A Method for Secure Communication Using a Discrete Wavelet Transform for Audio Data and Improvement of Speaker Authentication

Kouhei Nishimura[1], **Yasunari Yoshitomi**[2], **Taro Asada**[2], **and Masayoshi Tabuse**[2]
1: Nippon Telegraph and Telephone West Corp.
3-15 Bamba-cho, Chuo-ku, Osaka 540-8511, Japan
2:Graduate School of Life and Environmental Sciences, Kyoto Prefectural University,
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

Abstract

We developed a secure communication method using a discrete wavelet transform. Two users must each have a copy of the same piece of music to be able to communicate with each other. The message receiver can produce audio data similar to the sending user's speech by using our previously proposed method and the given recording of music. To improve the accuracy of speaker authentication, the quantization level for the scaling coefficients is increased. Furthermore, the amount of data sent to the message receiver can be remarkably reduced by exploiting the characteristics of this data.

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

1. Introduction

The elderly are often 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. In the present study, we propose a method for secure communication using a discrete wavelet transform (DWT) and thus improve speaker authentication; this is an enhancement of our previously proposed method.[1] It can be used with Internet protocol (IP) telephones, and it has the potential to help prevent telephone fraud.

2. Proposed Method

2.1. *Encoding*

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

In the course of our research,[1] we found that the histogram of the scaling coefficients for each domain of a multiresolution analysis (MRA) sequence is centered at approximately zero when a DWT is performed on audio data. Exploiting this phenomenon, we have developed a secure communication method using audio data.[1]

2.1.2. *Use of five quantization levels for scaling coefficients*

(1) *Parameter setting*

In our reported study,[1] we set the following coding parameters.

The values of $Th(\text{minus})$ and $Th(\text{plus})$ in Fig. 1 are chosen such that the nonpositive 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:

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.

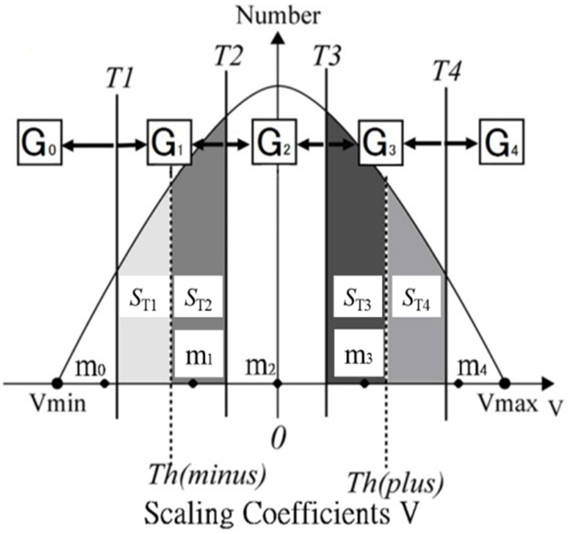

Fig. 1. Schematic diagram for demonstrating the selection of the scaling coefficients for encoding the audio data.[1]

(2) *Encoding*

In the preprocessing of the audio data prior to encoding, the scaling coefficients V of the MRA sequence are separated into five sets (G_0 to G_4), as shown in Fig.1, 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.

The scaling coefficients for the MRA sequence are encoded according to the following rules, where V_i denotes scaling coefficient *i*: $V_i \in G_0$, $c_i = 0$; when $V_i \in G_1$, $c_i = 1$; when $V_i \in G_2$, $c_i = 2$; when $V_i \in G_3$, $c_i = 3$; and when $V_i \in G_4$, $c_i = 4$. We represent the scaling coefficient for each set, G_j, by its average value, m_j. For the formation of audio data, we use a code C, which is the sequence of c_i and m_j defined above.

2.1.3. *Use of eight quantization levels for scaling coefficients*

Here, we define eight sets of $G_{8,0}$, to $G_{8,7}$, as follows:

- $G_{8,0} = \{V \mid V \in V^{SC}, V \le T1\}$,
- $G_{8,1} = \{V \mid V \in V^{SC}, T1 < V < Th(\text{minus})\}$,
- $G_{8,2} = \{V \mid V \in V^{SC}, Th(\text{minus}) \le V \le T2\}$,
- $G_{8,3} = \{V \mid V \in V^{SC}, T2 < V < 0\}$,
- $G_{8,4} = \{V \mid V \in V^{SC}, 0 \le V \le T3\}$,
- $G_{8,5} = \{V \mid V \in V^{SC}, T3 < V < Th(\text{plus})\}$,
- $G_{8,6} = \{V \mid V \in V^{SC}, Th(\text{plus}) \le V \le T4\}$,
- $G_{8,7} = \{V \mid V \in V^{SC}, T4 < V\}$.

Again, we let the representative value for each set, $G_{8,i}$, be its average, $m_{8,i}$. For the formation of audio data, we use the code C_8, which is the sequence of $c_{8,i}$ defined for eight quantization levels for scaling coefficients in the similar manner as c_i described in Section 2.1.2, and $m_{8,j}$ as defined above.

2.1.4. *Use of 16 quantization levels for scaling coefficients*

(1) *Parameter setting*

The values of *T*1*m*, *T*1*p*, *T*2*m*, *T*2*p*, *T*3*m*, *T*3*p*, *T*4*m*, and *T*4*p*, which are the parameters for controlling the authentication precision, are chosen to satisfy the following conditions:

1) $T1m < T1 < T1p < Th(\text{minus}) < T2m < T2 < T2p < 0 < T3m < T3 < T3p < Th(\text{plus}) < T4m < T4 < T4p$

2) The value of *T*1*m* is defined so that it equally divides the number of scaling coefficients in $[V\min, T1]$. *T*1*p*, *T*2*m*,…, *T*4*p* are defined similarly to *T*1*m*.

(2) *Encoding*

Sixteen sets of $G_{16,0}$ to $G_{16,15.}$ are defined as follows:

- $G_{16,0} = \{V \mid V \in V^{SC}, V \le T1m\}$,
- $G_{16,1} = \{V \mid V \in V^{SC}, T1m < V < T1\}$,
- $G_{16,2} = \{V \mid V \in V^{SC}, T1 \le V \le T1p\}$,
- $G_{16,3} = \{V \mid V \in V^{SC}, T1p < V < Th(\text{minus})\}$,
- $G_{16,4} = \{V \mid V \in V^{SC}, Th(\text{minus}) \le V \le T2m\}$,
- $G_{16,5} = \{V \mid V \in V^{SC}, T2m < V < T2\}$,

- $G_{16,6} = \{V \mid V \in V^{SC}, T2 \le V \le T2p\}$,
- $G_{16,7} = \{V \mid V \in V^{SC}, T2p < V < 0\}$,
- $G_{16,8} = \{V \mid V \in V^{SC}, 0 \le V \le T3m\}$,
- $G_{16,9} = \{V \mid V \in V^{SC}, T3m < V < T3\}$,
- $G_{16,10} = \{V \mid V \in V^{SC}, T3 \le V \le T3p\}$,
- $G_{16,11} = \{V \mid V \in V^{SC}, T3p < V < Th\,\text{(plus)}\}$,
- $G_{16,12} = \{V \mid V \in V^{SC}, Th\,\text{(plus)} \le V \le T4m\}$,
- $G_{16,13} = \{V \mid V \in V^{SC}, T4m < V < T4\}$,
- $G_{16,14} = \{V \mid V \in V^{SC}, T4 \le V \le T4p\}$,
- $G_{16,15} = \{V \mid V \in V^{SC}, T4p < V\}$.

As before, the value for each set, $G_{16,i}$, is represented by its average value, $m_{16,i}$. For the formation of audio data, we use the code C_{16}, which is the sequence of $c_{16,i}$ defined for 16 quantization levels for scaling coefficients in the similar manner as c_i described in Section 2.1.2, and $m_{16,j}$ defined above.

2.2. *Audio data formation using code replacement*

In this subsection, the formation of sound data is explained; for this example, we use five quantization levels for the scaling coefficient.[1] The scaling coefficient sequence for audio data A is expressed as $S(A)_k = \{x_1, x_2, x_3, \ldots, x_k\}$, where k is the total number of scaling coefficients of A at this level. Then, the 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, which indicates to which of the five sets of scaling coefficients x_i of A belongs. 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 $a_i \in \left\{m_0^A, m_1^A, m_2^A, m_3^A, m_4^A\right\}$ is the average of the scaling coefficients of A in the range denoted by $X_i \in \{0,1,2,3,4\}$ and is 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 the scaling coefficients of B in 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 in 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.

2.3 *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 $Y_i \neq X_i$, $z_p = [\,|y_i|\,] \bmod 256$, and the integer p is increased from 1 to n, in steps of size 1, when $Y_i \neq X_i$.[1] Here, $[x]$ signifies 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 for which $Y_i \neq X_i$ and $Z_p = X_i$.[1]

In communications between two users, the message sender and the receiver each have the secret key ***B***, and the sender sends $D1(B'_A)_n$ and $D2(B'_A)_n$ to the receiver.[1] Then, the receiver composes B''_A, which is defined in Section 2.4 and is expected to be similar to A.

2.4. *Audio data composition*

In this subsection, the processing of sound data formation is also explained using the case of five quantization levels, as an example, for the scaling coefficient.[1] The scaling coefficient sequence for audio data B is expressed as $S(B)_k = \{y_1, y_2, y_3, \ldots, y_k\}$, where k is the total number of scaling coefficients of B at this 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, which indicates to which of the five sets of scaling coefficients y_i of B belongs. $S(B')_k$ 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 $Y_i \in \{0,1,2,3,4\}$ and is obtained from B.

A sequence $D3(B)_k$ is defined as $D3(B)_k = \{z_{B,1}, z_{B,2}, \ldots, z_{B,k}\}$, where k is the total number of scaling coefficients of B at this level, and $z_{B,q} = [\,|y_q|\,] \bmod 256$. 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 using the inverse DWT (IDWT) of the scaling coefficient sequence $S(B''_A)_k$ and the value of zero for all wavelet coefficients at every level. The receiver composes B''_A from $D1(B'_A)_n$ and $D2(B'_A)_n$, which are determined by both A and B and are sent by the sender, and B, which the receiver has obtained prior to the conversation. B''_A is expected to be similar to A.

2.5. *Data reduction*

2.5.1. *Processing for* $D1$

Because $z_p = [\![y_i]\!] \bmod 256$, z_p is in the range from 0 to 255, and thus it can be expressed using 8 bits. In our computer, an integer is represented by 32 bits. Therefore, four values for z_p, each expressed using 8 bits, can be integrated into a single value expressed by 32 bits. For $D1(B'_A)_n = \{z_1, z_2, \ldots, z_n\}$, z'_j is defined as

$$z'_j = z_{4i-3} + z_{4i-2} \times 256 + z_{4i-1} \times 256^2 + z_{4i} \times 256^3,$$

where i, j are natural numbers. As a result, we obtain a sequence for $D1'(B'_A)_m = \{z'_1, z'_2, \ldots, z'_m\}$, where

$$m = \begin{cases} [n/4] & (n \bmod 4 = 0) \\ [n/4]+1 & (n \bmod 4 \neq 0). \end{cases}$$

When $n \bmod 4 \neq 0$, $z_{4m+k-2} = 0\,(k = 0, \ldots, |n \bmod 4 - 3|)$. Here, $[x]$ is defined as in Section 2.3. In the first case of the above formula on m, the total amount of data, $D1'$, stored in a computer is thus one quarter of that stored for $D1$. However, the total amount of data sent to a receiver depends on the way in which the data are expressed.

2.5.2. *Processing for* $D2$

(1) Case of five quantization levels

$D2(B'_A)_n = \{Z_1, Z_2, \ldots, Z_n\}$ and $D2'(B'_A)_l = \{Z'_1, Z'_2, \ldots, Z'_l\}$, where $Z'_j = Z_{13i-12} + Z_{13i-11} \times 5 + Z_{13i-10} \times 5^2 + \cdots + Z_{13i} \times 5^{12}$, are defined as described in Section 2.5.1.

(2) Case of eight quantization levels

$D2(B'_A)_n = \{Z_1, Z_2, \ldots, Z_n\}$ and $D2''(B'_A)_r = \{Z''_1, Z''_2, \ldots, Z''_r\}$, where $Z''_j = Z_{10i-9} + Z_{10i-8} \times 8 + Z_{10i-7} \times 8^2 + \cdots + Z_{10i} \times 8^9$, are defined as described in Section 2.5.1.

(3) Case of 16 quantization levels

$D2(B'_A)_n = \{Z_1, Z_2, \ldots, Z_n\}$ and $D2'''(B'_A)_s = \{Z'''_1, Z'''_2, \ldots, Z'''_s\}$, where $Z'''_j = Z_{8i-7} + Z_{8i-6} \times 16 + Z_{8i-5} \times 16^2 + \cdots + Z_{8i} \times 16^7$, are defined as described in Section 2.5.1.

3. Numerical Experiment

We applied the proposed method, using several voice recordings for A, and for B, we used two recordings of music, one classical and the other hip-hop. The music was taken from a copyright-free database.[2] In all cases, all of the produced B''_A were audible and sounded similar to A; each B''_A was made with five, eight, or 16 quantization levels. An increase in the quantization level improved the sound quality because a waveform made from B''_A with a higher quantization level was more similar to the original waveform than was one made with a lower quantization level, as shown in Fig. 2. For (1), (2), and (3) in Section 2.5.2, the data reduction for one minute of audio data at 44.1 kHz, 16 bits, a single channel, and volume of 87 KB was as follows:

$$(1)\ D1(75\,\text{KB}) \rightarrow D1'(48\,\text{KB}),\ D2(49\,\text{KB}) \rightarrow D2'(9\,\text{KB})$$

$$(2)\ D1(86\,\text{KB}) \rightarrow D1'(55\,\text{KB}),\ D2(57\,\text{KB}) \rightarrow D2''(21\,\text{KB})$$

$$(3)\ D1(92\,\text{KB}) \rightarrow D1'(59\,\text{KB}),\ D2(65\,\text{KB}) \rightarrow D2'''(29\,\text{KB})$$

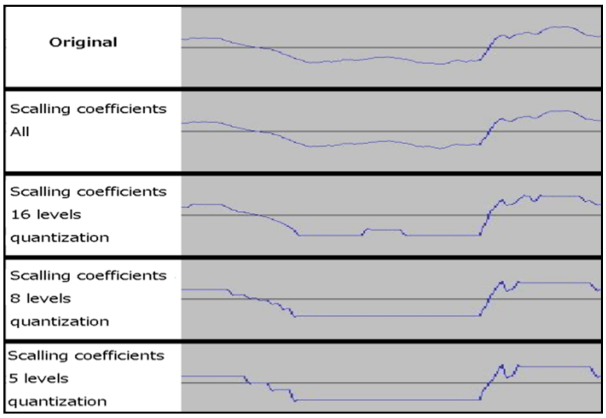

Fig. 2. Examples of waveform

4. Conclusion

We developed a secure communication method using a discrete wavelet transform for audio data; we used an increased number of quantization levels for the scaling coefficients along with a data reduction technique. The waveform produced by the proposed method was more similar to the original one than that produced by our previously proposed method.[1]

References

1. Y. Tsuda, K. Nishimura, H. Oyaizu, Y. Yoshitomi, T.Asada, and M.Tabuse, A method for secure communication using a discrete wavelet transform for audio data, *J. Robotics, Networking and Artif. Life*, **3**(3), 2016, pp. 193-196.
2. 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, pp 728-738.