Some Evaluations on a Digital Watermarking Technique for Music Data Using Distortion Effect*

Yuto MATSUNAGA†, Nonmember, Tetsuya KOJIMA††, Naofumi AOKI†, Senior Members, Yoshinori DOBASHI†, and Tsuyoshi YAMAMOTO†, Members

SUMMARY We have proposed a novel concept of a digital watermarking technique for music data that focuses on the use of sound synthesis and sound effect techniques. This paper describes the details of our proposed technique that employs the distortion effect, one of the most common sound effects frequently utilized especially for guitar and bass instruments. This paper describes the experimental results of evaluating the resistance of the proposed technique against some basic malicious attacks utilizing MP3 coding, tempo alteration, pitch alteration, and high-pass filtering. It is demonstrated that the proposed technique potentially has appropriate resistance against such attacks except for the high-pass filtering attack. A technique for increasing the resistance against the high-pass filtering attack is also supplementarily discussed.

key words: digital watermark, sound synthesis, sound effect, distortion effect

1. Introduction

To protect the copyright of music data, several approaches based on digital watermarks have been investigated [1]. Some ideas are studied from the viewpoints of the musical arrangement knowledge [2], [3]. These approaches embed secret data by adding extra notes based on the chord theory of music composition.

Although such a technique can work well without much loss of sound quality in a sense of subjective viewpoints, the musical arrangement itself is considered to be a kind of degradation process from the original music data in a sense of objective viewpoints. As shown in Fig. 1 (a), the stego data is not the same anymore as the original music data in such a process [4]. To avoid this problem, it is better to consider other approaches that embed secret data at the stage of creating the original music data instead of employing postprocesses such as in the musical arrangement approach.

For this purpose, we have proposed a novel concept of a digital watermarking technique for music data that focuses on the use of sound synthesis and sound effect techniques, quite ordinary processes generally employed in music creation. The concept of the proposed technique guarantees that the stego data is equal to the original music data itself, so that there is no degradation caused by its embedding procedure.

As shown in Fig. 1 (b), if sound synthesis techniques are considered to be the embedding procedure, it is possible to create stego data which is equal to the original music data itself. As a pilot study, we have proposed a technique that employs M sequences for synthesizing sound of hi-hat cymbal tone. The proposed technique extracts the secret data by taking advantage of the signal detectability of M sequences [5].

Also, as shown in Fig. 1 (c), if sound effect techniques are considered to be the embedding procedure, it is possible to create stego data which is equal to the original music data itself. As a pilot study, we have proposed a technique that employs the distortion effect for controlling the timbre of guitar and bass instruments. The proposed technique extracts the secret data by detecting the statistical differences in the waveform of stego data [5], [6].

This paper describes the details of our proposed technique that employs the distortion effect, one of the most common sound effects frequently utilized especially for guitar and bass instruments.

To demonstrate the validity of the proposed technique,
this paper describes the experimental results of evaluating the resistance of the proposed technique against some basic malicious attacks utilizing MP3 coding, tempo alteration, pitch alteration, and high-pass filtering. A technique for increasing the resistance against the high-pass filtering attack, which is considered to be one of the critical attacks against the proposed technique, is also supplementarily discussed.

2. Digital Watermark Using Distortion Effect

The distortion effect is one of the most common sound effects for creating music data. It controls the timbre of musical instruments by adding extra harmonics utilizing nonlinear signal processing [7], [8].

Practically, there are several techniques for applying the distortion effect to music data. Among them, waveform clipping is considered to be one of the most common processes. It clips waveform when its absolute amplitudes exceed clipping levels. These thresholds may be independently chosen at the positive and negative sides of the waveform.

The clipping levels determine the statistical properties of the waveforms. For example, the average of the clipped waveform becomes smaller than 0 when the clipping level of the positive side is smaller than that of the negative side. On the other hand, the average of the clipped waveform becomes larger than 0 when the clipping level of the positive side is larger than that of the negative side. These differences may be employed as a feature for embedding binary information required for implementing a digital watermark.

3. Proposed Technique

Focusing on the above-mentioned statistical differences of the waveform processed by the distortion effect, it may be possible to propose a digital watermarking technique for music data. Figure 2 shows the overall procedure of the proposed technique. The details are described as follows.

3.1 Embedding Procedure

In the embedding procedure, the positive-biased clipping method is applied when the secret data is 1 and the negative-biased clipping method is applied when the secret data is 0. Each binary information is embedded into the waveform in a frame-by-frame manner.

To avoid perceptible artifacts which may be caused by switching between positive and negative-biased clipping method, the frame size should be properly chosen according to the tempo of the music data. It is defined as follows.

\[
N = \frac{60f_s}{T}
\]

where \(N\) denotes the frame size in sample, \(f_s\) the sampling frequency in Hz, and \(T\) the tempo of the music data in BPM (beat per minute), respectively.

The data rate and the payload of the secret data are defined as follows.

\[
R = \frac{T}{60} \quad (2)
\]

\[
P = RL \quad (3)
\]

where \(R\) denotes the data rate in bps (bits per second), \(T\) the tempo of the music data in BPM, \(P\) the payload in bits, and \(L\) the length of the music data in second, respectively.

The proposed technique applies the distortion effect using two mapping functions defined with hyperbolic tangent [9]. For two intensity constants \(a_p (0 < a_p < 0.5)\) and \(a_n (0 < a_n < 0.5)\), a gain constant \(g (g \gg 1)\), the secret data of the \(k\)th frame \(w(k)\), and the music data \(x(n)\), the stego data \(y(n)\) after the embedding procedure is defined as follows.

When \(w(k) = 1\), \(y(n)\) is defined with the positive-biased mapping function as

\[
y(n) = \begin{cases} 
A - \frac{A^2 - 1}{g}B & \text{if} \quad x(n) > 1 - a_p \\
C & \text{otherwise} \\
D - \frac{D^2 - 1}{g}E & \text{if} \quad x(n) < -a_p 
\end{cases} \quad (4)
\]

where

\[
A = \tanh(1 - a_p) 
\]

\[
B = \tanh(g(x(n) - (1 - a_p))) 
\]

\[
C = \tanh(x(n)) 
\]

\[
D = -\tanh(a_p) 
\]

\[
E = \tanh(g(x(n) + a_p)). 
\]

On the other hand, when \(w(k) = 0\), \(y(n)\) is defined with the negative-biased mapping function as

\[
y(n) = \begin{cases} 
A - \frac{A^2 - 1}{g}B & \text{if} \quad x(n) > 1 + a_p \\
C & \text{otherwise} \\
D - \frac{D^2 - 1}{g}E & \text{if} \quad x(n) < -a_p 
\end{cases} \quad (4)
\]

\[
A = \tanh(1 + a_p) 
\]

\[
B = \tanh(g(x(n) - (1 + a_p))) 
\]

\[
C = \tanh(x(n)) 
\]

\[
D = -\tanh(a_p) 
\]

\[
E = \tanh(g(x(n) + a_p)). 
\]
where

\[
y(n) = \begin{cases} 
    A - \frac{A^2 - 1}{g} B & (x(n) > a_n) \\
    C & \text{(otherwise)} \\
    D - \frac{D^2 - 1}{g} E & (x(n) < -(1 - a_n)) 
\end{cases}
\]

with

\[
A = \tanh(a_n) \\
B = \tanh(g(x(n) - a_n)) \\
C = \tanh(x(n)) \\
D = -\tanh(1 - a_n) \\
E = \tanh(g(x(n) + (1 - a_n))).
\]

The timbre of the distorted sound is mainly controlled by \(a_p\) and \(a_n\). If these constants become close to 0, more distorted sound may be obtained. In addition, the robustness of the proposed technique is also controlled by \(a_p\) and \(a_n\). If these constants become close to 0.5, two types of the clipping methods are difficult to distinguish. To increase the robustness, these constants should be close to 0.

3.2 Extraction Procedure

In the extraction procedure, the secret data is retrieved from the stego data in a frame-by-frame manner.

For a detection threshold \(d\) and the stego data \(y(n)\), the secret data of the \(k\)th frame \(w(k)\) after the extraction procedure is defined as follows.

\[
w(k) = \begin{cases} 
    1 & (s \geq d) \\
    0 & (s < d)
\end{cases}
\]

where

\[
s = \frac{1}{N} \sum_{n=0}^{N-1} y(n).
\]

In this procedure, \(d\) must be appropriately chosen according to \(a_p\) and \(a_n\). A reasonable choice for \(d\) is 0 when \(a_p\) and \(a_n\) are equally chosen.

4. Subjective Evaluation

The proposed technique switches between positive and negative-biased clipping methods. To confirm if this process does not cause perceptible artifacts, subjective evaluation was performed.

The demonstrated sound was generated from four guitar phrases recorded in mono channel with 44.1 kHz sampling frequency and 16 bits quantization level. These were (1) chord backing guitar phrase in fast tempo (168 BPM), (2) melodic solo guitar phrase in fast tempo (168 BPM), (3) chord backing guitar phrase in slow tempo (75 BPM), and (4) melodic solo guitar phrase in slow tempo (75 BPM), respectively.

The duration of each phrase was 8 bars, so that the total size of the secret data was 32 bits. Figure 3 illustrates the patterns of the secret data. These were (a) constant, (b) single, (c) alternative, and (d) random patterns, respectively.

In addition, to investigate the influence by different intensity of the distortion effect, \(a_p\) and \(a_n\) were chosen to be 0.1, 0.2, 0.3, or 0.4, where \(g\) was 100.

All the combination of the above-mentioned conditions generated 64 sound clips. Employing these sound clips, ABX tests were conducted. Seven participants were asked to listen to the phrase A and B and identify the phrase X as the phrase A or B. Either the phrase A or B was the constant pattern as the reference in each trial. All the phrases were demonstrated with a headphone (Audio-Technica, ATH-M20x) with moderate loudness to each participant.

Figures 4 and 5 show the experimental results summarized for each phrase or each pattern of the secret data. The correctness around 0.5 indicates that the participants could not recognize well the perceptible artifacts caused by switching between positive and negative-biased clipping methods. It was found that the influence by different intensity of the distortion effect was not remarkable.

ANOVA (analysis of variance) tests were conducted to confirm if the subjective evaluations depended on the phrases or the secret data. Since \(p\)-value of the tests was 0.92 for the phrases and 0.73 for the secret data, it was found that there was no statistical difference by the phrases or the secret data at the significance level of 0.05.
Fig. 4 Subjective evaluation summarized for each phrase. $a_p$ and $a_n$ range from 0.1 to 0.4.

Fig. 5 Subjective evaluation summarized for each pattern of the secret data. $a_p$ and $a_n$ range from 0.1 to 0.4.

These results indicate that almost no perceptible artifacts are recognized in cases where only the guitar phrases are evaluated. It should become more difficult to recognize such artifacts against the music data processed by the mixing procedure combining the guitar phrases with other instrument phrases.

5. Objective Evaluation

To examine the robustness of the proposed technique, objective evaluation was also performed.

The examined sound was generated from 116 guitar phrases recorded in mono channel with 44.1 kHz sampling frequency and 16 bits quantization level. These were obtained from the sound database of Cubase, a commercial music creation software [10]. These were chord backing guitar phrases and melodic solo guitar phrases obtained from several music genres consisting of popular and rock music. No phrases included silence part. The tempo of each phrase was between 62 and 180 BPM, so that the data rate of the secret data was between 1.03 and 3 bps as defined in Eq. (2).

The duration of each phrase was 8 bars, so that the total size of the secret data was 32 bits. The secret data was randomly generated binary sequences. To investigate the influence by different intensity of the distortion effect, $a_p$ and $a_n$ were chosen to be 0.1, 0.2, 0.3, or 0.4, where $g$ was 100.

The examined sound was generated by mixing the stego guitar phrases with other instrument phrases consist-

ing of bass and drum obtained from the sound database of Cubase. The same bass and drum phrases were adopted through all the mixing procedure.

In the following experiments, 5 different SNR (signal to noise ratio) of the guitar phrases were examined. The SNR was calculated for the condition that the guitar phrases were signal and the other instrument phrases were noise. The SNR around $-10$ dB was assumed to be the normal level, since the loudness of its guitar phrases was perceived as moderate.

In the extraction procedure, 0 was chosen for $d$ since both $a_p$ and $a_n$ were equally chosen.

5.1 Resistance against Mixing Procedure

First of all, the mixing resistance was examined to evaluate the detectability of the secret data when the guitar phrases were mixed with the other instrument phrases.

For the proposed technique, the mixing procedure itself is considered to be a sort of attack since the extraction procedure using average calculation is perturbed by the other instrument phrases combined with the guitar phrases. The robustness of the proposed technique depends on how much the average calculation in the extraction procedure is stable even after the mixing procedure. If the average of the stego data is not changed drastically by the mixing procedure, the resistance of the proposed technique is maintained.

Figure 6 shows the experimental result in the form of the BER (bit error rate) for each condition. The averages and 95% confidence intervals are plotted.

As the general tendency, the BER decreases when the SNR increases. Also, the BER decreases when both $a_p$ and $a_n$ decrease. It appears that the proposed technique works well when $a_p$ and $a_n$ are close to 0.

5.2 Resistance against Some Basic Malicious Attacks

To evaluate the resistance of the proposed technique against some basic malicious attacks, we performed a series of experiments. Five types of attacks were examined. These attacks utilizes (1) MP3 coding at a constant bit rate of 128 kbps, (2) tempo alteration of 10% acceleration, (3) tempo alteration of 10% deceleration, (4) pitch alteration by the increment of one semitone, and (5) high-pass filtering of
Fig. 7  Resistance against MP3 coding attack. $a_p$ and $a_n$ range from 0.1 to 0.4.

Fig. 8  Resistance against tempo acceleration attack. $a_p$ and $a_n$ range from 0.1 to 0.4.

Fig. 9  Resistance against tempo deceleration attack. $a_p$ and $a_n$ range from 0.1 to 0.4.

Fig. 10  Resistance against pitch alteration attack. $a_p$ and $a_n$ range from 0.1 to 0.4.

Fig. 11  Resistance against high-pass filtering attack. $a_p$ and $a_n$ range from 0.1 to 0.4.

20 Hz cut-off frequency, respectively. Each condition was evaluated by using the BER as the index.

The experimental results against these attacks are shown in Figs. 7 through 11. Except for the high-pass filtering attack, it appears that the proposed technique has appropriate resistance against such attacks when $a_p$ and $a_n$ are close to 0.

6. Discussion

The payload of the proposed technique depends on the tempo and the length of the music data as defined in Eq. (3). For example, it is 480 bits for the music data of 120 BPM tempo and 4 minutes length. Considering this limitation in capacity, it should be noted that the proposed technique may be suited to transmitting short messages such as in steganography applications.

As shown in this paper, the proposed technique may potentially work well even if it is processed by some basic malicious attacks. However, it is not fully guaranteed against other advanced attacks. For example, the high-pass filtering attack is considered to be one of the critical attacks against the proposed technique, since its secret data is in general embedded into the low-frequency components of music data. To increase the resistance against the high-pass filtering attack, further improvements are necessary.

One of the remedies may be the use of other statistical properties of the stego data in the extraction procedure. As shown in Fig. 12, even after the high-pass filtering attack, the envelope of the stego data does not change drastically. It seems that the secret data can be estimated by detecting the envelope of the stego data.

From this viewpoint, the extraction procedure may be modified as shown in Fig. 13. The secret data of the $k$th frame $w(k)$ after the extraction procedure is defined as follows.

$$w(k) = \begin{cases} 1 & (s \geq d) \\ 0 & (s < d) \end{cases}$$  \hspace{1cm} (18)

where

$$s = \frac{1}{N} \sum_{n=0}^{N-1} (e_p(n) + e_n(n)).$$  \hspace{1cm} (19)
In this modification, $e_p(n)$ and $e_n(n)$ denote the positive and negative envelopes of the stego data, respectively. Both the envelopes are obtained by smoothing the stego data.

To increase the performance furthermore, both the envelopes should be truncated with appropriate thresholds $T_p$ and $T_n$ as follows.

\[
\begin{align*}
    e'_p(n) &= \begin{cases} 
        e_p(n) & (|e_p(n)| \geq T_p) \\
        0 & (|e_p(n)| < T_p)
    \end{cases} & (20) \\
    e'_n(n) &= \begin{cases} 
        e_n(n) & (|e_n(n)| \geq T_n) \\
        0 & (|e_n(n)| < T_n)
    \end{cases} & (21)
\end{align*}
\]

A reasonable choice for each threshold is the average of each envelope.

To evaluate the effectiveness of this modification, we performed another experiment using the same 116 sound clips employed in the objective evaluation. As shown in Fig. 14, it appears that the resistance against the high-pass filtering attack may be improved with this modification especially when $a_p$ and $a_n$ are close to 0.

7. Conclusions

In this paper, we have mentioned a novel concept of a digital watermarking technique for music data. The proposed technique embeds secret data at the stage of creating the original music data. In the proposed technique, the secret data is therefore considered to be a part of the original music data. Consequently, the proposed technique is free from the degradation that is inevitably caused by using other techniques based on postprocesses such as in the musical arrangement approach.

The experimental results show that the proposed technique potentially has appropriate resistance against some basic malicious attacks utilizing MP3 coding, tempo alteration, and pitch alteration. Also, as discussed in this paper, the resistance of the proposed technique against high-pass filtering attack may be improved by detecting the envelope of the stego data.

The main concern of the proposed technique is its limitation. It should be noted that the proposed technique is difficult to employ in principle if there is silence part where distorted sound is not present. Sound detection algorithms such as in voice activity detection techniques may be required for employing the proposed technique in practical applications.

The proposed technique may potentially work well if distorted sound is required in music creation. However, distorted sound is not always applicable to all the music genres. Thus, the effectiveness of the proposed technique depends on the type of music data.

In addition, the distortion effect may be switched on and off by the guitar players in practical cases. Also, the same intensity of the distortion effect is not always applied through an entire piece of music data. It should be noted that the proposed technique is difficult to employ in principle if the distortion effect is switched off. From the experimental results of the objective evaluation described in this paper, the robustness of the proposed technique decreases inevitably...
if the intensity of the distortion effect is not appropriately maintained.

To discuss the availability of the proposed technique in more detail, further verification is ongoing. In addition, expanding the proposed concept by employing other types of sound effect techniques is also under consideration.

References

[1] K. Kondo, Multimedia Information Hiding Technologies and Methodologies for Controlling Data, Information Science Reference, Hershey, PA, USA, 2013.

[2] H. Murata and A. Ogihara, “An audio watermarking method using tones of diatonic chord,” Proc. IHMSP 2015, Adelaide, Australia, pp.156–159, 2015.

[3] K. Kamada, T. Kojima, and U. Parampalli, “Tone code: A novel method for covert communications based on musical components,” Proc. ISITA 2016, Monterey, USA, pp.340–344, 2016.

[4] F.A.P. Petitcolas, R.J. Anderson, and M.G. Kuhn, “Information hiding – A survey,” Proc. IEEE, vol.87, no.7, pp.1062–1078, 1999.

[5] Y. Matsunaga and T. Kojima, “A digital watermarking technique based on musical characteristics and distortion effects,” IEICE Technical Report, EMM2016–78, 2017. (in Japanese)

[6] Y. Matsunaga, T. Kojima, N. Aoki, Y. Dobashi, and T. Yamamoto, “A digital watermarking technique for music data using distortion effect,” Acoustical Science and Technology, vol.39, no.1, pp.37–39, 2018.

[7] A.U. Case, Sound FX, Focal Press Elsevier, New York, 2007.

[8] U. Zölzer, DAFX: Digital Audio Effects, John Wiley & Sons, Baffins Lane, 2002.

[9] F. Eichas and U. Zölzer, “Black-box modeling of distortion circuits with block-oriented models,” Proc. Digital Audio Effects (DAFx-16), Brno, Czech Republic, pp.39–45, 2016.

[10] https://www.steinberg.net/en/products/cubase/start.html

Yuto Matsunaga received B.E. degree from Hokkaido University in 2018. He is a student at Graduate School of Information Science and Technology, Hokkaido University. He is engaged in the research on audio processing.

Naofumi Aoki received Ph.D. degree from Hiroshima University in 2000. He is an assistant professor at Graduate School of Information Science and Technology, Hokkaido University. He is engaged in the research on audio processing.

Yoshinori Dobashi received Ph.D. degree from Hiroshima University in 1997. He is an associate professor at the Graduate School of Information Science and Technology, Hokkaido University. He is engaged in the research on computer graphics.

Tsuyoshi Yamamoto received Ph.D. degree from Hokkaido University in 1986. He is a professor at Graduate School of Information Science and Technology, Hokkaido University. He is engaged in the research on information processing.

Tetsuya Kojima received Ph.D. degree from Hokkaido University in 1997. He is a professor at National Institute of Technology, Tokyo College. He is engaged in the research on information theory.