Journal of Robotics, Networking and Artificial Life, Vol. 3, No. 3 (December 2016), 193-196

A Method for Secure Communication Using a Discrete Wavelet Transform for Audio Data

Yuji Tsuda[1], Kouhei Nishimura[2], Haruka Oyaizu[3], Yasunari Yoshitomi[2], Taro Asada[2], and Masayoshi Tabuse[2]
1: Software Service, Inc.
Nishi-Miyahara, Yodogawa-Ku, Osaka 532-0004, Japan
2: Graduate School of Life and Environmental Sciences, Kyoto Prefectural University,
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan
E-mail: {yoshitomi, tabuse}@kpu.ac.jp, t_asada@mei.kpu.ac.jp}
http://www2.kpu.ac.jp/ningen/infsys/English_index.html
3: NHK Media Technology, Inc.
Kamiyama-cho, Shibuya-ku, Tokyo 150-0047, Japan

Abstract

We developed a secure communication method using a discrete wavelet transform. Two users must both have a copy of the same piece of music to be able to communicate with each other. The music and the sender's message are encoded using the scaling coefficients obtained from a discrete wavelet transformation. The message receiver can produce the audio data similar to the sending user's speech using an inverse discrete wavelet transform, together with information on the difference between these two codes.

Keywords: Secure communication, Audio data processing, Wavelet transform, Encoding.

1. Introduction

Recently, there has been an increase in certain kinds of fraud. Specifically, the elderly are often the targets of telephone fraud. The fraudster pretends to be a grandchild of the elderly person while talking on the phone, and appeals to the elderly person to send money, for example, through a bank transfer. When the elderly person mistakes the fraudster for a grandchild, the fraudster can obtain money. Even if the voice of the fraudster is not similar to that of the grandchild, the elderly victim might send money to the fraudster. This is because the fraudster skillfully convinces the elderly person, who cherishes their real grandchild, of a serious monetary problem, such as might exist following a traffic accident.

In the present study, we propose a method for secure communication using a discrete wavelet transform (DWT). The method can be used with Internet protocol (IP) telephones, and has the potential to help prevent telephone fraud.

2. Wavelet Transform

In this section, we provide a brief introduction to the DWT.[1]

Original audio data $s_k^{(0)}$,which are used as the level-0 wavelet decomposition coefficient sequence, where k denotes the element number, are decomposed into the elements of a multi-resolution representation (MRR) and the elements of a multi-resolution analysis (MRA) by repeatedly applying the DWT. The wavelet decomposition coefficient sequence $s_k^{(j)}$ at level j is decomposed into two wavelet decomposition coefficient sequences at level $j+1$ using the following equations:

$$s_k^{(j+1)} = \sum_n \overline{p_{n-2k}} s_n^{(j)} \tag{1}$$

$$w_k^{(j+1)} = \sum_n \overline{q_{n-2k}} s_n^{(j)}, \tag{2}$$

where p_k and q_k denote the scaling and wavelet sequences, respectively, and $w_k^{(j+1)}$ denotes the development coefficient at level $j+1$. The development coefficients at level J are obtained using (1) and (2) iteratively from $j = 0,...,J-1$. Figure 1 shows the process for the multi-resolution analysis by DWT.

In the present study, we use the Daubechies wavelet for the DWT.[2] As a result, we obtain the following relation between p_k and q_k:

$$q_k = (-1)^k p_{1-k} \qquad (3)$$

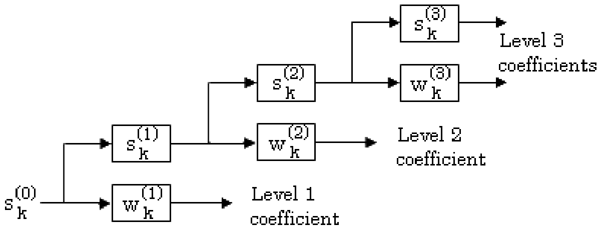

Fig. 1. Multi-resolution analysis by DWT.[1]

3. Proposed Method

3.1. *Encoding*

3.1.1. *Phenomenon exploited for the coding algorithm for audio data*

It is known that the histogram of the wavelet coefficients for each domain of the MRR sequences is centered at approximately zero when the DWT is performed on audio data.[1] In the present study, we found that the histogram of the scaling coefficients for each domain of the MRA sequences is also centered at approximately zero when the DWT is performed on audio data. Exploiting this phenomenon, we have developed a secure communication method using audio data.

3.1.2. *Parameter settings*

As with digital watermark (DW) techniques for images[2,3] and digital sounds,[4] we set the following coding parameters.

The values of $Th(\text{minus})$ and $Th(\text{plus})$ in Fig. 2 are chosen such that the non-positive scaling coefficients (S_m in total frequency) are equally divided into two groups by $Th(\text{minus})$, and the positive scaling coefficients (S_p in total frequency) are equally divided into two groups by $Th(\text{plus})$. Next, the values of *T1, T2, T3* and *T4,* which are the parameters for controlling the authentication precision, are chosen to satisfy the following conditions:

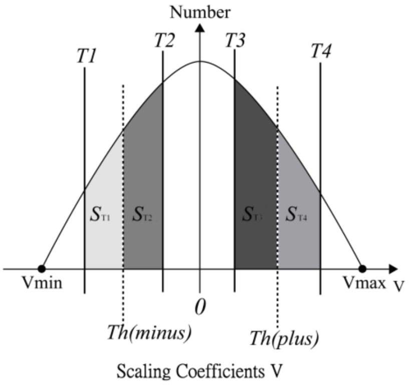

Fig. 2. Schematic diagram of the histogram of the MRA scaling coefficients.

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

2) The value of S_{T1}, which is the number of scaling coefficients in $(T1, Th(\text{minus}))$, is equal to S_{T2}, which is the number of scaling coefficients in $[Th(\text{minus}), T2)$, i.e., $S_{T1} = S_{T2}$.

3) The value of S_{T3}, the number of scaling coefficients in $(T3, Th(\text{plus})]$, is equal to S_{T4}, the number of scaling coefficients in $(Th(\text{plus}), T4)$, 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 was determined experimentally.

3.1.3. *Encoding*

In the preprocessing of the audio data before encoding, the scaling coefficients V of an MRA sequence are separated into five sets (hereinafter referred to as G_0, G_1, G_2, G_3, and G_4), as shown in Fig. 3, under the following criteria:

- $G_0 = \{V \mid V \in V^{SC}, V \le T1\}$,
- $G_1 = \{V \mid V \in V^{SC}, T1 < V < T2\}$,
- $G_2 = \{V \mid V \in V^{SC}, T2 \le V \le T3\}$,
- $G_3 = \{V \mid V \in V^{SC}, T3 < V < T4\}$,
- $G_4 = \{V \mid V \in V^{SC}, T4 \le V\}$,

where V^{SC} is the set of scaling coefficients in the audio data file.

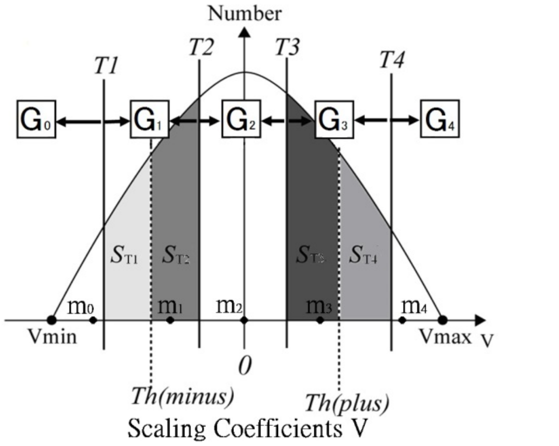

Fig. 3. Schematic diagram for demonstrating the selection of the scaling coefficients for encoding the audio data.

The scaling coefficients for an MRA sequence are encoded according to the following rules, in which V_i denotes one scaling coefficient:

When $V_i \in G_0$, $c_i = 0$.
When $V_i \in G_1$, $c_i = 1$.
When $V_i \in G_2$, $c_i = 2$.
When $V_i \in G_3$, $c_i = 3$.
When $V_i \in G_4$, $c_i = 4$.

Then, the representative value for each set, G_0, G_1, G_2, G_3, and G_4, is its average, m_0, m_1, m_2, m_3 and m_4, respectively. For audio data formation, we use a code $\boldsymbol{C}$ (hereinafter referred to as an original code), which is the sequence of c_i, and m_j defined above.

3.2. *Sound data formation using code replacement*

The scaling coefficient sequence for audio data, A, is expressed by

$$S(A)_k = \{x_1, x_2, x_3, \ldots, x_k\},$$

where k is the total number of scaling coefficient of A at a level. Then, a sequence

$$C(A)_k = \{X_1, X_2, X_3, \ldots, X_k\}$$

is determined, where $X_i \in \{0,1,2,3,4\}$ is the element index denoting which of the five sets of scaling coefficients x_i of A belongs to.

Next, the audio data A' is defined as having the scaling coefficient sequence $S(A')_k$ and a value of zero for all wavelet coefficient values at every level. $S(A')_k$ is defined as:

$$S(A')_k = \{a_1, a_2, a_3, \ldots, a_k\},$$

where $a_i \in \{m_0^A, m_1^A, m_2^A, m_3^A, m_4^A\}$ is the average of the scaling coefficients of A at the range denoted by $X_i \in \{0,1,2,3,4\}$ obtained from A.

Then, the audio data B'_A is defined as having the scaling coefficient sequence $S(B'_A)_k$ and a value of zero for all wavelet coefficient values at every level. $S(B'_A)_k$ is defined as:

$$S(B'_A)_k = \{b_{A,1}, b_{A,2}, b_{A,3}, \ldots, b_{A,k}\},$$

where $b_{A,i} \in \left\{m_0^B, m_1^B, m_2^B, m_3^B, m_4^B\right\}$ is the average of scaling coefficients of B at the range denoted by $X_i \in \{0,1,2,3,4\}$ obtained from A.

$S(B'_A)_k$ is obtained by replacing Y_i with X_i when $Y_i \neq X_i$, and then replacing b_i with $b_{A,i}$, where b_i is the average of the scaling coefficients of B at the range denoted by Y_i. Therefore, $C(B'_A)_k = C(A)_k$. As a result, B'_A is expected to be similar to A.

3.3. *Data for communication*

A sequence $D1(B'_A)_n$ is defined as:

$$D1(B'_A)_n = \{z_1, z_2, \ldots, z_n\},$$

where n is the total number of cases where $Y_i \neq X_i$, and $z_p = [\![|y_i|]\!] \bmod 256$, and the integer p is increased from 1 to n, in steps of size 1, when $Y_i \neq X_i$. Here, $[x]$ is the maximum integer that is not greater than x. Then, a sequence $D2(B'_A)_n$ is defined as:

$$D2(B'_A)_n = \{Z_1, Z_2, \ldots, Z_n\},$$

where n is the total number of cases where $Y_i \neq X_i$, and $Z_p = X_i$, in which the integer p is increased from 1 to n, in steps of size 1, when $Y_i \neq X_i$.

In communications between two users, both the message sender and the receiver have the secret key $\boldsymbol{B}$, and the sender sends $D1(B'_A)_n$ and $D2(B'_A)_n$ to the receiver. Then, the receiver composes B''_A, which is defined in Section 3.4 and expected to be similar to A.

3.4. *Audio data composition*

The scaling coefficient sequence for audio data B is expressed by

$$S(B)_k = \{y_1, y_2, y_3, \ldots, y_k\},$$

where k is the total number of scaling coefficient of B at the level . Then, a sequence

$$C(B)_k = \{Y_1, Y_2, Y_3, \ldots, Y_k\}$$

is determined, where $Y_i \in \{0,1,2,3,4\}$ is the element index denoting which of the five sets of scaling coefficients y_i of B belongs to. $S(B')_k$ is defined as:

$$S(B')_k = \{b_1, b_2, b_3, \ldots, b_k\},$$

where $b_i \in \left\{m_0^B, m_1^B, m_2^B, m_3^B, m_4^B\right\}$ is the average of the scaling coefficients of B at the range denoted by $Y_i \in \{0,1,2,3,4\}$ obtained from B.

A sequence $D3(B)_k$ is defined as:

$$D3(B)_k = \left\{z_{B,1}, z_{B,2}, \ldots, z_{B,k}\right\},$$

where k is the total number of scaling coefficient of B at the level, and $z_{B,q} = \left[\left|y_q\right|\right] \bmod 256$. B''_A is determined as follows; $S(B''_A)_k$ is calculated from $S(B')_k$ by replacing b_q with $m_{Z_p}^B$ when $z_{B,q} = z_p$, for $p = 1, \ldots, n$, then the audio data B''_A is composed by IDWT using the scaling coefficient sequence $S(B''_A)_k$ and a value of zero for all wavelet coefficients at every level. The receiver composes B''_A using $D1(B'_A)_n$ and $D2(B'_A)_n$, which are determined by both A and B, and are sent by the sender, in addition to B which the receiver has prior to the conversation. B''_A is expected to be similar to A.

3.5. *Communication of audio data with arbitrary length*

In general, the recording time for A is not the same as that for B. For B we use a unit recording time such as one second. Then, we apply the proposed method described in the above sections to A every unit recording time. When the recording time of A is indivisible by the unit recording time, the additional scaling coefficients that are needed for application of the proposed method are set to zero.

4. Numerical Experiment

We applied the proposed method using several voice recordings for A and for B we used (1) 'Classical', (2) 'Hiphop' music, with one second of recording time each. The music was taken from a copyright free database[5]. In all cases of the experiment, B''_A was audible and similar to A. However, B''_A contained low-frequency background noise. After erasing this noise, the audio recording became more audible and the tone of the sender's voice changed as if the speaker was a different person. Our next objective is to develop a method for decreasing the background noise while preserving the tone of the speaker.

5. Conclusion

We developed a secure communication method using a discrete wavelet transform for audio data. In this method, two users must each have a copy of the same piece of music, which has a length of one second and functions as a secret key, before communicating with each other. The music is transformed into a code before the conversation using the scaling coefficients obtained from a discrete wavelet transform. The audio data are transformed into another code using the same method as that for the music. Information on the difference between these two codes is sent from one user to the other. The user who receives the information can largely reconstruct the original recording using an inverse discrete wavelet transform with the code obtained from the music, the information on the difference between these two codes, and values of zero for all wavelet coefficients. The voice produced by the proposed method was audible.

References

1. Y. Yoshitomi, T. Asada, Y. Kinugawa, and M. Tabuse, An authentication method for digital audio using a discrete wavelet transform, *J. Inf. Sec.* **2**(2) (2011) 59-68.
2. D. Inoue and Y. Yoshitomi, Watermarking using wavelet transform and genetic algorithm for realizing high tolerance to image compression, *J. IIEEJ*, **38**(2) (2009) 136-144.
3. M. Shino, Y. Choi, and K. Aizawa, Wavelet domain digital watermarking based on threshold-variable decision, *Technical Report of IEIC*E, DSP2000-86, **100**(325) (2000) 29-34. (in Japanese)
4. S. Murata, Y. Yoshitomi, and H. Ishii, Audio watermarking using wavelet transform and genetic algorithm for realizing high tolerance to MP3 compression, *J. Inf. Sec.* **2**(3) (2011) 99-112.
5. M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, RWC music database: database of copyright-cleared musical pieces and instrument sounds for research purposes, *Trans. IPSJ*, **45**(3) (2004) 728-738.