Robust Zero-Watermark Algorithm Based on Audio Fundamental Frequency Features (Postprint)

Authors: Zheng Mengyi, Li Chen, Tian Lihua

Date: 2018-05-18T00:00:00+00:00

Abstract

To integrate digital audio watermarking technology with audio content, a robust zero-watermarking method is proposed by extracting the fundamental frequency from music based on the stability characteristics of the fundamental frequency in vocals and instruments. First, the fundamental frequency is extracted via the Normalized Subharmonic Summation algorithm, then the K-means algorithm is employed to encode the fundamental frequency features to enhance their stability, and finally a zero-watermark sequence is generated through XOR operation with the watermark image. Additionally, in the event of malicious tampering, the tampered region can be determined by comparing the fundamental frequency information in the zero-watermark to identify inconsistent parts. Experimental results demonstrate that this algorithm exhibits favorable stability under both conventional attacks and jitter attacks, and can achieve tamper detection.

Full Text

Robust Zero-Watermarking Algorithm Based on Audio Fundamental Frequency Feature

Zheng Mengyi, Li Chen, Tian Lihua†
(School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract

To integrate digital audio watermarking technology with audio content, this paper proposes a robust audio zero-watermarking method based on fundamental frequency features extracted from music, leveraging the stability of fundamental frequencies in human voice and musical instruments. The algorithm first extracts the fundamental frequency using the normalized subharmonic summation algorithm, then enhances its stability through K-means clustering-based

encoding of the fundamental frequency features, and finally generates a zero-watermark sequence via XOR operation with the watermark image. Additionally, when malicious tampering occurs, the tampered region can be identified by comparing the inconsistent portions of the fundamental frequency information contained in the zero-watermark. Experimental results demonstrate that the proposed algorithm exhibits excellent stability under both conventional attacks and jitter attacks, and can effectively detect and localize tampering.

Keywords: digital audio watermarking; robust zero watermark; normalized subharmonic summation; fundamental frequency; tamper detection

Introduction

The rapid development of computer and internet technologies has facilitated the creation, storage, and transmission of digital products, while simultaneously exacerbating piracy issues such as illegal copying. To address malicious infringement and copyright disputes of digital products, watermarking technology has been proposed as a copyright protection mechanism and has become an effective tool for solving such problems [1]. Digital audio watermarking technology embeds watermark information into audio signals while preserving auditory quality, thereby protecting digital audio products through copyright verification —ensuring transparency after watermark embedding [2].

However, existing approaches have limitations. Reference [3] proposed a zero-watermarking algorithm based on DCT coefficient signs, which demonstrates good robustness but lacks practical significance as audio features. Reference [4] introduced a DWT-SVD based zero-watermarking algorithm that requires audio length to exceed 1024 times the watermark image size, discarding any surplus audio and imposing strict requirements on test audio selection; furthermore, it shows weak robustness against jitter attacks. The DWT-SVD watermarking algorithm described in [5] shares a similar approach with [4] and exhibits good robustness against common attacks, but like [4], its extracted features cannot intuitively represent copyright information. Reference [6] presented a watermarking algorithm based on audio features and low-frequency coefficient minima, which requires selecting an appropriate threshold to discard partial audio frames, resulting in suboptimal stability and poor robustness against MP3 compression attacks. Reference [7] proposed a zero-watermarking algorithm for content authentication using low-frequency components as features, which demonstrates weak resistance to filtering attacks. Reference [8] developed a watermarking algorithm based on zero-crossing rate and short-term energy features of audio frames; although simple and intuitive, it shows poor robustness against noise and MP3 compression attacks.

Given that fundamental frequency features in audio exhibit stability and possess more intuitive practical meaning compared to various transform coefficients, this paper proposes a robust zero-watermarking algorithm based on audio fundamental frequency features. The algorithm first preprocesses the input audio

signal through wavelet approximation coefficient reconstruction, then calculates multiple fundamental frequencies per frame using the normalized subharmonic summation algorithm, and selects the frequency with maximum energy from the candidate fundamental frequency set. Since the fundamental frequency of signals is relatively stable, the resulting zero-watermark sequence demonstrates robust resistance to common attacks. As the fundamental frequency features of each frame are extracted independently without mutual influence, when audio suffers localized tampering, comparing the fundamental frequency sequences before and after tampering can identify the altered frames, with these frames constituting the tampered region. Consequently, the algorithm enables accurate detection of malicious tampering. [Figure 1: see original paper] illustrates the basic flow of feature extraction.

1.1 Zero-Watermarking

Zero-watermarking technology is a watermarking approach that does not require actual embedding of watermark information into the carrier signal. Instead, it generates a "zero-watermark" by combining feature information from the audio signal with the watermark image [9]. Therefore, zero-watermarking algorithms avoid the inherent conflict between imperceptibility and robustness present in traditional watermark embedding processes.

1.2 Musical Fundamental Frequency Feature

The fundamental frequency of music is the frequency at the lowest natural component (i.e., the fundamental tone) of the audio vibration system. It determines pitch and serves as one of the most basic features of audio signals, widely utilized in music information retrieval systems [10]. The primary reason for selecting fundamental frequency as the audio signal feature in this work is its stability and its capability to express audio content information.

1.3 Normalized Subharmonic Summation (NSHS) Algorithm

Based on the characteristic that audio harmonics are located at integer multiples of the fundamental frequency, the subharmonic summation (SHS) algorithm enhances the fundamental frequency energy by continuously accumulating energy from each harmonic to the fundamental frequency, thereby highlighting the fundamental tone [11]. The normalized subharmonic summation algorithm proposed by Hsu et al. [12] improves upon traditional SHS by multiplying each frequency by a coefficient $1/$, preventing excessive low-frequency energy accumulation caused by multiple superpositions in conventional methods. This achieves accurate fundamental frequency extraction without excessively increasing low-frequency and fundamental frequency energy, while reducing the influence of high-order harmonics on pitch detection [13,14]. The summation spectrum is defined as:

$$H(f) = \sum_{n=1}^{N_f} h_n \cdot P\left(\frac{f}{n}\right)$$

where $h_n = 0.84^{n-1}$, $N_f = \text{floor}\left(\frac{f_s}{f}\right)$, and f_s is the sampling frequency.

2 Audio Fundamental Frequency Feature Extraction

The primary task of the algorithm is to obtain multiple fundamental frequencies for each audio frame and then select the frequency with maximum energy from the candidate fundamental frequency set of each frame. Since fundamental frequency features are inherently stable, the zero-watermark sequence generated from these features exhibits good robustness against common attacks. As the fundamental frequency features of each frame are extracted independently, when audio suffers localized tampering, comparing the fundamental frequency sequences before and after tampering can identify altered frames, which constitute the tampered region. Thus, the algorithm enables tampering detection. [Figure 1: see original paper] describes the basic flow of feature extraction.

2.1 Preprocessing Since the fundamental frequencies of human voice and musical instruments typically reside in the low-to-mid frequency range, to reduce the impact of high-frequency information on fundamental frequency feature extraction, the original audio undergoes preprocessing via wavelet transform approximation coefficient reconstruction. Approximation coefficients represent the large-scale low-frequency components of the signal, which contain the most important information and basic characteristics of the signal. [Figure 2: see original paper] (top) shows the waveform of the original audio, while [Figure 2: see original paper] (bottom) displays the audio waveform after three-level wavelet transform approximation coefficient reconstruction. As shown in [Figure 2: see original paper], the reconstructed audio signal becomes smoother, with high-frequency components suppressed, thereby reducing noise effects and facilitating fundamental frequency feature extraction.

2.2 Fixed Frame Number Segmentation After preprocessing, the audio is segmented into frames with a fixed number of frames and overlapping. Overlapping segmentation is adopted because audio signal changes are continuous, and the characteristic parameters of segmented audio should transition smoothly. Overlapping frames enable gradual transitions between adjacent frames, preventing discontinuity in extracted feature parameters.

Fixed frame number segmentation is employed to minimize the impact of time-scale stretching on fundamental frequency feature extraction. Additionally, since zero-watermark construction requires extracting one feature point per frame to XOR with one bit of watermark information, the number of feature points must match the number of pixels in the watermark image. Therefore, the frame count is determined by the watermark image as a fixed value.

2.3 Fundamental Frequency Feature Extraction Each framed audio signal is transformed to the frequency domain via STFT, and the NSHS algorithm is applied to extract the candidate fundamental frequency set for each frame. Through NSHS, harmonics are shifted to the fundamental frequency, enhancing its energy. Consequently, the fundamental frequency energy peak becomes more prominent relative to other frequencies, facilitating feature extraction. [Figure 3: see original paper] (top) shows the spectrum of an original audio frame, while [Figure 3: see original paper] (bottom) displays the spectrum after normalized subharmonic summation. As shown in [Figure 3: see original paper], after NSHS, a distinct peak appears at the fundamental frequency (approximately 150 Hz), with non-peak portions remaining relatively stable and high-frequency energy significantly suppressed. Since the frequencies at local energy peaks constitute the candidate fundamental frequency set, peak detection and extraction are more effective on the harmonically summed signal. Moreover, [Figure 3: see original paper] demonstrates that energy at harmonics and other frequencies is lower than at the fundamental frequency. Therefore, for each NSHS-processed frame, the frequency with maximum energy in the extracted candidate set is selected as the frame' s fundamental frequency feature and added to the fundamental frequency feature sequence.

[Figure 4: see original paper] compares the fundamental frequency features F0 extracted under various attacks with those of the original audio. In the figure, solid lines represent original audio fundamental frequency features, while dashed lines represent features extracted from attacked audio. If attacked features match original features, the data points overlap, displaying only the original features; non-overlapping portions indicate deviations caused by attacks. Note that the extracted fundamental frequency features may not belong to the same sound source across frames. Selecting the strongest fundamental frequency avoids feature loss due to pauses in singing or instrumental performance, which could prevent operation with the zero-watermark image. Consequently, the proposed scheme may exhibit large fluctuations in certain frames, as shown in [Figure 9: see original paper].

As demonstrated in Figure 4: see original paper and (b), up-sampling and down-sampling have no effect on fundamental frequency features. Figure 4: see original paper-(f) show that while low-pass filtering, noise, MP3 compression, and jitter attacks introduce deviations, the attacked fundamental frequency feature curves remain largely consistent with the original overall. Thus, [Figure 4: see original paper] demonstrates that audio fundamental frequency features exhibit good stability under various attacks. Furthermore, since the fundamental frequency extracted from each frame is independent of others, tampering attacks only affect the tampered region' s features, leaving unaffected regions unchanged. This characteristic enables accurate tampering localization.

3 Watermark Embedding and Extraction

3.1 Generating Zero-Watermark Binary Sequence Based on Funda-

mental Frequency Features [Figure 5: see original paper] illustrates the generation process of the zero-watermark binary sequence, i.e., the watermark embedding process. Using a meaningful binary image of size P×Q as the watermark, where P and Q represent the pixel rows and columns respectively, and P×Q=n denotes the total pixel count, the watermark image can be expressed as I={w(a,b), 0 a P, 0 b Q}. The zero-watermark sequence generation proceeds as follows:

a) Convert the original two-dimensional watermark image into a one-dimensional sequence, yielding binary image I={I(i), 0 i n}.

b) Record the extracted fundamental frequency features as f(i), where i denotes frame number (0 i n) and n represents the total frame count. To express the overall trend of fundamental frequency variations, binary encoding (0,1) is adopted to represent features. This requires classifying features into two categories: one encoded as 1 and the other as 0. K-means clustering rapidly classifies fundamental frequency features f(i) into two classes C1 and C2, which are then encoded into a binary feature sequence K(i) according to equation (3):

$$K(i) = \begin{cases} 1, & f(i) \in C_1 \\ 0, & f(i) \in C_2 \end{cases}$$

c) Perform XOR operation between watermark image I and binary feature sequence K(i) to obtain binary sequence W containing features:

$$W(i) = I(i) \oplus K(i)$$

d) Record and store the resulting zero-watermark binary sequence W.

3.2 Zero-Watermark Image Extraction [Figure 6: see original paper] shows the zero-watermark image extraction process. Zero-watermark extraction does not require the original audio carrier; it only needs the fundamental frequency feature sequence K and binary watermark sequence W. The specific steps are similar to binary sequence generation:

a) Obtain the attacked fundamental frequency features f'(i) from the attacked audio signal (0 i n, where i is frame number and n is total frame count), and generate binary sequence K' (i) via K-means clustering.

b) XOR feature binary sequence K' (i) with the previously obtained zero-watermark binary sequence W to extract watermark image I' (i):

$$I'(i) = W(i) \oplus K'(i)$$

4 Experiments

The experimental section comprises robustness testing and tampering detection to evaluate both aspects of the zero-watermarking algorithm.

4.1 Robustness Testing Simulations were conducted using MATLAB 2010b on different types of test audio signals under various attacks: up-sampling (2× original sampling frequency fs), down-sampling (fs/2), low-pass filtering (cutoff frequency 5 kHz), additive white Gaussian noise (20 dB), MP3 compression (128 kbps), and jitter attack (100). Normalized correlation coefficient (NC) and bit error rate (BER) served as evaluation metrics. A meaningful 32×32 binary image was used as the watermark.

[Figure 7: see original paper] shows the watermark images extracted by the proposed algorithm under various attacks for a pop music test sample (sampling frequency fs=44,100 Hz, duration=27 s). The extracted watermark images demonstrate that the proposed zero-watermarking algorithm can recover clear watermark images under different attacks, proving strong stability of the fundamental frequency-based zero-watermarking algorithm against common attacks.

NC and BER values were calculated for watermark images extracted after various attacks. Higher NC and lower BER values indicate closer resemblance to the original watermark and stronger robustness. Experiments were performed on three different music genres (pop, light music, and rock) using the same audio segments, with robustness comparisons against algorithms from references [3] and [4]. Results are shown in -.

As shown in , for pop music, the proposed algorithm outperforms reference [3] in all robustness tests. Compared with reference [4], the proposed algorithm demonstrates stronger robustness under all conventional attacks except MP3 compression, particularly excelling in low-pass filtering and jitter attacks. indicates that for light music, the proposed algorithm maintains good robustness, while reference [3]' s algorithm shows poor stability under MP3 compression and jitter attacks, and reference [4]' s algorithm exhibits weak robustness against jitter attacks. demonstrates that for rock music, the proposed algorithm remains stable under various attacks, whereas reference [3]' s algorithm performs poorly against jitter attacks, and reference [4]' s algorithm shows very weak stability under MP3 compression and jitter attacks.

These results primarily stem from the inherent stability of fundamental frequency features and the K-means clustering-based encoding of these features in our algorithm, which reduces sensitivity to various attacks and further enhances watermark feature stability. Additionally, the fixed-frame-number overlapping segmentation method employed in this work features adjustable frame lengths according to audio variations, minimizing the impact of local time-scale stretching caused by attacks, particularly jitter attacks. Reference [3]'s non-overlapping segmentation method contributes to its inferior jitter attack resistance compared to our algorithm. Reference [4]' s algorithm requires both a fixed frame length

and frame count equal to watermark length, causing significant sampling point variations per frame when attacks alter the audio, with deviations increasing as frame count grows. Consequently, for all three tested music genres, the proposed fundamental frequency-based zero-watermarking algorithm exhibits superior robustness against both common and jitter attacks, with fundamental frequency features providing more practical and intuitive meaning for audio content compared to coefficient or sign features extracted in other works.

4.2 Tampering Detection and Localization Based on the characteristic that fundamental frequency features are extracted independently per frame, malicious tampering only alters the fundamental frequency features of the tampered portion. Therefore, comparing fundamental frequency features can confirm tampering occurrence. If certain frames show inconsistent features concentrated in a specific region, this indicates localized tampering, and the region can be identified as the tampered area.

[Figure 8: see original paper] (top) shows the original audio waveform, [Figure 8: see original paper] (middle) shows the waveform after muting the 5-6 s segment, and [Figure 8: see original paper] (bottom) shows the waveform after replacing the 2-3 s segment with the 5-6 s segment. To highlight the tampered portions, only the first 6 seconds are displayed.

Tampered fundamental frequency features and detection localization results are shown in [Figure 9: see original paper]. To magnify the tampered portions for observation, results for the first 500 frames are displayed. For clear comparison, [Figure 9: see original paper] shows both tampered features and detection localization in the same figure. Figure 9: see original paper (top) displays fundamental frequency features after muting frames 188-224 (approximately 5-6 s), while Figure 9: see original paper (bottom) shows the detection localization results on the audio signal. Figure 9: see original paper shows features after replacing frames 76-112 (2-3 s) with frames 188-224 (5-6 s) (top) and the corresponding tampering detection localization results (bottom).

As shown in Figure 9: see original paper, fundamental frequency features are lost at frames 188-224 under malicious attack 1 (muting), successfully localizing the muted portion. Figure 9: see original paper demonstrates that under malicious attack 2 (replacement), frames 76-114 are tampered and accurately localized. In summary, the proposed algorithm can effectively detect and locate tampered positions in audio signals under malicious tampering attacks.

5 Conclusion

The proposed algorithm constructs a binary watermark sequence by performing K-means clustering and encoding on fundamental frequency features extracted via the NSHS algorithm—features that represent audio content information—generating a 0-1 sequence matching the length of the dimension-reduced watermark image, and XORing it with the watermark image. The zero-watermarking

approach preserves original audio content, ensuring excellent imperceptibility. Experiments prove that due to the strong stability of audio fundamental frequency features, the constructed binary sequence exhibits robust resistance to various attacks. Furthermore, since extracted fundamental frequency features are independent of each other, only tampered regions are affected under malicious attacks. Experimental results demonstrate that the proposed scheme can achieve accurate tampering detection and localization.

References

[1] Kumar S, Dutta A. Performance analysis of spatial domain digital watermarking techniques [C]// Proc of IEEE International Conference on Information Communication and Embedded Systems. 2016: 1-4.
[2] Gao Hualing. Survey on key technologies of information hiding [J]. Electronic World, 2016 (9): 146-148.
[3] Sun Ruipeng, Xu Haitao. Audio zero-watermarking algorithm based on DCT coefficient signs [J]. Computer Technology and Development, 2014, 24 (5): 146-149.
[4] Cai Yongmei, Guo Wenqiang. Audio zero-watermarking algorithm based on DWT-SVD [J]. Computer Engineering and Design, 2014, 35 (1): 42-46.
[5] Gopalan K, Fu J. An imperceptible and robust audio steganography employing bit modification [C]// Proc of IEEE International Conference on Industrial Technology. 2015: 1635-1638.
[6] Yang Deguo, Li Zhi, Jiang Jindi. Watermarking algorithm based on audio features and low-frequency coefficient minima [J]. Computer Engineering, 2012, 38 (21): 10-13.
[7] Liu Guangyu, Zhang Xueying, Ma Chaoyang. Semi-fragile audio zero-watermarking algorithm for content authentication [J]. Computer Applications, 2012, 32 (4): 976-980.
[8] Wu Weina. Digital audio blind watermarking algorithm based on audio characteristic and scrambling encryption [C]// Advanced Information Technology, Electronic and Automation Control Conference. 2017: 1195-1198.
[9] Yang Yu, Lei Min, Cheng Mingzhi, et al. An audio zero-watermarking scheme based on energy comparison [J]. China Communications, 2014, 11 (7): 110-116.
[10] Zhang Xueyuan. Research on audio feature analysis methods for audio retrieval [D]. Guangzhou: South China University of Technology, 2015.
[11] Cheng T, Xu W, Tian Y, et al. Extracting singing melody in music with accompaniment based on harmonic peak and subharmonic summation [C]// Proc of Iet International Conference on Wireless, Mobile & Multimedia Networks. 2012: 200-205.
[12] Hsu C L, Chen L Y, Jang J S R, et al. Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement [C]// Proc of International Society for Music Information Retrieval Conference. 2009: 201-206.
[13] Ikemiya Y, Itoyama K, Yoshii K, et al. Singing voice separation and vocal

F0 estimation based on mutual combination of robust principal component analysis and subharmonic summation [J]. IEEE/ACM Trans on Audio Speech & Language Processing, 2016, 24 (11): 2084-2095.
[14] Song Yueyang. Research on music main melody extraction method based on single-source underdetermined speech separation [D]. Beijing: Beijing University of Posts and Telecommunications, 2012.

Note: Figure translations are in progress. See original paper for figures.