Journal of Robotics, Networking and Artificial Life
Vol. **6**(**2**); September (2019), *pp.* 133–137
DOI: https://doi.org/10.2991/jrnal.k.190828.013; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

An Authentication Method for Digital Audio using Wavelet Transform and Fundamental Frequencies

Yasunari Yoshitomi[1,*], Shohei Tani[2], Masaki Arasuna[3], Ryota Kan[4], Taro Asada[1], Masayoshi Tabuse[1]

[1]*Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, 1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan*
[2]*Fukuchiyama City Hall, 13-1 Miki, Fukuchiyama, Kyoto Prefecture 620-8501, Japan*
[3]*Nissay Information Technology Co., Ltd., 5-37-1 Kamata, Ohta-ku, Tokyo 144-8721, Japan*
[4]*Shimazu Business Systems Co., Ltd., 1 Kuwahara-cho Nishinokyo, Nakagyo-ku, Kyoto 604-8442, Japan*

ARTICLE INFO

Article History

Received 18 October 2018
Accepted 05 December 2018

Keywords

Authentication
audio
copyright protection
wavelet transform
fundamental frequency

ABSTRACT

Several digital watermarking techniques for audio files have been proposed for hiding data for protecting their copyrights. There is a tradeoff between the quality of watermarked audio and the tolerance of watermarks to signal processing methods, such as compression. To overcome the inevitable tradeoff, we previously developed an authentication method for digital audio. We have improved the method by determining the region to be authenticated in the audio data by making effective use of the fundamental frequency characteristics.

1. INTRODUCTION

Recent progress in digital media technology and distribution systems, such as the Internet and cellular phones, has enabled consumers to easily access, copy, and modify digital audio. Several Digital Watermarking (DW) techniques for audio files have been proposed for hiding data for protecting their copyrights. There is generally a tradeoff between the quality of watermarked audio and the tolerance of watermarks to signal processing methods, such as compression.

In previous research [1], to essentially overcome this issue, we developed an authentication method for digital audio to protect the copyrights. In contrast to DW, no additional information is inserted into the original audio by the previously proposed method, and the digital audio is authenticated using features extracted using a Discrete Wavelet Transform (DWT) and characteristic coding of the previously proposed method [1]. However, in the previously proposed method [1], it is indispensable to determine the region to be authenticated in the audio data by using the fixed length and the fixed starting time from the beginning of the audio data. Therefore, the authentication tolerance to clipping of the audio data is essentially insufficient for practical use.

In the present study, to overcome this issue, we have improved the method by determining the region to be authenticated in the audio data by using the fundamental frequency characteristics.

2. OBSERVED PHENOMENON UNDERPINS THE AUTHENTICATION METHOD

The procedure and algorithm of our previously proposed method [1] is reviewed in this section, because it is very important for the present study.

It has been observed that when a DWT is applied to audio data, in the histogram of the wavelet coefficients of the Multi-Resolution Representation (MRR), the center of the distribution is very close to zero [2]. We exploited this phenomenon to develop an authentication method for audio data [1]. For further information on the DWT used, see Inoue and Yoshitomi [3] and Taniguchi and Yoshitomi [4].

3. AUTHENTICATION RATIO

We set the authentication parameters as described below [1,5].

In Figure 1, Th'(minus) was chosen so that it divides the nonpositive wavelet coefficients (S'_m in total frequency) into two equal groups, and similarly Th'(plus) was chosen so that it divides the positive wavelet coefficients (S'_p in total frequency) into two equal groups. Next, the values of the parameters $T1'$–$T4'$, which control the authentication precision, were chosen such that the following conditions were satisfied:

(1) $T1' < Th'(\text{minus}) < T2' < 0 < T3' < Th'(\text{plus}) < T4'$.

[*]*Corresponding author. Email:* yoshitomi@kpu.ac.jp

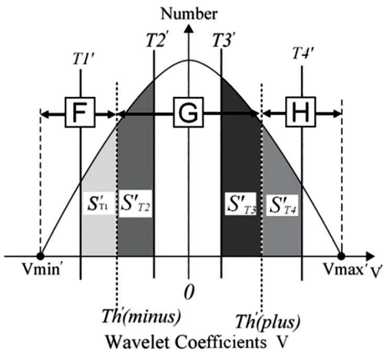


Figure 1 | Three sets (*F*, *G*, and *H*) of MRR wavelet coefficients used for authentication [1].

(2) The value of S'_{T1}, the number of wavelet coefficients in ($T1'$, Th'(minus)), is equal to S'_{T2}, the number of wavelet coefficients in [Th'(minus), $T2'$), i.e., $S'_{T1} = S'_{T2}$.

(3) The value of S'_{T3}, the number of wavelet coefficients in ($T3'$, Th'(plus)], is equal to S'_{T4}, the number of wavelet coefficients in (Th'(plus), $T4'$), i.e., $S'_{T3} = S'_{T4}$.

(4) $S'_{T1} / S'_m = S'_{T3} / S'_p$.

In the present study, the values of both S'_{T1} / S'_m and S'_{T3} / S'_p are set to 0.3, which is the same setting used for creating the code for the original audio data [1]. When preparing the authentication codes, the wavelet coefficients V' for each MRR sequence are divided as shown in Figure 1 into three sets, which are defined as follows:

- $F = \{V' | V' \in V'^{AC}, V' < Th'(minus)\}$
- $G = \{V' | V' \in V'^{AC}, Th'(minus) \le V' \le Th'(plus)\}$
- $H = \{V' | V' \in V'^{AC}, Th'(plus) < V'\}$,

where V'^{AC} is the set of wavelet coefficients from the target audio data that is used to create the authentication code.

The wavelet coefficients V'_i are then classified according to the following rules with the flags f_i used in creating the original code C:

When $f_i = 1$ and $V'_i \in G$, b'_i is set to 0.

When $f_i = 1$ and $V'_i \in (F \cup H)$, b'_i is set to 1.

When $f_i = 0$, b'_i is set to 0.5.

Note that the value 0.5 can be chosen arbitrarily, since the value of b_i that is the bit for creating the code for the original audio data [1] does not influence the method's performance. Finally, this sequence of b'_i values is used to form the authentication code C'.

The authentication ratio AR(%) is defined as follows [Equation (1)]:

$$AR = \frac{100\sum_{i=1}^{N} f_i(1-|b_i - b'_i|)}{\sum_{i=1}^{N} f_i}, \quad (1)$$

where N is the number of wavelet coefficients chosen to create the authentication code for the original audio data [1]. As can be seen in Equation (1), neither b_i nor b'_i influences the value of Authentication Ratio (AR) when $f_i = 0$, which occurs when the corresponding V_i that is the wavelet coefficient of the original audio is not selected for coding in the original audio data [1].

To use the proposed method, we need to store the flags f_i and the original code C for each copyrighted file that we want to protect. When calculating (1) to authenticate audio data, we do not use the original audio data; instead, we use the flags f_i and the code C for that file [1].

4. FUNDAMENTAL FREQUENCY CHARACTERISTICS OF AUDIO DATA

Figure 2 shows the fundamental frequency of the first entry of the rock music genre category in the music database RWC for research purposes [6]. As shown, several local maximums of the fundamental frequency exist within the stream of the music. We have proposed a method for determining the region to be authenticated in the audio data by using local maximums of the fundamental frequency of the audio data, as described in the next section.

5. PROPOSED METHOD FOR DETERMINING THE REGION TO BE AUTHENTICATED IN THE AUDIO DATA

The audio data clipping procedure for the authentication is as follows:

Step 1: The fundamental frequency $f(i)(1, 2, 3, ..., n)$ at the start time index i is measured every 0.01 s from the beginning to the end of T s of the original audio data, i.e., $n = 100T$. Then, the absolute value of difference $ADF(i) = |f(i + 1) - f(i)|$ is calculated for all $i(i = 1, 2, 3, ..., n - 1)$.

Step 2: The sum $S(i) = \sum_{j=0}^{999} ADF(i + j)$ is calculated for all $i(i = 1, 2, 3, ..., n - 999)$.

Step 3: The start time indexes i are determined by whether the value of $S(i)$ is among the top 10 of all $S(k)$ under the restriction $|i - j| \ge 1000$ for all j such that $S(i) < S(j)$.

Step 4: For the start time indexes i selected in Step 3, 10 s of audio data are clipped for the authentication.

6. EXPERIMENTS

6.1. Conditions

An experiment was performed in the following computational environment: the personal computer was a DELL OPTIPLEX CF-SX1

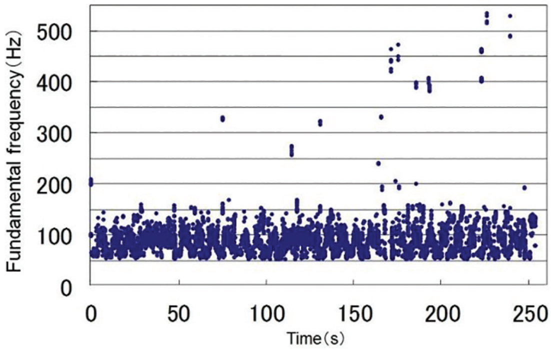


Figure 2 | Fundamental frequency of audio data.

(CPU: Core i7-2600 Duo, 3.40 GHz; main memory: 4.0 GB); the OS was Microsoft Windows XP; the development language was Microsoft Visual C++ 6.0.

Five music audio files, namely, the first entry of each of five genre categories—classical, jazz, popular, rock, and hiphop—in the music database RWC used for research purposes [6], were copied from CDs onto a personal computer as WAVE files with the following specifications: 44.1 kHz, 16 bits, and monaural. For each music audio file selected from the database, 10 sets of 10-second clips of music audio were produced using the proposed method described in Section 5.

For investigating the authentication tolerance to clipping, audio test data were produced by clipping one region from each music audio file selected from the database. The clipped regions were specified by all combinations of the lengths 0.01, 0.1, 0.5, 1.0, 2.0, 3.0, and 4.0 s and the starting times 0, 5, and 10 s (from the beginning of the original audio data), resulting in 21 clipping conditions. Then, for each audio test data file produced from each original audio data file, 10 sets of 10-s clips of music audio were produced using the proposed method described in Section 5, and the authentication procedure was performed using the previously proposed method [1]. The highest value of the authentication ratio of the original audio data to the audio test data (hereinafter referred to as HAR) of AR as described in Section 3 among those of the 100 combinations of the 10 clips of the original data and the 10 clips of test data was calculated.

As another method for comparison (hereinafter referred as AMFC), we chose audio data from the classical music genre category and produced one 10-s clip whose starting time gave the highest fundamental frequency for 0.01 s among those in the audio data. Next, the clipping regions from the beginning of the original audio data were specified by the lengths 0.00001, 0.00005, 0.0001, 0.0002, 0.0003, 0.0004, 0.0005, 0.001, and 0.01 s. Then, the calculation of AR for the original audio data and the audio data after clipping by the above each length was performed for one 10-s clip whose starting time from the beginning of the audio data was decided by the above method for the original audio data. For the DWT, we used Daubechies wavelets. Level 8 was chosen based on an analysis of preliminary experiments [1].

6.2. Results and Discussion

Tables 1 and 2 respectively show the AR values for AMFC and the HAR values for the proposed method. As shown in Table 1, AMFC showed poor authentication tolerance to clipping: clipping 0.01 s of audio caused AR to decline to 65.88%. On the other hand, as shown in Table 2, the authentication tolerance to clipping of the audio data was remarkably improved by adopting the proposed method.

Table 1 | AR with AMFC for clipping from the beginning of original classical audio data

Clipping length (s)	AR (%)
0.00001	100
0.00005	100
0.0001	99.22
0.0002	96.84
0.0003	87.75
0.0004	84.98
0.0005	78.66
0.001	69.57
0.01	65.88

Table 2 | HAR (%) with the proposed method under the 21 clipping conditions. Starting time for the test data from the beginning of the original audio data: (a) 0, (b) 5, and (c) 10 s

		Music				
		Classical	Jazz	Popular	Rock	Hiphop
		(a)				
Clipping length (s)	0.01	100	100	100	100	100
	0.1	100	100	100	100	85.49
	0.5	98.82	99.60	99.17	99.61	78.04
	1.0	98.08	98.81	92.24	98.43	80.63
	2.0	73.12	86.67	99.17	95.29	80.63
	3.0	67.19	89.33	100	93.37	80.63
	4.0	94.25	90.40	95.01	69.41	71.76
		(b)				
Clipping length (s)	0.01	95.62	98.82	96.40	99.72	94.90
	0.1	96.16	94.12	92.52	87.06	90.91
	0.5	82.75	89.02	93.91	84.71	83.40
	1.0	91.78	80.63	77.25	77.47	75.29
	2.0	86.03	74.67	75.10	95.44	73.52
	3.0	79.73	72.19	85.77	77.65	86.17
	4.0	64.38	71.76	76.28	74.12	73.12
		(c)				
Clipping length (s)	0.01	100	100	100	100	100
	0.1	96.44	100	100	99.61	100
	0.5	97.65	99.61	98.42	97.63	99.60
	1.0	93.70	86.27	96.08	95.65	94.47
	2.0	88.77	66.40	91.70	93.73	88.54
	3.0	80.00	73.33	80.29	87.84	91.70
	4.0	64.84	73.87	79.45	79.61	76.28

Furthermore, when 5-s clip was used for narrowing the 10-s clip decided as the region to be authenticated in the audio data, HAR was much improved, being almost 100% for all the clipping conditions used in this experiment. Even for a clipping length of 4.0 s, HAR improved to 100% except under one condition for classical audio data: clipping from a starting time of 5 s from the beginning of the original audio data. In the exceptional case, HAR was 76.25%.

7. CONCLUSION

In general, there is a tradeoff between the quality of watermarked audio and the tolerance of watermarks to signal processing methods, such as compression. To overcome this inevitable tradeoff, we previously developed an authentication method [1] for digital audio using a DWT. In the present study, we have improved the method by determining the region to be authenticated in the audio file by using the fundamental frequency characteristics. The experimental results show that the method has a high authentication tolerance to clipping small parts from the audio data.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

REFERENCES

[1] Y. Yoshitomi, T. Asada, Y. Kinugawa, M. Tabuse, An authentication method for digital audio using a discrete wavelet transform, J. Inform. Sec. 2 (2011), 59–68.

[2] S. Murata, Y. Yoshitomi, H. Ishii, Optimization of embedding position in an audio watermarking method using wavelet transform, in: Abstracts of The 2007 Fall National Conference of ORSJ, Tokyo, Japan, 2007, pp. 210–211 (in Japanese).

[3] D. Inoue, Y. Yoshitomi, Watermarking using wavelet transform and genetic algorithm for realizing high tolerance to image compression, J. IIEEJ 38 (2009), 136–144.

[4] T. Taniguchi, Y. Yoshitomi, Method for character domain extraction from image using wavelet transform, J. Robot. Netw. Artif. Life 2 (2015), 103–106.

[5] R. Fujii, Y. Yoshitomi, T. Asada, M. Tabuse, An authentication method using a discrete wavelet transform for a recaptured video, J. Robot. Netw. Artif. Life 3 (2016), 107–110.

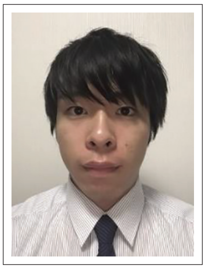
[6] M. Goto, H. Hashiguchi, T. Nishimura, R. Oka, RWC music database: database of copyright-cleared musical pieces and instrument sounds for research purposes, Trans. IPSJ 45 (2004), 728–738 (in Japanese).

Authors Introduction

Dr. Yasunari Yoshitomi

He received his B.E, M.E. and PhD degrees from Kyoto University in 1980, 1982 and 1991, respectively. He works as a Professor at the Graduate School of Life and Environmental Sciences of Kyoto Prefectural University. His specialties are applied mathematics and physics, informatics environment, intelligent informatics. IEEE, HIS, ORSJ, IPSJ, IEICE, SSJ, JMTA and IIEEJ member.

Mr. Masaki Arasuna



He received his B.S. degree from Kyoto Prefectural University in 2017. He works at Nissay Information Technology Co., Ltd.

Mr. Shohei Tani

He received his B.S. degree from Kyoto Prefectural University in 2013. He works at Fukuchiyama City Hall.

Mr. Ryota Kan

He received his B.S. degree from Kyoto Prefectural University in 2017. He works at Shimazu Business Systems Co., Ltd.

Dr. Taro Asada

He received his B.S., M.S. and PhD degrees from Kyoto Prefectural University in 2002, 2004 and 2010, respectively. He works as an Associate Professor at the Graduate School of Life and Environmental Sciences of Kyoto Prefectural University. His current research interests are human interface and image processing. HIS, IIEEJ member.

Dr. Masayoshi Tabuse

He received his M.S. and PhD degrees from Kobe University in 1985 and 1988 respectively. From June 1992 to March 2003, he had worked in Miyazaki University. Since April 2003, he has been in Kyoto Prefectural University. He works as a Professor at the Graduate School of Life and Environmental Sciences of Kyoto Prefectural University. His current research interests are machine learning, computer vision and natural language processing. IPSJ, IEICE and RSJ member.