mathematics

MDPI

Article

Desynchronization Resilient Audio Watermarking Based on Adaptive Energy Modulation

Weinan Zhu [1], Yanxia Zhou [2], Deyang Wu [3], Gejian Zhao [1], Zhicheng Dong [4], Jingyu Ye [3,5] and Hanzhou Wu [1,6,*]

[1] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China; zwn@shu.edu.cn (W.Z.); 23820171@shu.edu.cn (G.Z.)
[2] College of Information Science and Technology, Xizang University, Lhasa 850012, China; zyx9208@utibet.edu.cn
[3] College of Computer Science and Engineering, Guangxi Normal University, Guilin 541001, China; wdyang@gxnu.edu.cn (D.W.); newzolo@gmail.com (J.Y.)
[4] College of Information Science and Technology, Tibet University, Lhasa 850000, China; dongzc@utibet.edu.cn
[5] Digital Guangxi Group, Guilin 541001, China
[6] School of Big Data and Computer Science, Guizhou Normal University, Guiyang 550025, China
* Correspondence: hanzhou@shu.edu.cn

Abstract

With the rapid proliferation of social media platforms and user-generated content, audio data is frequently shared, remixed, and redistributed online. This raises urgent needs for copyright protection and traceability to safeguard the integrity and ownership of such content. Resilience to desynchronization attacks remains a significant challenge in audio watermarking. Most existing techniques face a trade-off between embedding capacity, robustness, and imperceptibility, making it difficult to meet all three requirements effectively in real-world applications. To address this issue, we propose an improved patchwork-based audio watermarking algorithm. Each audio frame is divided into two non-overlapping segments, from which mid-frequency energy features are extracted and modulated for watermark embedding. A linearly decreasing buffer compensation mechanism balances imperceptibility and robustness. Additionally, an optimization algorithm is incorporated to enhance watermark transparency while maintaining resistance to desynchronization attacks. During watermark extraction, each bit of the watermark is recovered by analyzing the intra-frame energy relationships. Furthermore, we provide a theoretical analysis demonstrating that the proposed method is robust against various types of attack. Extensive experimental results demonstrate that the proposed scheme ensures high audio quality, strong robustness against desynchronization attacks, and a higher embedding capacity than existing methods.

Keywords: desynchronization attack; patchwork; linearly decreasing buffer compensation

MSC: 94A60; 94A08; 68M25

check for updates

Academic Editors: Yi-Hui Chen and Ya-Fen Chang

Received: 31 July 2025
Revised: 13 August 2025
Accepted: 22 August 2025
Published: 26 August 2025

Citation: Zhu, W.; Zhou, Y.; Wu, D.; Zhao, G.; Dong, Z.; Ye, J.; Wu, H. Desynchronization Resilient Audio Watermarking Based on Adaptive Energy Modulation. *Mathematics* **2025**, *13*, 2736. https://doi.org/10.3390/math13172736

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the era of social networks and online media platforms, audio content is widely disseminated through short videos, live streaming, and podcasts. Protecting the copyright of such content and enabling reliable traceability has become a critical requirement for ensuring data security and rightful ownership in user-generated media. In recent years, the rapid development of internet technology has significantly facilitated the dissemination of

digital content, offering unprecedented access to information. However, this convenience has also heightened concerns over copyright protection, particularly for traditional media such as audio. The widespread adoption of large language models (LLMs) has further compounded these challenges, driving increased demand for content traceability and the advancement of specialized watermarking techniques [1]. Within this context, audio watermarking remains a crucial research focus, offering effective solutions for copyright verification and content tracking through the imperceptible embedding of information into audio signals. Nonetheless, the growing accessibility of audio editing tools has made operations such as playback speed adjustment and timbre modification increasingly common, which can severely impair the accuracy of watermark extraction. As a result, enhancing the robustness of audio watermarking against time-scale modification (TSM) and pitch-scale modification (PSM) has become a pressing challenge in current research.

Embedding and extracting audio watermarks present greater challenges compared to their image counterparts, primarily due to the heightened sensitivity of the human auditory system (HAS) relative to the human visual system (HVS). The HVS exhibits a degree of insensitivity to local variations in image details, allowing watermarks to be imperceptibly embedded in high-frequency regions. In contrast, the HAS is acutely sensitive to minute changes in amplitude, frequency, and phase, where even slight distortions can degrade perceptual audio quality. Furthermore, image watermarks are embedded in a two-dimensional space characterized by high spatial redundancy, a property widely exploited by various transform-domain techniques such as those based on DWT, SVD, and chaotic maps [2–4]. In contrast, audio signals are one-dimensional temporal sequences with lower redundancy and strong temporal-frequency correlations [5].

Over the past decade, numerous audio watermarking algorithms have been developed, which can be broadly categorized into time-domain and frequency-domain approaches [6,7]. Time-domain techniques, such as amplitude modulation, histogram-based methods, and echo hiding, embed watermarks by directly modifying the temporal characteristics of the audio signal [8–10]. In contrast, frequency-domain methods employ transformations like the discrete Fourier transform (DFT) [11,12], discrete cosine transform (DCT) [13,14], and discrete wavelet transform (DWT) [15–17] to embed watermark information by adjusting specific transform coefficients. Furthermore, auxiliary techniques including spread spectrum [18–20], patchwork schemes [21–23], quantized index modulation (QIM) [24–26], and singular value decomposition (SVD) [27–29] are frequently integrated to improve the robustness of watermarking systems. These methods aim to embed watermark data into the audio signal in an imperceptible way, while ensuring reliable extraction under various conditions.

Desynchronization attacks have long posed a significant challenge in audio watermarking by disrupting the alignment between the embedded watermark and the host signal, which greatly complicates accurate extraction. To address this issue, a variety of synchronization strategies have been explored, including exhaustive search, explicit synchronization, implicit synchronization, and invariant watermarking [30–32]. Exhaustive search techniques are straightforward in principle but often demand high computational cost, making them unsuitable for real-time scenarios. Explicit synchronization involves inserting additional markers or sequences into the audio signal to guide alignment during extraction. Although effective, this method typically reduces payload capacity and becomes vulnerable to signal processing operations that distort or remove the markers. In contrast, implicit synchronization determines alignment points by analyzing inherent audio features, such as amplitude fluctuations, energy variations, or changes in signal envelope. These features act as intrinsic landmarks for alignment, enabling robust synchronization without additional markers.

Several representative methods have emerged from these strategies. The technique proposed in ref. [33] employs a patchwork-based method, embedding watermark bits in the DCT domain and synchronization codes in the logarithmic DCT (LDCT) domain. Temporal distortions are compensated for by estimating scaling factors from the LDCT domain, leading to strong robustness against PSM, TSM, and jitter. However, the increased complexity of LDCT processing and the separation between synchronization and watermarking domains limit both efficiency and embedding capacity. To improve upon local-feature-based schemes, Jiang et al. [15] introduce a global-characteristic-based method that applies adaptive frame segmentation and embeds watermark data using statistical features from the wavelet domain. Indirect synchronization is achieved through a logical frame index, yielding solid robustness to 30% TSM and 1/10 jitter, though the method remains less effective against amplitude variations. Another noteworthy contribution is the audio-lossless robust watermarking (ALRW) framework in ref. [34], which embeds watermark information into the histogram structure of the low-frequency DWT subband. The distortion introduced during embedding is fully reversible in the absence of attacks, ensuring lossless recovery. This method offers excellent resistance to desynchronization attacks such as cropping, TSM, and jitter, yet its limited embedding capacity restricts broader applicability. A further improvement is presented in ref. [35], which combines explicit synchronization with invariant feature embedding. Watermark and synchronization codes are embedded into frequency-domain logarithmic mean (FLDM) features using a patchwork structure. This enables frame realignment when temporal distortions occur. Despite improved robustness to misalignment, the method offers limited performance under amplitude scaling and suffers from constrained capacity. Recently, some research has focused on constructing watermarking schemes in the time domain to resist desynchronization attacks. Among them, the scheme proposed by Zhao et al. [36] based on the Time Domain Fragment Energy Relationship is a representative work. This method achieves effective resistance against TSM attacks by modifying the energy ratio between adjacent audio fragments. However, directly modifying full-band signal samples in the time domain poses a significant challenge to maintaining the imperceptibility of the watermark.

Despite significant progress in audio watermarking, achieving an effective balance among robustness, imperceptibility, and embedding capacity remains a persistent challenge. Many existing methods perform poorly under desynchronization attacks such as jitter, TSM, and PSM, or lack stability when subjected to common signal processing operations. To address these limitations, this paper proposes a novel audio watermarking algorithm that combines robustness with efficiency. The method first segments the audio into frames and equally divides each frame into two parts. Each part is processed using DCT, and mid-frequency coefficients are selected for embedding. Watermark bits are embedded based on the energy comparison between the two segments. To further improve resistance to TSM and PSM while maintaining audio quality, an optimization strategy and a linearly decreasing buffer compensation mechanism are introduced to reduce distortion caused by energy adjustment. During extraction, the decoder reverses the embedding procedure by analyzing the inter-segment energy relationships to accurately recover the watermark. Theoretical analysis demonstrates the method's reliability against a variety of attacks, and extensive experiments confirm its ability to maintain a favorable trade-off among audio fidelity, embedding rate, and robustness, showing strong practical applicability and adaptability.

The main contributions of this paper are as follows:

- We propose a linearly decreasing buffer compensation mechanism that enhances the robustness of audio watermarking against desynchronization attacks, such as TSM and PSM, while reducing the distortion caused by energy adjustment.

- We theoretically analyze the resilience of the proposed method against both common signal processing and desynchronization attacks, and experimentally validate its effectiveness through comprehensive evaluations.
- Experimental results show that the proposed method maintains stable performance under common signal processing operations, with the highest BER being only 4.01%, and exhibits enhanced robustness against desynchronization attacks such as jitter, TSM, and PSM, with the BER never exceeding 12.31% under all tested conditions.

2. Proposed Methods

Figure 1 shows the general framework of the proposed method, which mainly consists of two modules that are energy feature extraction and watermark embedding. In the following, we provide the details and give an analysis in terms of parameter optimization and robustness.

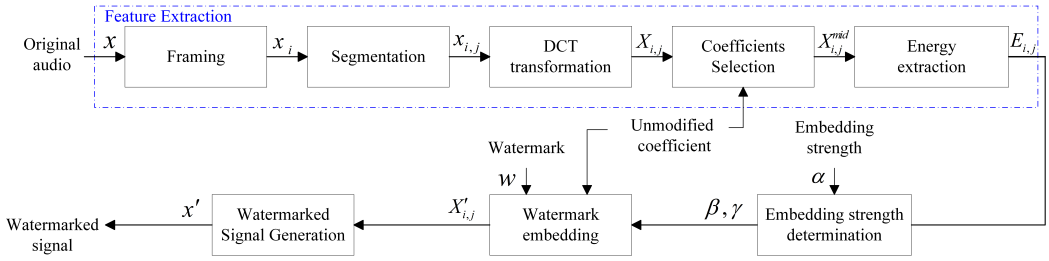


Figure 1. Sketch for the proposed method which consists of two modules, i.e., energy feature extraction and watermark embedding.

2.1. Energy Feature Extraction

Given the original audio signal $\mathbf{x} = \{x_1, x_2, \ldots, x_L\}$, where L is the total number of samples, we adopt a patchwork-based segmentation strategy to divide $\mathbf{x}$ into N non-overlapping frames of equal length M such that $L = NM$. For simplicity, we assume that $M \mid L$; otherwise, the remaining samples at the end are discarded. Each frame can be denoted as $\mathbf{x}_i = \{x_{(i-1)M+1}, x_{(i-1)M+2}, \ldots, x_{iM}\}$, where $i \in [1, N]$. Each frame $\mathbf{x}_i$ is further divided into two disjoint segments of equal length, i.e., $\mathbf{x}_i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}]$, where $\mathbf{x}_{i,j} = \{x_{(i-1)M+(j-1)S+1}, x_{(i-1)M+(j-1)S+2}, \ldots, x_{(i-1)M+jS}\}$, $j \in \{1, 2\}$, and $S = M/2$. Here we also assume $S \mid M$. Thereafter, each segment $\mathbf{x}_{i,j}$ is transformed using the DCT to obtain its frequency representation $\mathbf{X}_{i,j} = \{X_{i,j}(1), X_{i,j}(2), \ldots, X_{i,j}(S)\}$.

Owing to the energy compaction property of the DCT, most signal energy is concentrated in the low-frequency components. Modifying these components may degrade the perceived audio quality, whereas high-frequency components are more susceptible to signal processing operations such as compression, filtering, and noise, which may lead to the loss of embedded watermark information. To achieve a balance between robustness and perceived audio quality, the watermark is embedded into the mid-frequency band. Mathematically, the mid-frequency range can be defined by $[f_{\text{low}}, f_{\text{high}}]$, where f_{low} and f_{high} are the lower and upper bounds, respectively. These are computed as $f_{\text{low}} = \eta_1 S$ and $f_{\text{high}} = \eta_2 S$, where η_1 and η_2 are proportional parameters. The selected mid-frequency coefficients can be expressed as

$$\mathbf{X}_{i,j}^{\text{mid}} = \{X_{i,j}(f_{\text{low}}), X_{i,j}(f_{\text{low}} + 1), \ldots, X_{i,j}(f_{\text{high}})\}. \tag{1}$$

The energy feature of the segment is then computed over the mid-frequency range as

$$E_{i,j} = \sum_{k=f_{\text{low}}}^{f_{\text{high}}} X_{i,j}^2(k). \tag{2}$$

TSM and PSM attacks can be approximately seen as linear scaling operations along the frequency axis, during which the signal's energy features also change linearly. This characteristic gives patchwork-based watermarking methods a degree of robustness against scaling attacks. In addition, the temporal and spectral stability of energy features, as well as their resistance to noise, further strengthens the overall robustness of watermarking schemes.

2.2. Watermark Embedding

The proposed method builds upon a patchwork-based strategy and incorporates further improvements inspired by Ref. [37]. In this method, watermark information is embedded by adjusting the energy ratio between two adjacent segments within the mid-frequency band of each frame. Let $\mathbf{w} = \{w_1, w_2, \ldots, w_N\} \in \{0,1\}^N$ represent the binary watermark sequence. The general embedding rule is defined as follows:

$$\begin{cases} E'_{i,1} \geq \alpha E'_{i,2}, & \text{if } w_i = 0, \\ E'_{i,2} \geq \alpha E'_{i,1}, & \text{otherwise.} \end{cases} \tag{3}$$

In this formulation, α is the embedding strength factor, and $E'_{i,1}$ and $E'_{i,2}$ denote the energies of the two modified segments. By creating a buffer zone, α improves extraction reliability. However, a small α may reduce robustness, while a large value may cause audible distortion. The energy comparison is conducted based on the magnitude of the DCT coefficients in the frequency domain. The detailed embedding procedure is described as follows:

Let $E_{i,1}$ and $E_{i,2}$ denote the energies of two non-overlapping segments within a frame. In case $w_i = 0$, the watermark bit is embedded by adaptively adjusting the energy relationship between the two segments. If the original energy values satisfy $E_{i,1} > \alpha E_{i,2}$, no modification is applied. Otherwise, the segment energies are adjusted to meet Equation (4) using two scaling factors β and γ, i.e.,

$$\begin{cases} E'_{i,1} = \beta^2 E_{i,1}, \\ E'_{i,2} = \gamma^2 E_{i,2}. \end{cases} \tag{4}$$

To minimize the auditory distortion caused by watermark embedding, β and γ are optimized and defined as follows:

$$\begin{cases} \beta = \dfrac{\alpha E_{i,2} + E_{i,2}\sqrt{\alpha E_{i,1} E_{i,2}}}{\alpha E_{i,2} + E_{i,1} E_{i,2}}, \\ \gamma = \dfrac{\sqrt{\alpha E_{i,1} E_{i,2}} + E_{i,1} E_{i,2}}{\alpha E_{i,2} + E_{i,1} E_{i,2}}. \end{cases} \tag{5}$$

The details of the optimization strategy are described in Section 2.4.

In case $w_i = 1$, the energy adjustment follows a similar procedure. If the condition $E_{i,2} > \alpha E_{i,1}$ is satisfied, the energy values remain unchanged. Otherwise, the energies are modified as

$$\begin{cases} E'_{i,1} = \beta^2 E_{i,1}, \\ E'_{i,2} = \gamma^2 E_{i,2}, \end{cases} \tag{6}$$

where β and γ are given by

$$\begin{cases} \beta = \dfrac{E_{i,2} + E_{i,2}\sqrt{\alpha E_{i,1} E_{i,2}}}{E_{i,2} + \alpha E_{i,1} E_{i,2}}, \\ \gamma = \dfrac{\sqrt{\alpha E_{i,1} E_{i,2}} + \alpha E_{i,1} E_{i,2}}{E_{i,2} + \alpha E_{i,1} E_{i,2}}. \end{cases} \tag{7}$$

To enhance the robustness against TSM and PSM attacks, a buffer compensation mechanism is introduced to ensure a smooth transition between the watermarked and non-watermarked frequency regions. Since the scaling factors β and γ only affect the frequency range $[f_{\text{low}}, f_{\text{high}}]$, setting a large embedding strength factor α may lead to significant differences between the modified coefficients within this band and the unmodified coefficients outside it. Some attacks may cause the feature extraction process to involve coefficients beyond the embedding band, thereby introducing severe distortion to the extracted features. To address this issue, we define a buffer zone with width ξ on both sides of the embedding band and apply a gradual scaling adjustment to the coefficients within this zone, achieving a smooth transition between the embedded and non-embedded regions. The scaling process is defined as

$$X'_{i,j}(k) = \begin{cases} \left[1 + \left(\sqrt{\dfrac{E'_{i,j}}{E_{i,j}}} - 1\right) \cdot \dfrac{k - (f_{\text{low}} - \xi)}{\xi}\right] \cdot X_{i,j}(k), & k \in [f_{\text{low}} - \xi, f_{\text{low}}], \\ \sqrt{\dfrac{E'_{i,j}}{E_{i,j}}} \cdot X_{i,j}(k), & k \in [f_{\text{low}}, f_{\text{high}}], \\ \left[1 + \left(\sqrt{\dfrac{E'_{i,j}}{E_{i,j}}} - 1\right) \cdot \left(1 - \dfrac{k - f_{\text{high}}}{\xi}\right)\right] \cdot X_{i,j}(k), & k \in [f_{\text{high}}, f_{\text{high}} + \xi]. \end{cases} \tag{8}$$

Here, $X'_{i,j}(k)$ denotes the watermarked coefficient at frequency index k after energy adjustment. It should be noted that the frequency coefficients within the $[f_{\text{low}}, f_{\text{high}}]$ band strictly follow the predefined embedding rules and represent the region of coefficients used for energy extraction during the watermark detection process. In summary, the buffer compensation mechanism ensures a smooth energy transition at the boundaries of the embedding region, thereby improving the robustness of the proposed method under frequency distortions.

After the above adjustments, the modified DCT coefficients are concatenated with the unaltered coefficients to reconstruct the complete frame. Applying the inverse DCT to all frames yields the final watermarked audio signal, denoted by $\mathbf{x}'$.

2.3. Watermark Extraction

Let $\mathbf{x}''$ denote the received audio at the decoder, which may be an attacked version of the original watermarked signal $\mathbf{x}'$. Clearly, if no attack occurs, we have $\mathbf{x}'' = \mathbf{x}'$. Without loss of generality, suppose we are extracting a watermark bit w''_i from the segment $\mathbf{x}''_i$. According to the feature extraction procedure in Section 2.1, the energy features $E''_{i,1}$ and $E''_{i,2}$ are first calculated within the frequency range $[f_{\text{low}}, f_{\text{high}}]$. The watermark bit w''_i is then extracted as

$$w''_i = \begin{cases} 0, & \text{if } E''_{i,1} \geq E''_{i,2}, \\ 1, & \text{otherwise.} \end{cases} \tag{9}$$

Based on Equation (9), the full watermark sequence $\mathbf{w}''$ can be extracted from $\mathbf{x}''$. Ideally, we expect that the extracted watermark $\mathbf{w}''$ equals the originally embedded watermark $\mathbf{w}$.

2.4. Parameter Optimization

In this section, we discuss our optimization strategy for selecting the parameters β and γ, with the objective of minimizing the distortion introduced to the DCT coefficients while preserving the watermark embedding conditions. Without loss of generality, we assume the embedded watermark bit is $w_i = 0$; the case for $w_i = 1$ can be addressed similarly. To ensure that the embedding rule remains unchanged, we aim to satisfy the following condition:

$$\frac{E'_{i,1}}{E'_{i,2}} = \alpha. \tag{10}$$

Using the energy modulation rules, this can be rewritten as

$$\beta^2 E_{i,1} = \alpha\gamma^2 E_{i,2}. \tag{11}$$

Solving for β, we obtain the relation between β and γ:

$$\beta = \gamma\sqrt{\frac{\alpha E_{i,2}}{E_{i,1}}}. \tag{12}$$

We define the distortion D_i introduced by modifying the DCT coefficients as

$$D_i = \sum_{j\in[1,2]} \sum_{k\in[f_{\text{low}},f_{\text{high}}]} \frac{(X'_{i,j}(k) - X_{i,j}(k))^2}{\left|\mathbf{X}_{i,j}[f_{\text{low}}, f_{\text{high}}]\right|}. \tag{13}$$

The distortion metric in Equation (13) is computed over the core embedding band $[f_{\text{low}}, f_{\text{high}}]$, where the embedding rule and detection features are defined. The adjacent buffer zones $[f_{\text{low}} - \xi, f_{\text{low}})$ and $(f_{\text{high}}, f_{\text{high}} + \xi]$ apply a linear taper (Equation (8)) to ensure smooth coefficient transitions. The resulting per-coefficient distortion in the buffer is attenuated compared to the core band, and its total energy is much smaller due to the narrow width of the buffer. Therefore, we only consider the total distortion over the core embedding band. Based on Equation (11) and the coefficient scaling rule, the distortion can be reformulated as

$$\begin{aligned} D_i &= \sum_{j\in[1,2]} \left(\sqrt{\frac{E'_{i,j}}{E_{i,j}}} - 1\right)^2 E_{i,j} \\ &= (\beta-1)^2 E_{i,1} + (\gamma-1)^2 E_{i,2} \\ &= \left(\gamma\sqrt{\frac{\alpha E_{i,2}}{E_{i,1}}} - 1\right)^2 E_{i,1} + (\gamma-1)^2 E_{i,2}. \end{aligned} \tag{14}$$

Obviously, D_i is a function of γ. To minimize D_i, we take the derivative with respect to γ and set it to zero:

$$\frac{\mathrm{d}D_i}{\mathrm{d}\gamma} = 2\left(\gamma\sqrt{\frac{\alpha E_{i,2}}{E_{i,1}}} - 1\right)\sqrt{\frac{\alpha E_{i,2}}{E_{i,1}}}\, E_{i,1} + 2(\gamma-1)E_{i,2} = 0. \tag{15}$$

Solving Equation (15) yields

$$\gamma = \frac{\sqrt{\alpha E_{i,1}E_{i,2}} + E_{i,1}E_{i,2}}{\alpha E_{i,2} + E_{i,1}E_{i,2}}. \tag{16}$$

Substituting Equation (16) into Equation (11) gives

$$\beta = \frac{\alpha E_{i,2} + E_{i,2}\sqrt{\alpha E_{i,1} E_{i,2}}}{\alpha E_{i,2} + E_{i,1} E_{i,2}}. \tag{17}$$

It is evident that the above results satisfy Equation (5).

2.5. Theoretical Analysis of Robustness

This section theoretically analyzes the robustness of the proposed method against common signal processing and desynchronization attacks. When the embedded watermark bit is $w_i = 0$, the energy features after embedding satisfy the condition

$$\frac{E'_{i,1}}{E'_{i,2}} \geq \alpha. \tag{18}$$

Without loss of generality, we consider the equality case

$$\frac{E'_{i,1}}{E'_{i,2}} = \alpha, \tag{19}$$

as the basis for robustness analysis. After the audio signal is subjected to various attacks, the intra-frame energy relationship can be modeled as

$$\frac{E''_{i,1}}{E''_{i,2}} = \frac{E'_{i,1} + \Delta_1}{E'_{i,2} + \Delta_2}, \tag{20}$$

where $\Delta_1 = E''_{i,1} - E'_{i,1}$ and $\Delta_2 = E''_{i,2} - E'_{i,2}$. It is evident that the proposed method is inherently resistant to closed-loop attacks, in which $\Delta_1 = \Delta_2 = 0$. For low-pass filtering, MP3 compression, and AAC compression attacks, which primarily affect high-frequency components, the mid-frequency band used for feature embedding remains mostly intact. Thus, we can reasonably assume that $\Delta_1 \approx \Delta_2 \approx 0$ under these conditions. In the case of amplitude scaling attacks, all frequency coefficients within a frame are scaled by the same factor σ, resulting in

$$\frac{E''_{i,1}}{E''_{i,2}} = \frac{\sigma^2 E'_{i,1}}{\sigma^2 E'_{i,2}} = \frac{E'_{i,1}}{E'_{i,2}}, \tag{21}$$

which demonstrates that the proposed method is robust to amplitude scaling. For echo addition, resampling, quantization, and additive noise attacks, the intra-frame energy ratio can be rewritten as

$$\frac{E''_{i,1}}{E''_{i,2}} = \frac{\alpha E'_{i,2} + \Delta_1}{E'_{i,2} + \Delta_2}. \tag{22}$$

To ensure correct watermark extraction (i.e., $w''_i = 0$), we require

$$\frac{E''_{i,1}}{E''_{i,2}} \geq 1, \tag{23}$$

which leads to the condition

$$E'_{i,2} \geq \frac{\Delta_2 - \Delta_1}{\alpha - 1}. \tag{24}$$

Since Δ_1 and Δ_2 reflect the energy perturbation in adjacent sub-segments within the same frame, we can reasonably assume $\Delta_2 - \Delta_1 \approx 0$. Moreover, $E'_{i,2}$ captures the dominant energy in the sub-segment, and is typically far greater than $\Delta_2 - \Delta_1$. Additionally, the threshold α can be adjusted to introduce a margin of tolerance, making the inequality easily

satisfied. Therefore, the proposed method is robust against echo addition, resampling, quantization, and noise attacks.

As stated in Ref. [28], time-scale and pitch-scale modification attacks can be modeled as stretching or compressing operations along the frequency axis. These distortions may cause frequency components outside the embedding region to be involved in feature extraction, while some originally embedded coefficients may be excluded. To mitigate this issue, the proposed buffer compensation mechanism smooths the transition between watermarked and non-watermarked regions, thereby reducing the influence of incorrectly involved coefficients during extraction. Assuming the embedded watermark bit is 0 (the analysis for bit 1 is analogous), we have

$$E''_{i,j} = E'_{i,j} + \Delta^{(j)}_{\text{interf}}. \tag{25}$$

Here, $\Delta^{(j)}_{\text{interf}}$ represents the interference energy caused by the inclusion of non-embedded coefficients and the loss of embedded ones. To correctly extract the watermark bit, the following condition must hold:

$$\frac{E''_{i,1}}{E''_{i,2}} = \frac{E'_{i,1} + \Delta^{(1)}_{\text{interf}}}{E'_{i,2} + \Delta^{(2)}_{\text{interf}}} \geq 1. \tag{26}$$

This can be simplified as

$$\alpha \geq \frac{1 + \Delta^{(2)}_{\text{interf}} - \Delta^{(1)}_{\text{interf}}}{E'_{i,2} - 1}. \tag{27}$$

Next, we examine $\Delta^{(2)}_{\text{interf}}$ and $\Delta^{(1)}_{\text{interf}}$ in conjunction with Equation (8). A linear scaling factor smaller than 1 is applied to $\Delta^{(2)}_{\text{interf}}$, while a factor greater than 1 is applied to $\Delta^{(1)}_{\text{interf}}$. This results in a reduced numerator; when the TSM and PSM attack intensities are moderate, $E'_{i,2} - 1$ is typically much greater than α, ensuring that the above inequality is satisfied.

In the case of moderate jittering attacks, where a small proportion of samples is randomly removed, we also assume $\Delta_1 - \Delta_2 \approx 0$. This suggests that the proposed method retains its robustness against jittering as well.

3. Experimental Results and Analysis

3.1. Setup

To evaluate the robustness and imperceptibility of the proposed method, all experiments are conducted using the same audio dataset as in Ref. [16]. The dataset consists of audio clips from 25 different genres. Each audio sample is uniformly processed to a duration of 20 s, with a sampling rate of 44.1 kHz and a quantization precision of 16 bits. All audio samples are embedded with watermarks using the proposed algorithm. Specifically, the lower and upper frequency bounds are set to $f_{\text{low}} = 3000$ Hz and $f_{\text{high}} = 4000$ Hz, corresponding to the proportional parameters $\eta_1 = 0.1361$ and $\eta_2 = 0.1814$, respectively. The buffer compensation width is configured as $\xi = 500$ Hz, and the embedding strength parameter is set to $\alpha = 8$. All experiments are conducted on a workstation equipped with a 12th Gen Intel(R) Core(TM) i7-12700H CPU at 2.30 GHz, 16 GB RAM, and an NVIDIA RTX 3060 GPU, using MATLAB R2020b as the software environment.

Metrics

To assess the perceptual quality of the watermarked audio, two commonly used objective metrics are employed: Signal-to-Noise Ratio (SNR) and Objective Difference

Grade (ODG). According to the guidelines established by the International Federation of the Phonographic Industry (IFPI), an SNR value above 20 dB is generally required to ensure acceptable perceived audio quality. The SNR is computed as follows:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{\sum_{i=1}^{n} X(i)^2}{\sum_{i=1}^{n} (X(i) - X'(i))^2} \right), \tag{28}$$

ODG is a key output of the Perceptual Evaluation of Audio Quality (PEAQ) standard, which simulates human auditory perception based on psychoacoustic models. ODG values typically range from -4 (very annoying) to 0 (imperceptible), with values between -1 and 0 generally indicating no perceptible degradation in audio quality.

To evaluate the robustness of the embedded watermark against common signal processing attacks, two additional metrics are adopted: Bit Error Rate (BER) and Normalized Cross-Correlation (NCC).

3.2. Imperceptibility Experiment

To further validate the transparency of the proposed embedding strategy, we compare the waveforms before and after watermark embedding for six different music genres at an embedding rate of 30 bps. As depicted in Figure 2, the waveforms remain visually indistinguishable before and after watermarking, demonstrating that our method introduces negligible perceptual distortion.

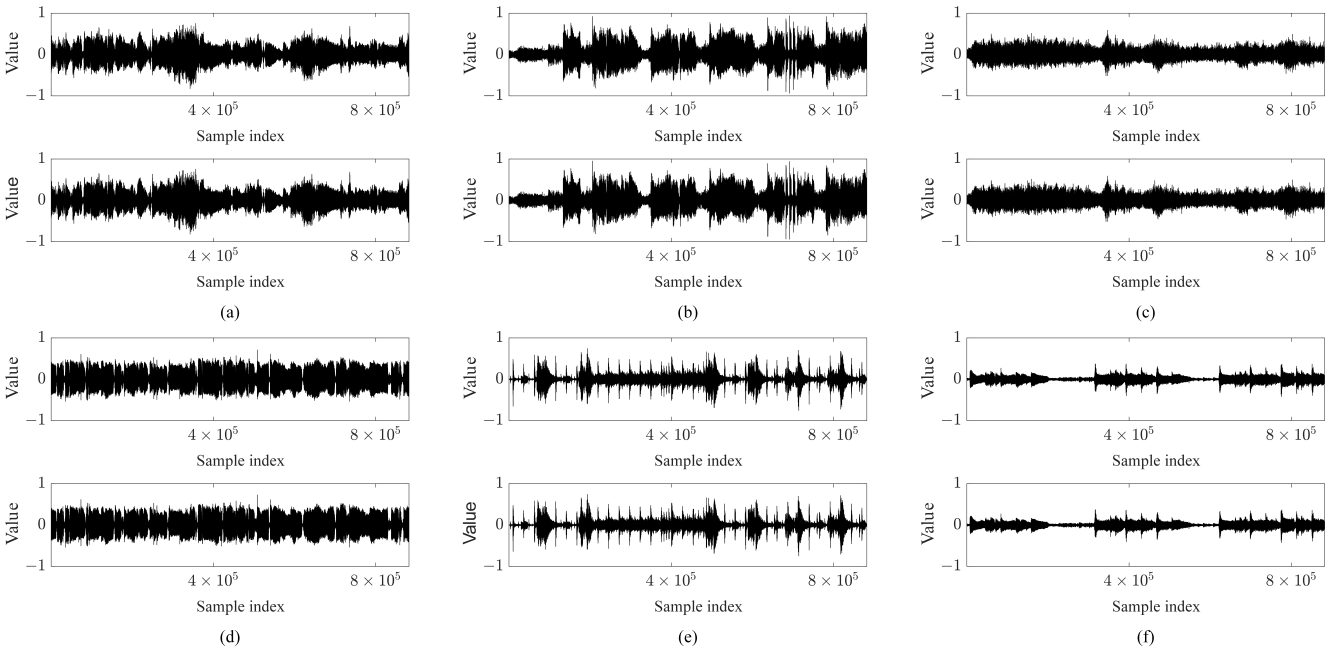


Figure 2. Comparison of waveform segments for various types of audio signals at an embedding rate of 30 bps. The original audio (top) and the corresponding watermarked audio (bottom) are shown for each genre: (**a**) hip-hop, (**b**) popular music, (**c**) folk music, (**d**) classical, (**e**) jazz, and (**f**) piano.

Figure 3a shows the mean SNR values under different embedding strength factors α. As expected, both higher embedding rates and stronger embedding strengths negatively affect the SNR. Specifically, the SNR gradually decreases as α increases. Nevertheless, the proposed method maintains a high perceptual quality even under large α values.

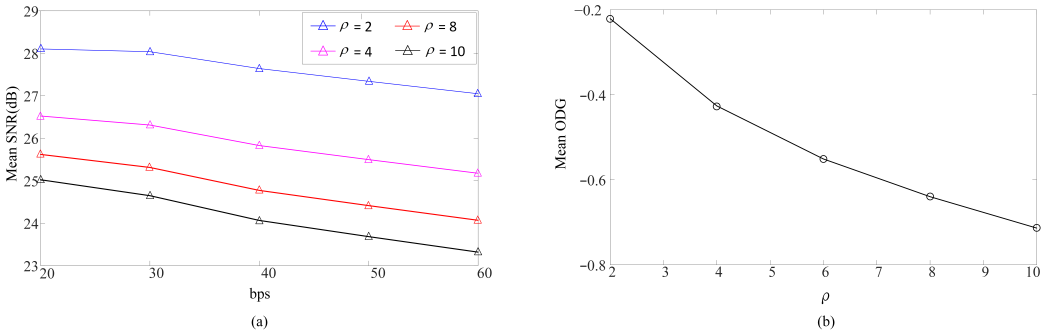


Figure 3. Mean SNRs and ODGs under different embedding strengths and embedding rates. (**a**) The mean SNRs under varying embedding rates and strength parameters (**b**) The mean ODGs under different embedding strengths.

Figure 3b illustrates the mean ODG values at 30 bps for different values of α. As anticipated, ODG decreases with increasing embedding strength. However, even at a relatively large embedding strength (e.g., $\alpha = 10$), the ODG remains greater than -1, indicating that the proposed method maintains high perceptual transparency. This result can be attributed to our embedding optimization strategy, which minimizes the distortion power while preserving the desired embedding proportion.

3.3. Robustness Against Common Attacks

To evaluate the robustness of the proposed watermarking method under realistic signal processing operations, we conduct a series of experiments involving various common audio attacks. These attacks are representative of distortions that may occur during typical transmission, compression, and playback scenarios. The goal is to assess whether the embedded watermark can be reliably extracted after such distortions.

- Non-attack: The watermark is extracted directly from the watermarked audio signal without any modification.
- Noise attack: Additive white Gaussian noise (AWGN) is added to the watermarked audio.
- Quantization attack: The watermarked audio is re-quantized before watermark extraction.
- Amplitude attack: The amplitude of the watermarked audio is scaled.
- Echo attack: An echo signal with a time delay is added to the watermarked audio.
- MP3 attack: The watermarked audio is compressed using MPEG-1 Layer III encoding.
- AAC attack: The watermarked audio is compressed using MPEG-4 Advanced Audio Coding (AAC).
- Resampling attack: The watermarked audio is first downsampled and then upsampled.
- Low-pass filter attack: A low-pass filter is applied to the watermarked signal.
- High-pass filter attack: A high-pass filter is applied to the watermarked signal.

We compare the proposed method with three advanced audio watermarking algorithms, namely those presented in Refs. [21,28,35]. To ensure a fair comparison, the embedding strength of each method is adjusted such that the resulting watermarked audio achieves a similar perceptual quality, with the SNR uniformly set to 25 dB. Typical parameters are selected for each type of attack. Specifically, for the noise attack, the SNR between the watermarked audio and the added noise is set to 20 dB and 30 dB, respectively. In the quantization attack, the bit depth is reduced from 16 bits to 8 bits. For amplitude scaling, scaling factors of 0.8 and 1.2 are applied. For the echo attack, the delay is set to 0.5 s, and the echo amplitude is 20% of the original signal. For MP3 compression, bitrates of 128 kbps and 96 kbps are used, with the same settings applied for AAC compression. In

the resampling attack, the watermarked audio is downsampled from 44.1 kHz to 22.05 kHz and then upsampled back to 44.1 kHz. For low-pass filtering, the cutoff frequency is 8 kHz, while for high-pass filtering, it is 0.5 kHz.

As shown in Table 1, the proposed method, along with the methods in Refs. [21,28], demonstrates strong robustness against common signal processing attacks, achieving low BER. In contrast, the method in Ref. [35] exhibits weaker robustness under noise, low-pass filtering, and resampling attacks. This limitation stems from its reliance on residual-based FDLM features, which are inherently sensitive to noise and sampling variations. Moreover, its feature extraction is performed on the top 75% of the frequency band coefficients, which includes a significant proportion of high-frequency components, rendering the method more susceptible to low-pass filtering.

Table 1. The mean BERs (%) and mean NCC of different methods against common attacks at an embedding rate of 30 bps.

Common Attacks		Ref. [21]		Ref. [35]		Ref. [28]		Proposed	
		BER	NCC	BER	NCC	BER	NCC	BER	NCC
Non-attack		0.00	1.000	0.00	1.000	0.00	1.000	0.00	1.000
Noise	30 dB	0.05	0.999	22.59	0.688	0.17	0.998	0.13	0.999
	20 dB	1.27	0.987	42.44	0.300	2.08	0.978	1.45	0.985
Quantization		0.11	0.999	14.37	0.856	0.19	0.998	0.13	0.999
Amplitude	0.8	0.00	1.000	0.00	1.000	0.00	1.000	0.00	1.000
	1.2	0.00	1.000	0.00	1.000	0.00	1.000	0.00	1.000
Echo		1.04	0.989	7.29	0.919	2.12	0.978	4.01	0.960
MP3	128 kbps	0.00	1.000	2.05	0.976	0.00	1.000	0.00	1.000
	96 kbps	0.00	1.000	8.41	0.902	0.00	1.000	0.00	1.000
AAC	128 kbps	0.00	1.000	0.02	1.000	0.00	1.000	0.00	1.000
	96 kbps	0.00	1.000	0.02	1.000	0.00	1.000	0.00	1.000
Resampling		0.00	1.000	30.78	0.593	0.00	1.000	0.00	1.000
Low-pass filtering		0.00	1.000	0.04	1.000	0.00	1.000	0.00	1.000
High-pass filtering		0.00	1.000	44.15	0.355	0.03	1.000	0.00	1.000

One of the most significant contributions of this paper is its robustness against desynchronization attacks. Common types of desynchronization attacks considered include the following:

- Jitter: A certain proportion of sample points from each segment of the watermarked signal is randomly removed.
- TSM: The duration of the signal is modified while the pitch remains unchanged.
- PSM: The pitch of the signal is altered while maintaining its duration.

Representative parameters are selected for each attack. Specifically, for the jittering attack, 1 sample is randomly removed from every 100 samples. For TSM, the time scaling factor is selected from the range [0.8, 1.2]. Similarly, for PSM, the pitch scaling factor is varied within [0.8, 1.2] while keeping the duration unchanged. The experimental results under desynchronization attacks are presented in Table 2. It can be observed that the proposed method achieves significant superiority in resisting desynchronization, with all BERs remaining below 13%, which aligns well with the theoretical analysis discussed in

Section 2.5. In comparison, the method in [21] exhibits partial robustness against jitter and TSM attacks, but fails completely under PSM, while the method in [35] suffers from BERs exceeding 20% under TSM and PSM due to the high embedding rate of 30 bps, which exceeds its operational capacity. Both the proposed method and the method in [28] show strong resistance to TSM and PSM attacks; however, our method consistently outperforms [28], owing to the incorporation of a linearly decreasing buffer compensation mechanism. In general, desynchronization attacks disrupt watermark embedding positions, significantly increasing the difficulty of accurate extraction. The compared methods fail to effectively address this challenge due to the lack of robust synchronization features or mechanisms. In contrast, our method exhibits the best overall performance, demonstrating its superior robustness against various forms of desynchronization.

Table 2. The mean BERs (%) and mean NCC of different methods under desynchronization attacks at an embedding rate of 30 bps.

Desynchronization Attacks		**Ref. [21]**		**Ref. [35]**		**Ref. [28]**		**Proposed**	
		BER	**NCC**	**BER**	**NCC**	**BER**	**NCC**	**BER**	**NCC**
Jitter (1%)		3.39	0.963	37.93	0.446	9.58	0.906	10.55	0.902
TSM	80%	0.25	0.997	25.47	0.676	1.83	0.981	1.24	0.987
	90%	0.13	0.999	22.26	0.724	1.91	0.980	0.68	0.993
	110%	0.15	0.998	19.19	0.767	2.27	0.976	1.93	0.980
	120%	0.17	0.998	21.33	0.737	6.51	0.933	7.40	0.923
PSM	80%	31.94	0.639	33.18	0.544	11.36	0.882	7.03	0.928
	90%	49.13	0.441	27.83	0.639	3.49	0.963	2.17	0.978
	110%	52.10	0.382	20.47	0.750	8.02	0.916	3.71	0.962
	120%	30.83	0.663	24.11	0.697	16.48	0.829	12.31	0.872

3.4. Runtime Analysis

In addition to robustness and imperceptibility evaluations, we also assess the computational overhead of the proposed embedding algorithm to verify its suitability for real-time applications. The runtime measurements were conducted on a 12th Gen Intel(R) Core(TM) i7-12700H CPU with 16 GB RAM, running MATLAB 2020b.

In our implementation, each audio clip has a length of 20 s sampled at 44.1 kHz, resulting in a total of 882,000 samples. The signal is divided into $N_f = 600$ equal-length frames, giving $L_f = 1470$ samples per frame, corresponding to approximately 33.33 ms of audio. Table 3 presents the average runtime per frame for each main processing step. The results show that the proposed method achieves an average embedding runtime of 0.209 ms per frame (std: 0.016 ms), corresponding to a throughput of approximately 4784 frames/s, i.e., more than 160 s of audio per second of processing time. Given that typical real-time audio processing requires fewer than 100 frames/s, the proposed method easily satisfies real-time requirements. Notably, the DCT/IDCT operations dominate the total runtime, while buffer scaling accounts for only a small fraction. The mean SNR remains at 25.61 dB, indicating that the imperceptibility is preserved.

Table 3. Average runtime per frame for the proposed method. Std values are only reported for the total runtime.

Step	Avg Runtime (ms)	Std (ms)	Percentage
DCT + IDCT	0.185	–	88.5%
Buffer Scaling	0.007	–	3.3%
Other Operations	0.017	–	8.2%
Total	0.209	0.016	100%

3.5. Ablation Study

To verify the effectiveness of the optimization strategy, we conduct comparative experiments under scenarios with and without parameter optimization. We evaluate the audio quality using the Signal-to-Noise Ratio (SNR) and Objective Difference Grade (ODG) as objective metrics. As shown in Figure 4, the proposed optimization yields significant improvements in both SNR and ODG across varying embedding strengths. This improvement can be attributed to the reduced energy distortion achieved by the analytical optimization of β and γ.

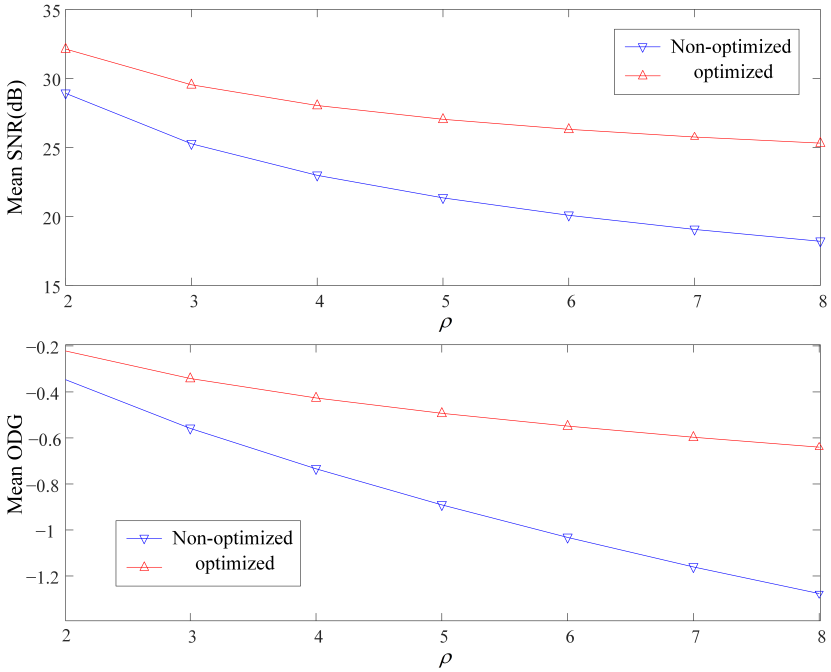


Figure 4. The mean SNRs and ODGs with and without parameter optimization.

To further evaluate the impact of buffer window length ξ on watermarking performance, a detailed analysis is conducted under four different desynchronization attack scenarios. As illustrated in Figure 5a, the mean BERs are plotted for TSM and PSM attacks with scaling factors of 0.8 and 1.2, respectively, across a range of ξ values from 0 to 0.75 kHz. The results show that BER consistently decreases as ξ increases under all conditions. In particular, when $\xi = 0$, the method still exhibits a certain degree of robustness, but the BER is noticeably higher—especially under PSM with a scaling factor of 1.2. This demonstrates that the buffer compensation mechanism can effectively alleviate the distortions caused by desynchronization attacks. As ξ increases, more adjacent coefficients are involved in the embedding process, providing a buffer compensation effect that suppresses desynchronization and significantly reduces BER, thereby enhancing overall robustness.

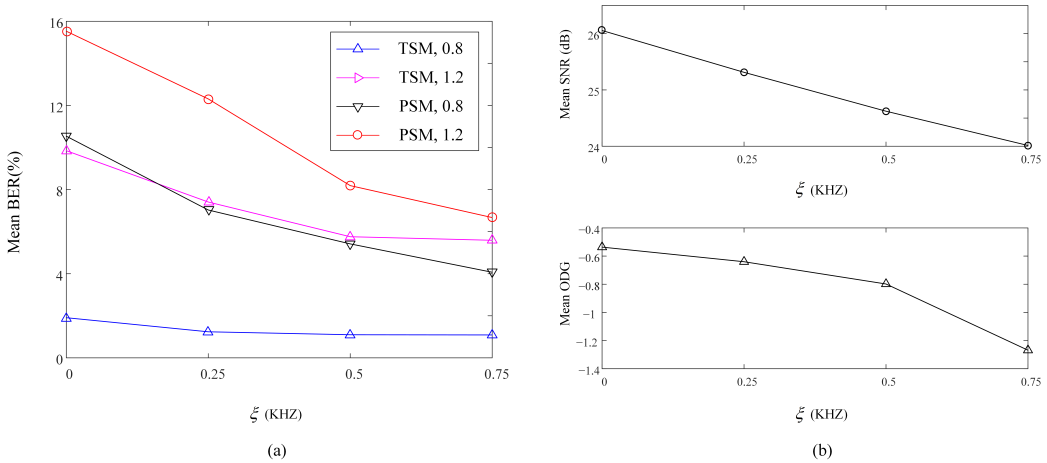


Figure 5. Impact of buffer window length on watermarking performance under TSM and PSM attacks.

We also investigate the influence of increasing ξ on perceptual audio quality. As shown in Figure 5b, both the average SNR and ODG decrease as ξ grows. This trend is expected, as a wider embedding region introduces greater energy modifications. Notably, when $\xi \geq 0.75$ kHz, the ODG value drops below -1.0, suggesting that the distortion introduced by watermarking may become perceptible to human listeners. Nevertheless, the SNR remains above 24 dB, and the ODG does not fall below typical acceptability thresholds.

In summary, increasing the buffer window length significantly improves robustness, particularly under severe desynchronization attacks such as PSM. However, it also introduces a trade-off in perceptual audio quality. To balance robustness and audio quality, a buffer length of $\xi = 0.5$ kHz is selected in our experiments.

To further examine the limitations of the proposed method, we conduct supplementary tests under extremely strong desynchronization conditions (130% TSM/PSM). Figure 6 presents the mean BER versus ξ for TSM 0.7, TSM 1.3, PSM 0.7, and PSM 1.3. The results show that for PSM 0.7, the BER remains above 20% when $\xi \leq 0.5$ kHz, and for TSM 1.3, the BER exceeds 8% even at $\xi = 0.75$ kHz. These findings confirm that while the buffer compensation mechanism effectively mitigates moderate desynchronization, extremely strong scaling factors still cause noticeable degradation. At the same time, the results validate our choice of $\xi = 0.5$ kHz, which offers a good balance between robustness improvement and imperceptibility under both normal and extreme attack scenarios.

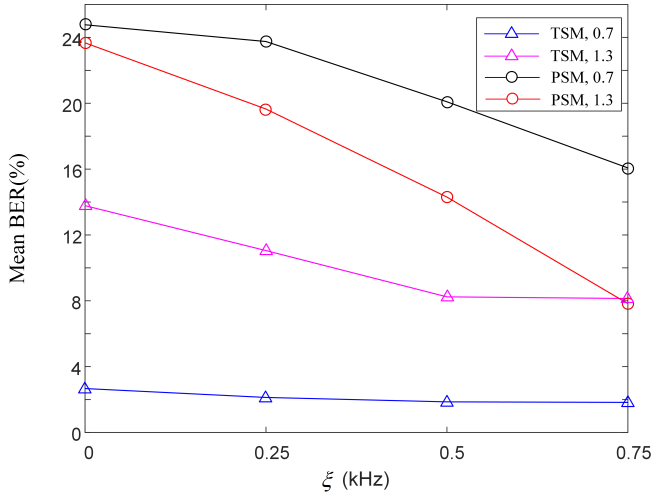


Figure 6. Performance under extremely strong desynchronization attacks (130% TSM/PSM) with varying buffer window length ξ.

4. Conclusions and Discussion

This paper presents a novel patchwork-based audio watermarking method designed to address the persistent challenge of desynchronization attacks in robust audio watermarking. The proposed method enhances resistance to typical desynchronization distortions, including TSM, PSM, and jitter, while maintaining high audio quality and a reasonable embedding capacity. Specifically, the audio signal is first divided into frames, and each frame is further split into two non-overlapping segments. Watermark bits are embedded by modulating the energy ratio of mid-frequency DCT coefficients between the two segments. A key innovation of this method lies in the introduction of a linearly decreasing buffer compensation mechanism, which alleviates boundary inconsistencies between watermark and non-watermark coefficients. By gradually reducing the scaling factors near the boundaries, this mechanism not only mitigates distortion caused by coefficient modification but also suppresses interference from non-watermark coefficients during feature extraction, thereby significantly improving watermark detection accuracy under TSM and PSM attacks. Experimental results demonstrate that the proposed method achieves a favorable balance among imperceptibility, embedding rate, and robustness. Under a wide range of desynchronization attacks, including TSM, PSM, and jitter, the BER consistently remains below 13%, validating the effectiveness of the buffer compensation mechanism and optimization strategy. Compared with existing state-of-the-art techniques, the proposed method exhibits superior robustness and perceptual transparency, confirming its applicability to practical audio watermarking scenarios.

Author Contributions: Conceptualization, W.Z. and H.W.; Methodology, W.Z.; Software, W.Z. and G.Z.; Validation, W.Z., Y.Z. and D.W.; Formal analysis, W.Z.; Investigation, W.Z.; Resources, H.W.; Data curation, Z.D. and J.Y.; Writing—original draft preparation, W.Z.; Writing—review and editing, H.W. and Y.Z.; Visualization, W.Z.; Supervision, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the 2024 Xizang Autonomous Region Central Guided Local Science and Technology Development Fund Project under Grant Number XZ202401YD0015, the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant Number 24ZR1424000, the National Natural Science Foundation of China (NSFC) under Grant Number U23B2023, and the Basic Research Program for Natural Science of Guizhou Province under Grant Number QIANKEHEJICHU-ZD[2025]-043.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The audio dataset used to support the findings of this study is openly available in Ref. [16]. The source code and further experimental data are available on reasonable request from the corresponding author.

Conflicts of Interest: Author Jingyu Ye is affiliated with Digital Guangxi Group. The authors declare no conflicts of interest.

References

1. Yang, Z.; Zhao, G.; Wu, H. Watermarking for large language models: A survey. *Mathematics* **2025**, *13*, 1420. [CrossRef]
2. Wang, B.; Zhao, P. An adaptive image watermarking method combining SVD and Wang-Landau sampling in DWT domain. *Mathematics* **2020**, *8*, 691. [CrossRef]
3. Zheng, Q.; Liu, N.; Wang, F. An adaptive embedding strength watermarking algorithm based on Shearlets' capture directional features. *Mathematics* **2020**, *8*, 1377. [CrossRef]
4. Ye, C.; Tan, S.; Wang, J.; Shi, L.; Zuo, Q.; Xiong, B. Double security level protection based on chaotic maps and SVD for medical images. *Mathematics* **2025**, *13*, 182. [CrossRef]

5. Wang, Y.; Xue, Y.; Liu, X.; Wen, J. An adaptive audio watermarking with frame-wise control parameter searching. *Digit. Signal Process.* **2025**, *160*, 105025. [CrossRef]
6. Zhao, J.; Zong, T.; Xiang, Y.; Hua, G.; Lei, X.; Gao, L.; Beliakov, G. Frequency spectrum modification process-based anti-collusion mechanism for audio signals. *IEEE Trans. Cybern.* **2023**, *53*, 5510–5522. [CrossRef]
7. Kim, H.-J.; Choi, Y.-H. A novel echo-hiding scheme with backward and forward kernels. *IEEE Trans. Circuits Syst. Video Technol.* **2003**, *13*, 885–889. [CrossRef]
8. Xiang, S.; Huang, J. Histogram-based audio watermarking against time-scale modification and cropping attacks. *IEEE Trans. Multimed.* **2007**, *9*, 1357–1372. [CrossRef]
9. Huang, X.; Ito, A. Imperceptible and reversible acoustic watermarking based on modified integer discrete cosine transform coefficient expansion. *Appl. Sci.* **2024**, *14*, 2757. [CrossRef]
10. Hua, G.; Goh, J.; Thing, V.L. Cepstral analysis for the application of echo-based audio watermark detection. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 1850–1861. [CrossRef]
11. Wang, S.; Yuan, W.; Zhang, Z.; Wang, J.; Unoki, M. Synchronous multi-bit audio watermarking based on phase shifting. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2700–2704.
12. Dronyuk, I.; Fedevych, O.; Kryvinska, N. Constructing of digital watermark based on generalized Fourier transform. *Electronics* **2020**, *9*, 1108. [CrossRef]
13. Hu, H.-T.; Hsu, L.-Y. Robust, transparent and high-capacity audio watermarking in dct domain. *Signal Process.* **2015**, *109*, 226–235. [CrossRef]
14. Saadi, S.; Merrad, A.; Benziane, A. Novel secured scheme for blind audio/speech norm-space watermarking by arnold algorithm. *Signal Process.* **2019**, *154*, 74–86. [CrossRef]
15. Jiang, W.; Huang, X.; Quan, Y. Audio watermarking algorithm against synchronization attacks using global characteristics and adaptive frame division. *Signal Process.* **2019**, *162*, 153–160. [CrossRef]
16. Karajeh, H.; Khatib, T.; Rajab, L.; Maqableh, M. A robust digital audio watermarking scheme based on dwt and schur decomposition. *Multimed. Tools Appl.* **2019**, *78*, 18395–18418. [CrossRef]
17. Wu, Q.; Wu, M. Adaptive and blind audio watermarking algorithm based on chaotic encryption in hybrid domain. *Symmetry* **2018**, *10*, 284. [CrossRef]
18. Xiang, Y.; Natgunanathan, I.; Peng, D.; Hua, G.; Liu, B. Spread spectrum audio watermarking using multiple orthogonal pn sequences and variable embedding strengths and polarities. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *26*, 529–539. [CrossRef]
19. Zhong, J.; Huang, S. An enhanced multiplicative spread spectrum watermarking scheme. *IEEE Trans. Circuits Syst. Video Technol.* **2006**, *16*, 1491–1506. [CrossRef]
20. Xiang, Y.; Natgunanathan, I.; Rong, Y.; Guo, S. Spread spectrum-based high embedding capacity watermarking method for audio signals. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 2228–2237. [CrossRef]
21. Natgunanathan, I.; Xiang, Y.; Hua, G.; Beliakov, G.; Yearwood, J. Patchwork-based multilayer audio watermarking. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2176–2187. [CrossRef]
22. Natgunanathan, I.; Xiang, Y.; Rong, Y.; Peng, D. Robust patchwork-based watermarking method for stereo audio signals. *Multimed. Tools Appl.* **2014**, *72*, 1387–1410. [CrossRef]
23. Natgunanathan, I.; Xiang, Y.; Rong, Y.; Zhou, W.; Guo, S. Robust patchwork-based embedding and decoding scheme for digital audio watermarking. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 2232–2239. [CrossRef]
24. Malvar, H.; Florencio, D. Improved spread spectrum: A new modulation technique for robust watermarking. *IEEE Trans. Signal Process.* **2003**, *51*, 898–905. [CrossRef]
25. Hwang, M.-J.; Lee, J.; Lee, M.; Kang, H.-G. Svd-based adaptive qim watermarking on stereo audio signals. *IEEE Trans. Multimed.* **2017**, *20*, 45–54. [CrossRef]
26. Khaldi, K.; Boudraa, A.-O. Audio watermarking via emd. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 675–680. [CrossRef]
27. Erfani, Y.; Pichevar, R.; Rouat, J. Audio watermarking using spikegram and a two-dictionary approach. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 840–852. [CrossRef]
28. Zhao, J.; Zong, T.; Xiang, Y.; Gao, L.; Zhou, W.; Beliakov, G. Desynchronization attacks resilient watermarking method based on frequency singular value coefficient modification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2282–2295. [CrossRef]
29. Zhang, G.; Zheng, L.; Su, Z.; Zeng, Y.; Wang, G. M-sequences and sliding window based audio watermarking robust against large-scale cropping attacks. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1182–1195. [CrossRef]
30. Lei, B.; Soon, I.Y.; Tan, E.-L. Robust SVD-based audio watermarking scheme with differential evolution optimization. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 2368–2378. [CrossRef]
31. Gui-jun, N.; Shuxun, W. Robust adaptive audio watermarking algorithm in cepstrum. *J. Jilin Univ.* **2008**, *26*, 55–61.

32. Dessein, A.; Cont, A. An information-geometric approach to real-time audio segmentation. *IEEE Signal Process. Lett.* **2013**, *20*, 331–334. [CrossRef]
33. Xiang, Y.; Natgunanathan, I.; Guo, S.; Zhou, W.; Nahavandi, S. Patchwork-based audio watermarking method robust to de-synchronization attacks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1413–1423. [CrossRef]
34. Li, J.; Xiang, S. Audio-lossless robust watermarking against desynchronization attacks. *Signal Process.* **2022**, *198*, 108561. [CrossRef]
35. Liu, Z.; Huang, Y.; Huang, J. Patchwork-based audio watermarking robust against de-synchronization and recapturing attacks. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 1171–1180. [CrossRef]
36. Zhao, J.; Zong, T.; Natgunanathan, I.; Xiang, Y.; Song, X.; Hua, G.; Gao, L.; Zhou, W. Fragment-energy audio watermarking resilient to de-synchronization attacks. *Expert Syst. Appl.* **2025**, *296*, 128980. [CrossRef]
37. Zhao, J.; Zong, T.; Xiang, Y.; Gao, L.; Hua, G.; Sood, K.; Zhang, Y. Ssvs-ssvd based desynchronization attacks resilient watermarking method for stereo signals. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 448–461. [CrossRef]

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