme provides an intelligent technical solution for audio coding and speech denoising with high practical application value.

Keywords: Intelligent Audio Analysis, Adaptive Coding, Random Forest, Spectral Subtraction, Wiener Filtering, Speech Denoising.

1. Introduction

With the rapid development of digital audio technology, audio signals have been increasingly applied in music storage, voice communication, streaming media transmission and other scenarios [1]. In audio coding, how to minimize file size on the premise of ensuring sound quality and achieve the optimal trade-off between sound quality and file size is one of the core issues in the field of audio processing [2]. Meanwhile, in the process of voice acquisition and transmission, the introduction of environmental noise severely degrades speech intelligibility [3]. How to achieve efficient denoising without clean reference speech is also a key technical difficulty in speech signal processing [4].

Most existing audio coding optimization methods adopt fixed-parameter coding without targeted optimization for the acoustic differences between music and speech, making it difficult to realize the optimal balance between sound quality and file size. Traditional speech denoising algorithms such as spectral subtraction tend to introduce musical noise [5], while Wiener filtering relies on accurate noise estimation [6], a single algorithm can hardly balance noise suppression and speech fidelity.

To address the above problems, this paper constructs an integrated audio processing technical system. On one hand, machine learning is used to realize intelligent audio type recognition and adaptive optimization of coding parameters. On the other hand, a cascade denoising algorithm is adopted to achieve efficient enhancement of noisy speech. In addition, objective indicators are designed to complete quantitative evaluation of the full-process performance, providing a full-link solution for audio processing.

2. Methods

This study proposes a closed-loop intelligent audio processing system integrating four core modules: intelligent audio analysis, adaptive coding optimization, speech denoising and enhancement, and quantitative effect evaluation. The system uses machine learning for audio classification and coding parameter optimization, a spectral-domain cascade filtering algorithm for denoising, and objective metrics for performance evaluation, forming a complete technical framework.

2.1. Intelligent Audio Analysis and Adaptive Coding Optimization

Thirty-four music and speech audio samples were used. By constructing a comprehensive evaluation indicator and combining machine learning models, coding parameter importance analysis, audio type classification, and adaptive optimal parameter matching were completed through four sequential steps.

2.1.1. Construction of Audio Features and Evaluation Indicators

Original uncompressed 48kHz/24bit music and speech audio are used as benchmark samples. A total of 34 groups of compressed audio samples with different coding formats (WAV/MP3/AAC) and technical parameters (sampling rate, bit depth, bit rate) are collected, and the coding attributes and physical features of each sample are extracted. The Peak Signal-to-Noise Ratio (PSNR) is adopted to quantify the sound quality loss of compressed audio relative to the original audio. This indicator is calculated based on the mean square error between the original and compressed audio, and a higher value indicates smaller sound quality distortion after audio

compression. To comprehensively measure the trade-off performance between sound quality and file size, the Quality-Size Index (QSI) is constructed, defined as the ratio of the Peak Signal-to-Noise Ratio to the file size, so as to realize the normalized evaluation of sound quality and file size. The formula is given as follows:

$$\mathrm{QSI} = \frac{\mathrm{PSNR(dB)}}{\mathrm{V(KB)}} \tag{1}$$

where V represents the file size of the compressed audio, in kilobytes (KB).

2.1.2. Importance Analysis of Coding Parameters

A random forest regression model (300 decision trees) was built, taking sampling rate, bit depth, bit rate (continuous features) and compression algorithm (categorical, one-hot encoded) as inputs, and QSI as output. Feature splitting contribution quantified parameter importance, providing a basis for optimal parameter screening.

2.1.3. Intelligent Music/Speech Classification

A random forest classification model (200 decision trees, max depth 10) was constructed using five acoustic features (zero-crossing rate, spectral centroid, spectral bandwidth, root-mean-square energy, onset strength). Features were standardized, and the dataset was split 8:2 by stratified sampling, achieving precise audio type recognition.

2.1.4. Adaptive Optimal Coding Parameter Matching

Based on the Quality-Size Index (QSI), music and speech audio samples with different coding parameters are sorted in descending order respectively to screen the optimal coding parameter combinations for each audio type, establishing a mapping relationship between audio types and optimal coding parameters. The mapping covers core parameters such as optimal coding format, sampling rate, bit depth (only for WAV format), and bit rate (for MP3/AAC format) corresponding to each audio type.

For unknown audio, the trained music/speech classification model first identifies its type, and then automatically matches the corresponding optimal coding parameters according to the established mapping relationship, realizing adaptive recommendation of audio coding parameters. This ensures that the file size is minimized on the premise of meeting sound quality requirements.

2.2. Speech Denoising Enhancement and Quantitative Effect Evaluation

For two sets of noisy speech signals without clean reference speech, an optimized spectral subtraction–Wiener filtering cascade denoising algorithm is designed to achieve efficient suppression of background noise. Meanwhile, an objective evaluation indicator system is constructed to quantitatively analyze the denoising effect from two dimensions: noise suppression capability and speech clarity. The technical process consists of three sequential steps: automatic noise segment extraction, implementation of the cascade denoising algorithm, and objective quantitative evaluation of denoising performance. All links are closely connected to realize integrated processing of noisy speech from noise extraction and denoising to effect evaluation.

2.2.1. Automatic Noise Segment Extraction

Energy-based Voice Activity Detection (VAD) is adopted to automatically identify silent segments in noisy speech as pure noise reference samples without manual annotation. By calculating the short-time energy of the audio signal and setting an energy threshold of 20 dB, audio segments with energy below the threshold are determined as silent segments. This signal segment is extracted as a noise reference sample, which lays a foundation for noise power spectrum estimation in the subsequent spectral-domain denoising algorithm and ensures the objectivity and accuracy of noise estimation.

2.2.2. Optimized Spectral Subtraction–Wiener Filtering Cascade Denoising Algorithm

A two-stage processing structure of rough denoising followed by fine denoising is adopted. First, preliminary background noise suppression is realized using optimized spectral subtraction. Then, based on the output of spectral subtraction, fine denoising is completed via Wiener filtering. This effectively solves the problem of musical noise easily generated by traditional spectral subtraction while preserving the integrity of speech details.

Both noisy speech and noise reference samples are converted between the time domain and frequency domain using Short-Time Fourier Transform (STFT) [7] and Inverse Short-Time Fourier Transform (ISTFT). The frame length is set to 512 points, the frame shift to 256 points, and a Hanning window is used to reduce spectral leakage, ensuring consistency in spectral-domain processing.

1.Optimized spectral subtraction

STFT is performed on noisy speech and noise reference samples separately to obtain spectral amplitude and phase information. The average amplitude spectrum of the noise reference sample is calculated and converted into a power spectrum, and the power spectrum of noisy speech is optimized using temporal smoothing with a window size of 11 frames. A power-law spectral subtraction strategy is applied: the power spectra of noisy speech and noise are raised to the second power before subtraction, with a subtraction coefficient of 2.0, and the power spectrum is restored via inverse power transformation. A minimum amplitude limit coefficient of 0.001 is set to avoid negative power spectrum values. Secondary smoothing is performed on the subtracted power spectrum to further suppress musical noise. Finally, the phase information of the original noisy speech is retained, and preliminary denoised speech is reconstructed through ISTFT.

2.Wiener filtering

The preliminary denoised speech output by optimized spectral subtraction is taken as the estimated clean speech, and the difference between noisy speech and the estimated clean speech as the estimated noise. STFT is applied to both signals to compute their power spectra. A Wiener filtering gain function is constructed as the ratio of the noisy speech power spectrum to the sum of the noisy speech power spectrum and the weighted noise power spectrum. The noise weighting coefficient is set to 0.5 and the bias term to 0.01. The gain function is used to weight and modify the spectrum of noisy speech for fine denoising. Finally, the modified spectrum is converted back to a time-domain signal via ISTFT to obtain the final denoised speech.

2.2.3. Objective Quantitative Evaluation of Denoising Effects

Under the condition of no clean reference speech, two objective evaluation indicators—SNR improvement and Spectral Flatness Measure (SFM)—are established to quantitatively analyze the denoising effect in terms of noise suppression capability and speech clarity. The STFT

parameters for indicator calculation are kept consistent with those of the denoising algorithm to ensure the objectivity and consistency of the evaluation results.

1.SNR Improvement

Calculated based on the energy variation of speech signals before and after denoising, it is defined as 10 times the logarithm of the ratio of the energy of the original noisy speech to the energy of the denoised speech. A higher value indicates a more significant noise suppression effect. The formula is as follows:

$$\mathrm{SNR}_{\mathrm{imp}} = 10\log_{10}\left(\frac{\mathrm{E}_{\mathrm{noisy}}}{\mathrm{E}_{\mathrm{denoised}}}\right) \quad (2)$$

where $\mathrm{E}_{\mathrm{noisy}}$ represents the energy of the original noisy speech, and $\mathrm{E}_{\mathrm{denoised}}$ represents the energy of the denoised speech.

2.Spectral Flatness Measure (SFM)

Reflects the smoothness of the audio spectrum. It is calculated as 10 times the logarithm of the ratio of the geometric mean to the arithmetic mean of the spectrum of each frame. A smaller spectral flatness indicates a more significant peak-valley difference in the speech spectrum, less musical noise, and higher speech clarity. After performing STFT on the audio, the spectral flatness is calculated frame by frame, and the final result is the average value of the spectral flatness across all frames.

3. Experimental Results and Analysis

In this study, simulation experiments were conducted to verify the full-process technical scheme of intelligent audio analysis and adaptive coding optimization, speech denoising and enhancement, and quantitative effect evaluation. The experimental data include 34 groups of music and speech audio with different coding parameters, as well as 2 groups of noisy speech samples. Each experimental result corresponds to the technical method one-to-one. By combining quantitative analysis and feature analysis, the numerical characteristics, variation trends and technical significance of the experimental results in each link are clarified, which fully verifies the effectiveness, stability and practicability of the intelligent audio signal processing scheme proposed in this study.

3.1. Experimental Results of Intelligent Audio Analysis and Adaptive Coding Optimization

3.1.1. Results of Coding Parameter Importance Analysis

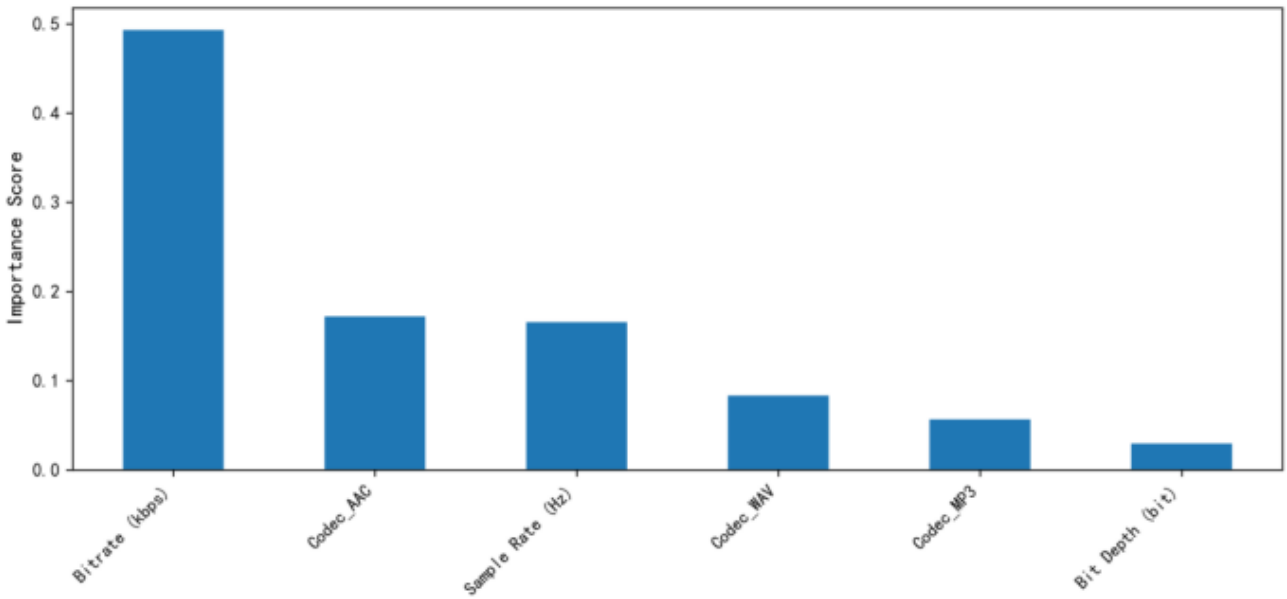

Figure 1. Audio coding parameter importance

As shown in Figure 1, the random forest regression model quantifies the influence of different coding parameters on the audio Quality-Size Index (QSI), revealing significant differences in the priority of each parameter. Ranked in descending order of importance score, the parameters are bit rate, compression algorithm, sampling rate, and bit depth. Among them, bit rate achieves a much higher importance score than other parameters, making it the core factor affecting audio quality-size efficiency; compression algorithm and sampling rate are secondary influencing factors; while bit depth has the lowest impact and can be weakened in coding optimization. These results clarify the core direction of audio coding optimization, that is, prioritizing the optimization of bit rate and compression algorithm can effectively control sound quality loss while minimizing file size, providing a scientific quantitative basis for the screening and optimization of coding parameters.

3.1.2. Performance Results of Music/Speech Intelligent Classification Model

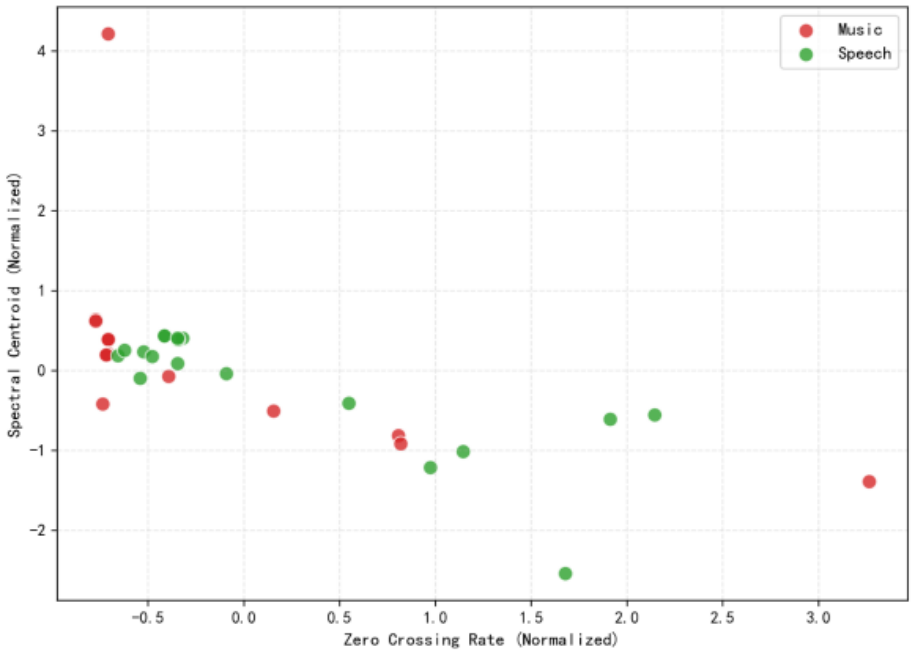

Figure 2. Music vs speech feature distribution

As shown in Figure 2, the two types of audio exhibit obvious separability in two core acoustic features: normalized zero-crossing rate and normalized spectral centroid. The feature points of music and speech samples form relatively independent clustering regions in the feature space, with no significant overlap. This result verifies that the extracted acoustic features can effectively capture the essential acoustic differences between music and speech, laying a solid feature foundation for the subsequent random forest-based intelligent music/speech classification model and ensuring the recognition accuracy of the classification model.

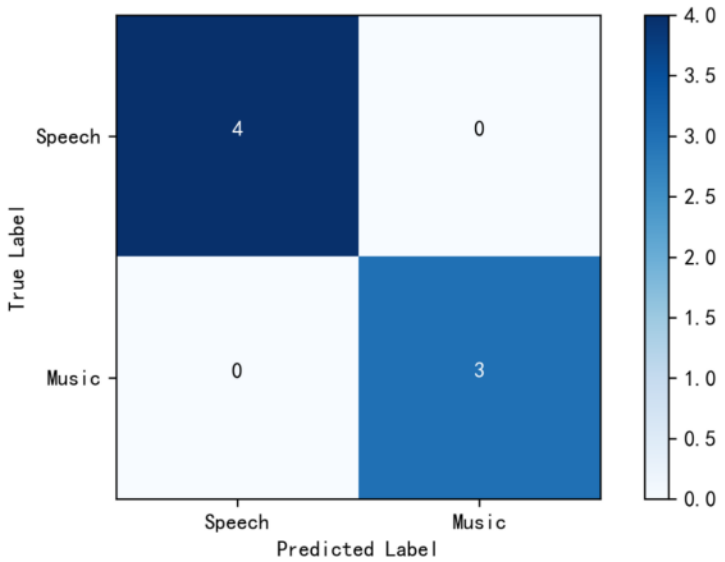

Figure 3. Confusion matrix of the music/speech classification model

As shown in Figure 3, the constructed random forest classification model achieves 100% classification accuracy on the test set: all 4 speech samples are correctly identified as speech with no misclassification, and all 3 music samples are correctly recognized as music without any false judgment. This result fully verifies the excellent classification performance of the model, proving that the random forest model trained on the extracted acoustic features can accurately distinguish between music and speech audio, fully meeting the requirements of intelligent audio type recognition and providing reliable classification support for the subsequent adaptive coding parameter matching.

3.1.3. Results of Adaptive Optimal Coding Parameter Matching

Table 1. Optimal Coding Parameters and Performance Metrics for Speech Audio

Type	Compression Algorithm	Sampling Rate (Hz)	Bit Rate (kbps)	PSNR (dB)	QSI
Speech	AAC	44100	96	15.99	0.48191
Speech	AAC	44100	128	15.99	0.37342
Speech	MP3	16000	64	17.44	0.36661

Table 2. Optimal Coding Parameters and Performance Metrics for Music Audio

Type	Compression Algorithm	Sampling Rate (Hz)	Bit Depth (bit)	Bit Rate (kbps)	PSNR (dB)	QSI
Music	MP3	44100	None	64	9.48	0.120595
Music	WAV	48000	16	None	92.64	0.098811
Music	WAV	48000	24	None	100	0.071109

Significant differences exist in the optimal coding parameter combinations for music and speech audio, as screened by QSI ranking. As shown in Table 1 and Table 2, the optimal coding format for music audio is mainly high-bit-rate MP3/AAC with a sampling rate of 48 kHz, which ensures the frequency-domain integrity and sound quality of music. For speech audio, medium-low bit-rate MP3/AAC with a reduced sampling rate of 16 kHz can be adopted, which greatly reduces file size while meeting speech clarity requirements.

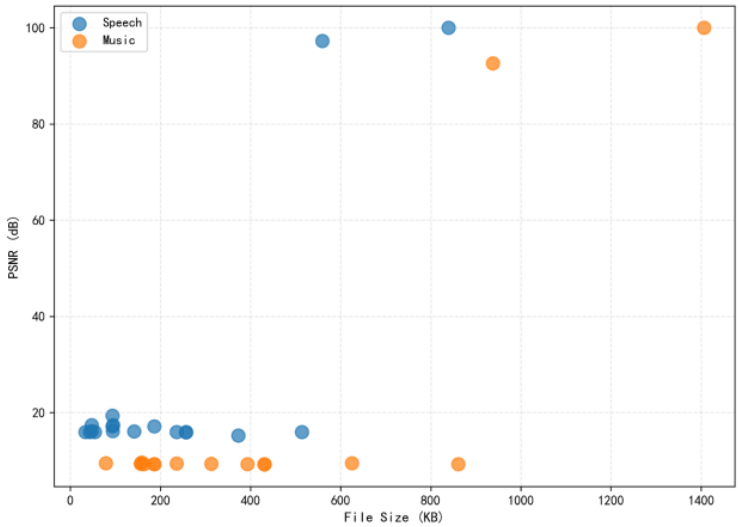

Figure 4. Distribution of PSNR and file size for music and speech samples

As illustrated in Figure 4, the quality-size trade-off scatter plot intuitively presents the distribution relationship between PSNR and file size for different audio types and coding parameters. Music audio generally exhibits the characteristics of high PSNR and large file size, while speech audio shows low PSNR and small file size. Moreover, the optimal parameter combinations of both audio types are distributed in the Pareto [8] optimal region of "high PSNR and small size", verifying the rationality and effectiveness of screening optimal coding parameters based on QSI.

3.2. Experimental Results of Speech Denoising Enhancement and Quantitative Effect Evaluation

3.2.1. Visualization Results of Cascade Denoising Algorithm

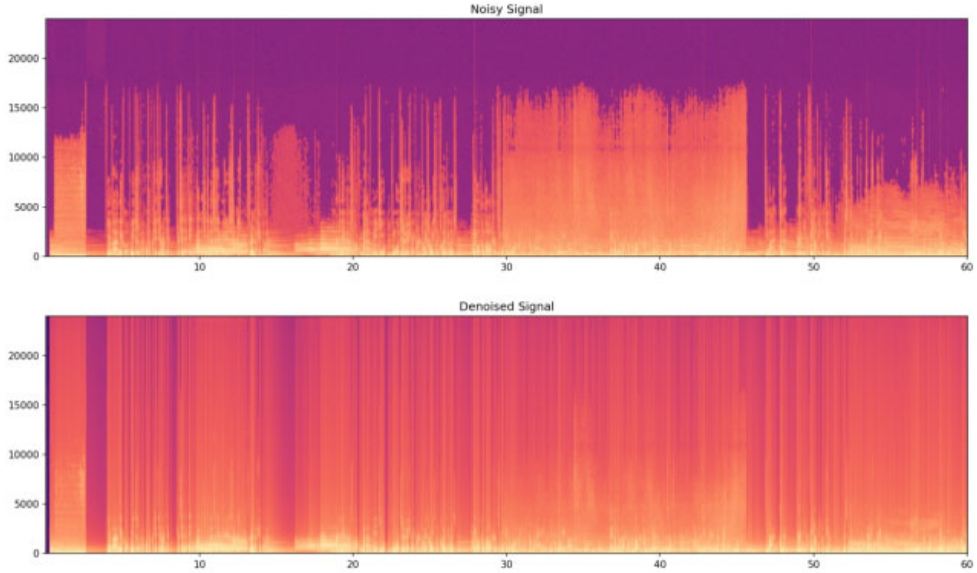

Figure 5. Noisy spectrogram and enhanced spectrogram of part1.wav

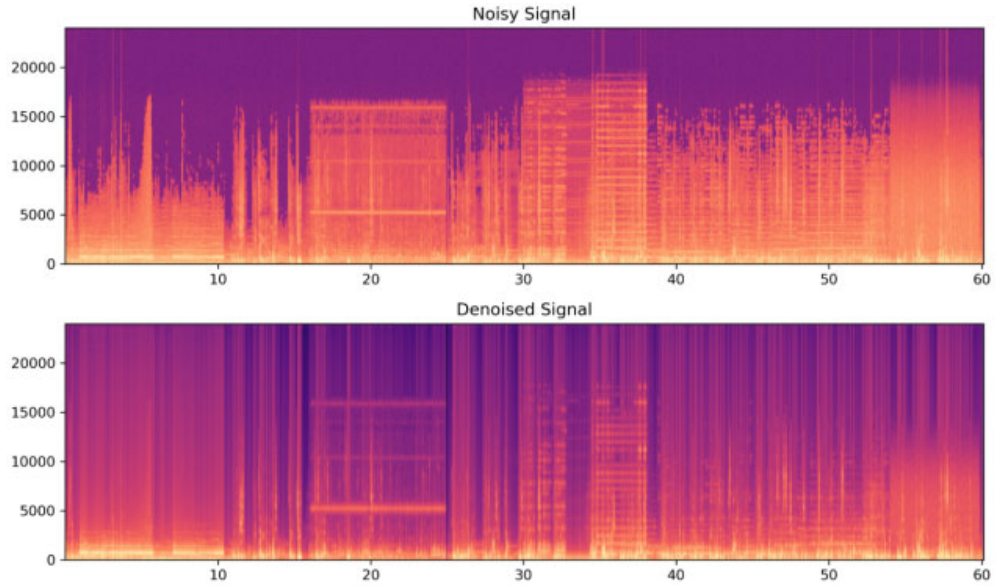

Figure 6. Noisy spectrogram and enhanced spectrogram of part2.wav

As can be seen from Figure 5 and Figure 6, which compare the spectrograms of two groups of noisy speech and denoised speech, the spectrograms of the original noisy speech contain obvious background noise spectra. The noise covers the entire frequency domain with disordered distribution, which partially masks the frequency-domain features of speech. After processing by the optimized spectral subtraction–Wiener filtering cascade algorithm, the background noise spectra in the denoised speech spectrograms are significantly suppressed, and the high-frequency and low-frequency noise components are almost completely eliminated. The core frequency-domain features of speech, such as formants, are clearly visible, without obvious musical noise being generated. These results intuitively demonstrate that the proposed cascade denoising algorithm can effectively suppress background noise while retaining the original frequency-domain features of speech to the maximum extent and avoiding speech distortion, thus solving the technical problem that traditional spectral subtraction tends to generate musical noise.

3.2.2. Results of Objective Quantitative Indicators for Denoising Effects

Denoising processing was performed on two groups of noisy speech samples with different noise characteristics, and the quantitative analysis of the denoising effect was completed using two core indicators: SNR improvement and Spectral Flatness Measure (SFM). The experimental results are shown in Table 3.

Table 3. Quantitative Evaluation Results of Speech Noise Reduction Effect

Speech	Original Noisy Speech SFM (dB)	Denoised Speech SFM (dB)	SNR Improvement (dB)
Part1.wav	-21.73	-22.87	6.22
Part2.wav	-18.15	-20.99	5.51

As can be seen from Table 3, after processing by the proposed optimized spectral subtraction–Wiener filtering cascade denoising algorithm, both speech samples achieved significant SNR improvement: the SNR of Part1.wav increased by 6.22 dB, and that of Part2.wav increased by 5.51 dB, with the SNR improvement of both samples exceeding 5 dB. This indicates that the algorithm can effectively suppress background noise with different characteristics, significantly enhance the SNR of speech signals, and exhibits excellent adaptability to various noisy speech environments.

In terms of spectral flatness, the original noisy speech samples had high SFM values: -21.73 dB for Part1.wav and -18.15 dB for Part2.wav. After denoising, the SFM values of both samples decreased significantly, dropping to -22.87 dB and -20.99 dB respectively, with a reduction of more than 1 dB for both. This demonstrates that the cascade algorithm effectively suppresses musical noise, reduces the spectral flatness of the denoised speech, highlights the frequency-domain characteristics of speech, and significantly improves speech intelligibility.

The quantitative results in Table 3 are highly consistent with the spectrogram visualization analysis, jointly verifying the denoising performance of the proposed cascade algorithm from the two dimensions of numerical quantification and graphical intuition, and achieving the dual goals of effective noise suppression and significant improvement of speech intelligibility.

4. Discussion

4.1. Analysis of Method Advantages

Compared with traditional methods, the proposed scheme has obvious technical advantages. In coding optimization, the QSI index and random forest are used to realize quantitative evaluation of coding parameters and adaptive coding for music and speech, overcoming the limitation of fixed-parameter coding. In speech denoising, the cascade structure of optimized spectral subtraction and Wiener filtering is designed, which suppresses musical noise and improves the accuracy of noise estimation, achieving effective denoising without clean reference signals. In evaluation, the proposed objective indicators can be directly used for real noisy speech without clean references, showing strong practicability.

4.2. Research Limitations and Future Prospects

This study still has several limitations. First, the

experimental dataset is small, including only 34 groups of audio and 2 groups of noisy speech. Future work can expand the data scale to verify the generalization ability. Second, only music and speech are considered, and the scheme can be extended to more audio types. Third, the denoising algorithm uses empirical parameters, which can be optimized adaptively by intelligent algorithms in the future.

5. Conclusion

In response to the practical demands of audio coding optimization and speech denoising, this paper constructs a complete intelligent audio signal processing system and verifies the effectiveness of the scheme through experiments. The main conclusions are as follows: Based on the QSI indicator and random forest regression model, bit rate and compression algorithm are identified as the core influencing parameters of audio coding, providing a quantitative basis for coding optimization. The random forest classification model based on acoustic features enables high-precision classification between music and speech with an accuracy of 100%, offering technical support for adaptive coding. The optimized cascaded denoising algorithm combining spectral subtraction and Wiener filtering can effectively suppress background noise in noisy speech, improve the signal-to-noise ratio by 5–7 dB, significantly reduce spectral flatness, eliminate musical noise, and enhance speech clarity. The full-process technical scheme proposed in this paper realizes the integrated processing of intelligent audio analysis, coding optimization, speech denoising and effect evaluation, which exhibits favorable practical application value and provides an efficient and intelligent technical solution for the field of audio processing.

Acknowledgment

I sincerely thank my family and friends for their constant support, encouragement and care throughout this journey.

References

[1] Peng M.The application of digital media technology in the post-production of film and television animation[J].Media and Communication Research,2024,5(2).

[2] M. K L,Dai P,Rachel B, et al.Assessing the Efficiency and Quality of Audio-Coding Versus Transcript Coding[J].Nursing Research,2026.

[3] Liu L,Liang L,Huang K, et al.Spatial Distribution and Influencing Factors of Speech Intelligibility in Round-Table Conversation Scenarios[J].Buildings,2026,16(6):1258-1258.

[4] Dionelis N,Brookes M.Modulation-Domain Kalman Filtering for Monaural Blind Speech Denoising and Dereverberation. [J]. IEEE/ACM Trans. Audio, Speech & Language Processing, 2019, 27(4): 799-814.

[5] Shenghuan Z,Ye C.Masking and noise reduction processing of music signals in reverberant music [J]. Journal of Intelligent Systems, 2022, 31(1):420-427.

[6] Nahavandi M A.Noise segmentation for improving performance of Wiener filter method in spectral reflectance estimation[J].Color Research & Application,2018,43(3):341-348.

[7] Lee M,Kim M,Lee J, et al.Real-time battery safety diagnosis via Siamese convolutional neural network combined with online passive electrochemical impedance spectroscopy pattern extraction based on driving data and short-time Fourier transform[J].Journal of Energy Storage,2026,154(PC):121274-121274.

[8] L. J A,A. L K.Effect of the use of music on definitional knowledge in an introductory statistics course: Evidence from a Pareto chart activity[J].Decision Sciences Journal of Innovative Education,2021,19(4):265-274.