IEICE TRANSACTIONS

# on Information and Systems

**VOL. E99-D NO. 1**
**JANUARY 2016**



A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER Special Section on Enriched Multimedia—Creation of a New Society through Value-added Multimedia Content—

# Reversible Audio Data Hiding Based on Variable Error-Expansion of Linear Prediction for Segmental Audio and G.711 Speech

**Akira NISHIMURA**[†a)], ***Senior Member***

**SUMMARY** Reversible data hiding is a technique in which hidden data are embedded in host data such that the consistency of the host is perfectly preserved and its data are restored during extraction of the hidden data. In this paper, a linear prediction technique for reversible data hiding of audio waveforms is improved. The proposed variable expansion method is able to control the payload size through varying the expansion factor. The proposed technique is combined with the prediction error expansion method. Reversible embedding, perfect payload detection, and perfect recovery of the host signal are achieved for a framed audio signal. A smaller expansion factor results in a smaller payload size and less degradation in the stego audio quality. Computer simulations reveal that embedding a random-bit payload of less than 0.4 bits per sample into CD-format music signals provide stego audio with acceptable objective quality. The method is also applied to G.711 $\mu$-law-coded speech signals. Computer simulations reveal that embedding a random-bit payload of less than 0.1 bits per sample into speech signals provide stego speech with good objective quality.

***key words:*** *steganography, audio coding, speech coding, performance evaluation, watermarking*

## 1. Introduction

Reversible data hiding is a technique for embedding hidden data in host data such that the consistency of the host data is perfectly preserved. The host data are then restored to their original form after the hidden data are retrieved using an extraction process. Several methods have been proposed for the reversible data hiding of audio data. These methods can be classified into three categories according to the domain of the embedded data: waveform domain [1], [2], spectral domain [3], [4], and compressed data domain [5].

In general, reversible data hiding achieves perfect recovery of the host data and perfect extraction of the payload from the unmodified stego data. If the recovered data are not identical to the host data, the technique is called semi-reversible. The current study only discusses perfect reversible techniques.

Reversible data hiding is considered to be useful for authentication, metadata recording, tamper detection [3], and covert communications, in which the host signal should not be modified for forensic use or should maintain high audio quality for commercial use. In addition, reversible data hiding provides a re-embedding feature through which a payload can be repeatedly embedded and removed. This technique is particularly useful for recording and rewriting metadata. In contrast to reversible data hiding, irreversible data hiding cannot recover the host signal, and re-embedding generally impairs the quality of the stego signal more, except for the LSB substitution. The requirements for this technology are minimal degradation of the stego signal quality, the ability to embed large payloads, and the undetectable concealment of the hidden data. In addition, a data hiding algorithm with a small computational load enables real-time embedding and detection. The techniques for the reversible data hiding of audio data in the waveform domain are typically simple and require less computational load compared with the techniques for hiding data in the other domains.

Veen *et al.* [1] proposed an amplitude expansion method that shifts bits of the amplitude data toward the most significant bit (MSB). However, the concealment of payload data is imperfect using this method because payload data are always represented by the least significant bit (LSB) data in the stego waveform. Yan and Wang [2] proposed a prediction error expansion method using linear prediction. In this method, the difference between the current and predicted amplitude is doubled, and the result is summed with the payload. This expanded difference is then added to the predicted amplitude to obtain the stego sample. These processes form the "prediction-error expansion". A location map, which indicates the non-expanded samples that prevent over/underflow of the stego amplitude, is embedded as overhead data. However, the concealment of payload data is inadequate because only five patterns of secret keys are used to represent the prediction coefficients [6].

The author has proposed an improved error expansion of the linear prediction technique for the reversible data hiding of audio waveforms [6]. In this technique, the errors when deriving the predicted amplitudes are reduced by using floating-point calculations and rounding the resulting output such that the degradation in the quality of the stego audio is minimized. Because a location map, which is employed to prevent amplitude overflow, is not embedded, the improved method also achieves a storage capacity of nearly 1 bit per sample as the payload. However, the conventional method [2] and the author's improved method [6] can only be applied to the entire host signal. Dividing the host signal into the framed signals of equal length and applying these methods to the framed signals to get the concatenated stego signal are possible, but payload extraction and recovery of the host signal from the unmodified part of a partially mod-

Manuscript received March 27, 2015.
Manuscript revised August 7, 2015.
Manuscript publicized October 21, 2015.

[†]The authors is with Department of Informatics, Faculty of Informatics, Tokyo University of Information Sciences, Chiba-shi, 265–8501 Japan.
a) E-mail: akira@rsch.tuis.ac.jp

DOI: 10.1587/transinf.2015MUP0009

ified stego signal is not possible, because locating the initial sample of the frame in the corresponding stego signal is required. In addition, controlling the sound quality of the stego signal by varying the payload size is challenging.

In order to resolve the aforementioned two drawbacks of the previous methods, the author has proposed an embedding method applied separately to the framed host signals [7]. The initial sample in the framed stego signal can be detected from an arbitrary portion of the entire stego signal. In other words, payload extraction and recovery of the host signal are available from the unmodified part longer than the frame length in the partially modified stego signal. These features are very important for tamper detection and for recovering the host signal from partially copied, pasted, cut, or lost stego signals, i.e., segmented stego signals. Furthermore, the objective sound quality of the stego signal is controlled by varying the payload size.

However, the previous paper [7] did not complete variable error expansion, because the definition of embedding using expansion and extraction using compression written in the paper was incorrect. The method was not generalized and can only be applied to the specific cases.

This paper redefines a reversible hiding method using variable expansion that expands an integer variable $\alpha$-fold. The redefined variable expansion method is applied to the error of the linear prediction. Computer simulations that use musical signals and G.711-coded speech signals as host signals were conducted to show relationships between the payload size and degradation in the sound quality of stego audio.

## 2. Variable Expansion Method

The previous expansion methods always expand the integer variable, such as amplitude [1], prediction errors [2], [6], or spectral coefficient [3], two-fold. A variable expansion method expands an integer variable $\alpha$-fold. This section describes the procedures for embedding and extracting data using the variable expansion method. The following section introduces this method into the linear prediction error method.

The variable expansion method expands the integer variable $\alpha$-fold ($1 < \alpha \leq 2$) and rounds the resulting value, and then it subtracts or adds the payload bit. $\alpha$ is a variable for controlling the stego quality and payload size. If $\alpha = 2$, the proposed method is identical to the previous two-fold expansion method, with the exception of inverting subtraction and the addition of a payload bit [6].

Equation (2) demonstrates how the payload data $q \in \{0, 1\}$ are embedded into the integer host data $h$ to obtain the stego data $s$. Equation (1) is an expansion and rounding function $G(.)$.

$$G(h) = \mathrm{round}(\alpha h). \tag{1}$$

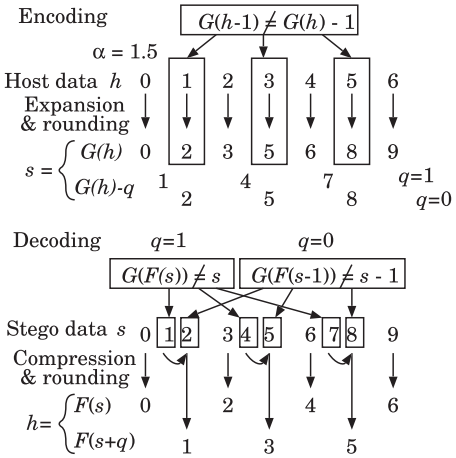


**Fig. 1** Examples of embedding, extraction, and recovery using the variable expansion technique.

$$s = \begin{cases} G(h) - q & \text{if } G(h-1) \neq G(h) - 1 \text{ and } h > 0, \\ G(h) + q & \text{if } G(h+1) \neq G(h) + 1 \text{ and } h < 0, \\ G(h) & \text{otherwise, i.e. not embeddable.} \end{cases} \tag{2}$$

Extraction of the payload, which is defined by (4), requires a compressive and rounding function $F(.)$, which is defined by (3).

$$F(s) = \mathrm{round}(s/\alpha), \tag{3}$$

$$q = \begin{cases} 0 & \text{if } G(F(s \pm 1)) \neq s \pm 1, \\ 1 & \text{if } G(F(s)) \neq s, \\ null & \text{otherwise.} \end{cases} \tag{4}$$

Introducing a sign function, which is defined by Eq. (5), the host data are recovered by the following:

$$\mathrm{sign}(s) = \begin{cases} 0 & s = 0, \\ 1 & s > 0, \\ -1 & s < 0. \end{cases} \tag{5}$$

$$h = \begin{cases} F(s) & \text{if } q = null, \\ F(s + \mathrm{sign}(s)q) & \text{otherwise.} \end{cases} \tag{6}$$

Examples of embedding, extraction, and recovery for $0 \leq h$ are schematically illustrated in Fig. 1.

## 3. Prediction-Error Expansion

There are two practical problems to be solved in the prediction error expansion method; expansion methods cannot avoid overflow and/or underflow in the amplitude domain. To solve this problem, a marking bit is introduced. This bit indicates whether the prediction-error of the current sample is expandable. The marking bit is embedded into the sample prior to the current sample. The location map in the current study is an array that points to embeddable samples. This map is derived by the prediction-error and the expansion factor in the embedding process. It is also derived by the expanded prediction-error, the expansion factor, and the

marking bits extracted during the extraction process. Therefore, the location map is not embedded into the stego signal. The location map proposed in the previous studies is an array that points to non-expandable samples. It resembles the current location map because it indicates embeddable and non-embeddable samples. However, the previous study [2] required the location map to be embedded into the stego signal.

Another problem is that partial recovery of the host waveform from an arbitrary part of the stego signal is challenging because the prediction errors depend on the previous host samples that are cut away. To overcome this problem, embedding into the framed host signals and frame synchronization between the framed host waveform and the stego waveform are required. The following sections describe over/underflow prevention and frame synchronization.

### 3.1 Embedding

We define an $n$th-order autoregressive linear prediction $p(t)$ $(t = 1, 2, \cdots, N-1)$ as

$$p(t) = \sum_{i=1}^{n} \mathrm{round}(a(i)x(t-i)), \tag{7}$$

for a discrete time series of integer host data $x(t)$, where $t = 0, 1, \cdots, N-1$ and $N$ is the number of the host data. Equation (7) differs from general linear prediction because a rounding function is introduced to obtain an integer value for $p(t)$. Floating-point array data $a(i)$, which can be derived by applying the Burg method [8] to all $x(t)$, are the prediction coefficients. Each set of $n$th-order $a(i)$ is considered to be a host-specific secret key that must be referenced during the extraction process. The prediction error $d(t)$ $(t = 1, 2, \cdots, N-1)$ is defined as

$$d(t) = x(t) - p(t). \tag{8}$$

The variable expansion method combined with linear prediction is established by considering $d(t)$ to be an integer variable $h$ in (2) and by embedding the time series of a bit to be embedded $b(t) \in \{0, 1\}$ as a payload bit $q$ in (2). Therefore, (2) is rewritten to derive the expanded and embedded prediction error $d'(t)$.

$$d'(t) = \begin{cases} G(d(t)) - b(t) & \text{if } G(d(t)-1) \neq G(d(t)) - 1 \\ & \text{and } d(t) > 0, \\ G(d(t)) + b(t) & \text{if } G(d(t)+1) \neq G(d(t)) + 1 \\ & \text{and } d(t) < 0, \\ G(d(t)) & \text{otherwise, i.e. not embeddable.} \end{cases} \tag{9}$$

The stego signal $y(t)$ is obtained by

$$y(t) = \begin{cases} p(t) + d'(t) & (0 < t \leq N-1), \\ x(t) & (t = 0). \end{cases} \tag{10}$$

### 3.2 Extraction and Recovery

In the extraction process, $p(t)$ is recovered from (7) using $x(t-i)$ and $a(i)$, where $t > 0$ and $1 \leq i \leq n$, because $x(0) = y(0)$ from Eq. (10). Then, the expanded prediction error $d'(t)$ is derived by subtracting $p(t)$ from $y(t)$:

$$d'(t) = y(t) - p(t). \tag{11}$$

Equation (4) is also rewritten as

$$b(t) = \begin{cases} 0 & \text{if } G(F(d'(t) \pm 1)) \neq d'(t) \pm 1, \\ 1 & \text{if } G(F(d'(t))) \neq d'(t), \\ null & \text{otherwise.} \end{cases} \tag{12}$$

Then, the host signal is recovered as follows:

$$x(t) = \begin{cases} p(t) + F(d'(t)) & \text{if } b(t) = null, \\ p(t) + F(d'(t) + \mathrm{sign}(d'(t))b(t)) & \text{otherwise.} \end{cases} \tag{13}$$

Calculating (7), (11), (12), and (13), where $t \geq 1$, successively extracts the payload bits and recovers the host signal.

### 3.3 Preventing Amplitude Overflow and/or Underflow

The expansion of $d(t)$ can cause amplitude overflow and/or underflow. In the beginning of the embedding process, a location map $m(t), (t = 1, 2, \cdots, N-1)$, which represents embeddable samples, is prepared. Equation (9) is evaluated to locate samples that can be embedded. If the sample at $t$ is decided as not embeddable according to Eq. (9), $m(t)$ is set to 0; otherwise, it is set to 1. If condition (14) is satisfied, the error expansion causes overflow or underflow at $y(t)$. In this case, the expansion and embedding are canceled—that is, $y(t) = x(t), b(t) = null$, and $m(t) = 0$ at the sample. INT_MAX and INT_MIN are the maximum and the minimum number of the signed integer, respectively. If $x(t)$ is linearly quantized in 16-bit, INT_MAX is 32767 and INT_MIN is –32768.

$$\begin{aligned} &p(t) + G(d(t)) > \mathrm{INT\_MAX} \quad \text{or} \\ &\quad p(t) + G(d(t)) < \mathrm{INT\_MIN}. \end{aligned} \tag{14}$$

If condition (14) is satisfied, error expansion is canceled in the embedding process; however, the cancellation cannot be correctly detected from the given $y(t)$ and $p(t)$ in the extraction process. Figure 2 shows examples of (a) not expanded and (b) expanded conditions that yield identical values of $y(t)$ in the embedding process. These two conditions are not discriminable in the extraction process. Therefore, the condition that does not satisfy Fig. 2(a) but satisfies Fig. 2(b) is defined by (15) as the expandable condition. To discriminate these conditions (14) and (15) in the extraction process, a marking bit that indicates whether the prediction error of the current sample was expandable is embedded into the sample prior to the current sample.

$$\mathrm{INT\_MAX} - G(G(d(t))) < p(t) \leq \mathrm{INT\_MAX} - G(d(t)) \quad \text{or}$$

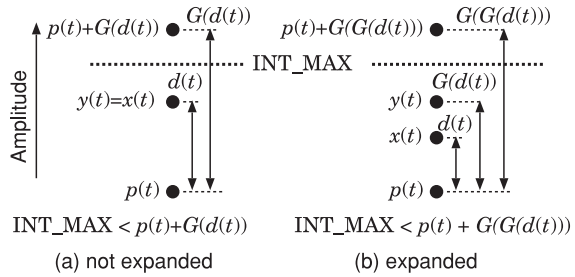


**Fig. 2** Examples of (a) not expanded and (b) expanded conditions that yield identical values of $y(t)$ in the embedding process. These two conditions are not discriminable in the extraction process when $y(t)$ and $p(t)$ are given.

$$\text{INT\_MIN} - G(d(t)) \le p(t) < \text{INT\_MIN} - G(G(d(t))). \quad (15)$$

In the embedding process, let $k$ be an index of the marking offset, with an initial value of 1. If condition (14) or (15) is satisfied, $k$ is incremented from 1 until $m(t-k) = 1$ to locate an available sample for embedding a marking bit. At this point, $m(t-k)$ is set equal to 0, which means that a marking bit is embedded into $d(t-k)$ rather than a payload bit. If condition (14) is satisfied, the marking bit $b(t-k)$ is set equal to 0 and is embedded into $d(t-k)$, which means that $d(t)$ was not expanded. If condition (15) is satisfied, the marking bit $b(t-k)$ is set equal to 1 and is embedded into $d(t-k)$, which means that $d(t)$ was expanded.

If $t-k < 1$ is satisfied, there is no room to embed the marking bit. In this case, extended marking bits $e(k-t)$, where $0 \le k-t$, which play the same role as the marking bits, are introduced. The $r$-bit $e(j)$, $(j = 0, 1, \cdots, r-1)$ are replaced by LSBs of the stego signal, as shown in Sect. 3.4. Figure 3 shows a flowchart of the embedding process, excluding extended marking bits in order to avoid complexity of the chart.

An alternative solution to prevent amplitude overflow and/or underflow is to cancel the error-expansion process when both conditions (14) and (15) hold. But it may cause sound quality degradation, because frequent and abrupt changes are induced in the residual noise components as a result of the prediction-error expansion, which is discussed in the last paragraph in Sect. 6.1.

In the extraction process, $m(t)$ and $k$ are initialized for the same purpose as the embedding process. If condition (16) is satisfied, $k$ is incremented from 1 until $m(t-k) = 1$ to locate the sample where the marking bit was embedded. At this point, if $b(t-k)$ is 1, then $d'(t)$ was expanded. Therefore, $b(t)$ can be extracted by (12) and $x(t)$ is recovered by (13). Otherwise, $d'(t)$ was not expanded and not embedded, that is, $x(t) = y(t)$, $b(t) = null$ and $m(t) = 0$. Figure 4 shows a flowchart of the extraction and recovery process, excluding the use of extended marking bits.

$$p(t) + G(d'(t)) > \text{INT\_MAX} \quad \text{or}$$
$$p(t) + G(d'(t)) < \text{INT\_MIN}. \quad (16)$$

Note that $m(t-k)$ and $b(t-k)$ can be generated prior to time series data $x(t)$ or $y(t)$, indicating that the embedding and extraction processes are causal.

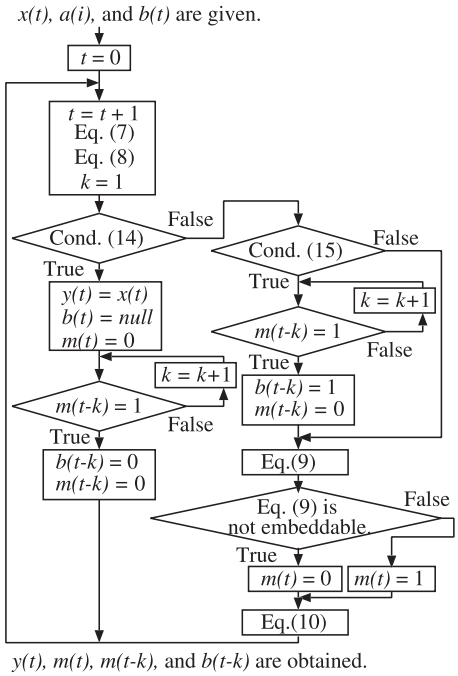


**Fig. 3** A flowchart of the embedding process, excluding extended marking bits.

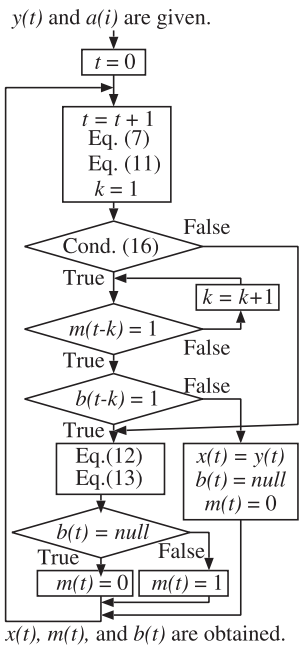


**Fig. 4** A flowchart of the extraction and recovery process, excluding the use of extended marking bits.

### 3.4 Concealment of Overhead Data and Frame Synchronization Technique

The variable prediction error expansion method requires the following overhead data for decoding: an $r$-bit extended marking bit, where $r$ is an 8-bit unsigned integer number; a

set of prediction coefficients; an expansion factor; and signature data. The prediction coefficients $a(i)$ and the expansion factor $\alpha$ are expressed as 16-bit IEEE 754 half-precision floating-point numbers. Because the length of the extended marking bits $r$ is variable, $r$ is expressed as an 8-bit unsigned integer number.

Signature data are required to detect the frame signal from the arbitrary extracted stego signal. These data are generated by applying an exclusive or (XOR) operation of the $L$-bit secret key on the masking $L$ bits. These masking bits are selected from the LSBs whose positions are determined from the initial stego samples using the secret key in which the payload is embedded. The XOR operation is applied to the remainder of the overhead data in the same manner. $r + 8 + 16 + 16n + L$ bits of overhead data are replaced with the LSBs contained in the final part of the stego signal.

The author's previous method [6] scrambles the $a(i)$ data and replaces them with the LSBs of the stego signal. However, if this method is applied to framed signals, identical bit values are repeatedly found in the LSB data, implying that data are hidden. Utilizing the XOR operation is expected to prevent the appearance of constant bit values in the LSBs of the stego signal.

The entire process flow of the embedding is shown in Fig. 5:

1. Examine $x(t)$ and generate $m(t)$ according to (9) and conditions (14) and (15).
2. Embed the marking bits and payload bits using (10) into the initial and final parts of the host signal.
3. Store the LSB data of the final part of the stego signal.
4. Embed the stored data into the host samples immediately before the stored part.
5. Select the masking bits from the LSBs in the initial part of the stego signal.
6. Evaluate the XOR operations on the selected masking $L + 8$ bits with the secret key to generate the signature and the XOR operations on the selected masking $(r + 16 + 16n)$ bits with the corresponding overhead data.
7. Replace the XOR-operated overhead data with the LSBs in the final part of the stego signal.

In the detection process, the detector evaluates the XOR operations on the LSBs, which constitute the masking bits, with selected bits of the signature because the relative positions of the masking and signature bits are known through the secret key. The resulting bits are identical to the secret key if the selected positions are synchronized with those in the embedding process. Step-by-step searching through shifting of the selected positions achieves frame synchronization with only a small computational load. $r$ can be decoded using the same procedure as for extracting the signature bits. Then, the remainder of the overhead $r + 16 + 16n$ bits are decoded.

Failure to locate the signature bits in a stego segment longer than the frame length implies the existence of tampered masking bits and/or signature bits. Embedding the hash data of the framed host signal achieves detection of the tampered stego frames by comparing the extracted hash data with the hash data of the recovered host signal [3].

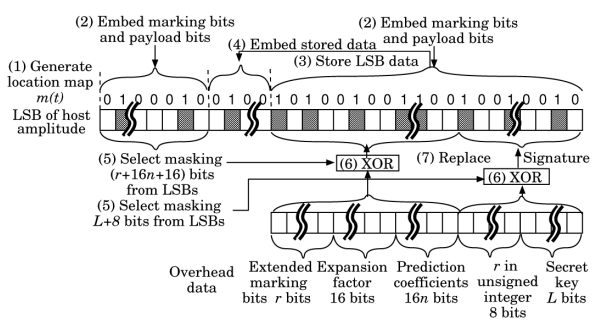


**Fig. 5** Schematic of embedding process flow.

## 4. Embedding Payload Into G.711 Speech Signal

G.711 is an ITU-T standard for audio compounding [9]. This standard is primarily used not only in telephony but also in Voice-over-IP (VoIP). The $\mu$-law and A-law algorithms encode 14-bit and 13-bit signed linear pulse code modulation (PCM) samples, respectively, to logarithmic 8-bit samples. Thus, the G.711 encoder create a 64 kbits/s bit-stream for a signal sampled at 8 kHz. $\mu$-law is used in North America and Japan, and A-law is used in Europe and the rest of the world.

ITU-T recently published Recommendation G.711.0, which describes a variable bit rate and lossless compression scheme of a G.711 bitstream aimed primarily for transmission over IP [10]. The target applications of G.711.0 are high-quality voice communication services, such as distant voice meetings, IP telephony, audio and visual communication, and streaming multimedia. Applying reversible data hiding to the G.711 codec is useful for the high-quality voice communication services that the G.711 codec can realize, providing compatibility to the conventional G.711 bitstream with a hidden communication channel rather than providing the variable bit rate feature of G.711.0.

In this section, error expansion of linear prediction for a G.711 ($\mu$-law) speech signal is introduced. Let $M(z)$ be an encoding function of G.711 for an integer input whose amplitude range is $-8159 \le z \le 8159$. Let $M^{-1}(u)$ be a decoding function of G.711 for an integer input whose range is $-127 \le u \le 127$. The host signal $X(t)$ is the G.711-coded data. Equations (7), (8), and (10) are rewritten as follows:

$$p(t) = \sum_{i=1}^{n} \mathrm{round}(a(i)M^{-1}(X(t-i))) \tag{17}$$

$$d(t) = X(t) - M(p(t)), \tag{18}$$

$$Y(t) = \begin{cases} M(p(t)) + d'(t), & (t > 0) \\ X(t), & (t = 0) \end{cases} \tag{19}$$

where $Y(t)$ is G.711-encoded stego data. INT_MAX and

INT_MIN in (14), (15), and (16) are 127 and –127 respectively. The technique of overflow and/or underflow prevention and frame synchronization are the same as described in Sect. 3.3 and 3.4.

If the host signal is packets of VoIP, frame synchronization is achieved by the Internet protocol and the prediction coefficient set $a(i)$ is commonly prepared and applied to all speech signals. If the host signal is a G.711-coded file, a prediction coefficient set is calculated by using entire speech in a file and embedded once for each speech file.

The payload is extracted by (12), and host data are recovered by (20).

$$X(t)=\begin{cases} M(p(t))+F(d'(t)) & \text{if } b(t)=null, \\ M(p(t))+F(d'(t)+\text{sign}(d'(t))b(t)) & \text{otherwise.} \end{cases} \quad (20)$$

## 5. Evaluation

The proposed expansion method was applied to framed host signals. The performance of the method was measured under the simulated host conditions through the amount of payload and the objective sound quality of the stego signal.

### 5.1 Embedding Payload into Music Signal

A total of 100 pieces of music from a database containing various types of music (RWC-MDB-G-2001) [11] served as host data. Explicitly, the 20 s between the initial 40 and 60 s, 44.1 kHz sampling, 16-bit resolution and stereo-channel signal of each piece was used. To model the hidden data, random-bit data were embedded into the host signal. Embedding was conducted independently to left and right channel.

Objective quality degradation of the stego audio was evaluated in terms of signal-to-noise ratio (SNR) and perceptual audio quality evaluation (PEAQ). An objective difference grade (ODG), which corresponds to the degree of subjective quality degradation of the stego audio signal when compared with the original audio signal, was obtained using PEAQ software [12] included in AFsp_v9r0 package. In terms of subjective quality, an ODG value of 0 corresponds to no difference, –1 corresponds to a perceptible but not annoying difference, and –2 corresponds to a slightly annoying difference.

The evaluation of the method was conducted by varying the length of the frame $N$ from 11,025 (0.25 s) to 88,200 (2.0 s) and by varying the expansion factor $\alpha$ from 1.2 to 2.0. The maximum order of the prediction coefficient $a(i)$ for the proposed method was $n = 8$. The length of the secret key was $L = 128$. Therefore, the constant overhead per frame was 280 bits. To reduce the amplitude overflow and underflow caused by the error expansion in the initial samples, the prediction coefficient and order were fixed to $a(1) = 1$ and $n = 1$ for the initial seven host samples. The prediction coefficient set was obtained using the entire host signal (20 s) that served for all frames in common. This condition is called 'common coefficient set' hereafter. Alternatively, the prediction coefficient set was obtained using the framed host signal (0.25 to 2.0 s) that served for all frames independently. This condition is called 'independent coefficient set' hereafter. The independent coefficient set condition adaptively changes the coefficient sets to the local audio frames in an audio file with the aim to reduce prediction-error. It may improve the quality of stego audio compared with the common coefficient set condition. However, changes in the prediction coefficient sets between successive frames may change the spectral characteristics of the residual noise components at the border of the successive frames, which might result in a degradation of the stego quality. Testing the two coefficient set conditions will clarify the better condition in terms of payload size and stego quality.

### 5.2 Results

Table 1 presents the results of the mean payload rate per channel (in units of kbits/s). Because the mean payload rates obtained under the common coefficient set condition were almost identical to those obtained under the independent coefficient set, Table 1 only lists the results obtained using the common coefficient set. The results show that the mean payload rate approaches the theoretical payload rate without overhead data (i.e., $(\alpha - 1)$ bits per sample) as the frame length increases because the ratio between the amounts of the overhead data and payload decreases.

Table 2 shows the SNRs in dB under the common coefficient set. The SNRs obtained under the independent coefficient set were slightly higher (by approximately 1 dB) than those obtained under the common coefficient set. The frame length did not affect the SNRs. As expected, the SNR was increased by decreasing the expansion factor.

Figures 6 and 7 show the median, $10^{th}$ and $90^{th}$ percentiles, and ranges of ODG values when the frame lengths are 11,025 and 88,200, respectively. At an expansion factor of 1.2, the quality degradation of all stego signals is extremely small. The objective quality under the independent coefficient set is slightly worse compared with that under the common coefficient set because changes in the prediction coefficient set cause changes in the spectral character-

**Table 1** Mean payload rates for the common coefficient set. The mean payload rates obtained under the common coefficient set condition were almost identical to those obtained under the independent coefficient set.

| Payload bit-rate per channel [kbits/s] | Expansion factor $\alpha$ | | | | |
|---|---|---|---|---|---|
| Frame length | 1.2 | 1.4 | 1.6 | 1.8 | 2.0 |
| 11,025 | 7.6 | 16.4 | 25.2 | 33.9 | 42.7 |
| 22,025 | 8.2 | 17.0 | 25.7 | 34.5 | 43.3 |
| 44,100 | 8.5 | 17.3 | 26.0 | 34.8 | 43.6 |
| 88,200 | 8.6 | 17.4 | 26.2 | 34.9 | 43.7 |

**Table 2** SNRs for the common coefficient set.

| All frame lengths | Expansion factor $\alpha$ | | | | |
|---|---|---|---|---|---|
| | 1.2 | 1.4 | 1.6 | 1.8 | 2.0 |
| SNR [dB] | 35.1 | 29.0 | 25.5 | 23.0 | 21.1 |

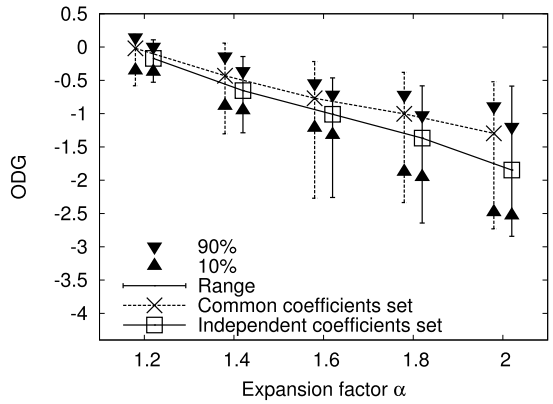


**Fig. 6** Median ODG values for a frame length of 11,025. Data points are slightly shifted along the horizontal axis for clarity.

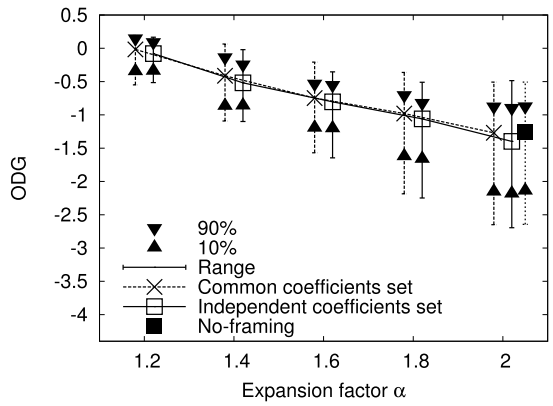


**Fig. 7** Median ODG values for a frame length of 88,200. Data points are slightly shifted along the horizontal axis for clarity.

istics of the residual noise components at the border of the successive frames. The frame length has a weak effect on the ODG values under the common coefficient set. Figure 7 also presents the results of no-framing, that is, embedding into the entire 20-s host signal at once. The 2-s framing of the host signal has a negligible effect on the ODG values compared to the no-framing condition. In summary, the best ODGs were obtained at the 2-s frame length and the common coefficient set. Almost all stego music at the expansion factor $\alpha = 1.4$ exhibited ODG values of greater than –1, which corresponds to subjective quality of perceptible but not annoying.

### 5.3 Embedding into G.711 Speech Signal

Speech files from the speech database published as ITU-T P.50 Appendix I [13], which is primarily used for the objective evaluation of speech processing systems and devices, served for computer simulations of the payload and of the objective quality of the stego speech signals. This database includes 112 speech files spoken by seven languages, American English, Arabic, Chinese, Danish, French, German, and Japanese, each of which consists of eight female speech files and eight male speech files. The duration of the speech ranged from 3.2 to 17.3 seconds, including silence intervals. The format of the files is 16-kHz sampling, 16-bit quantization, and single channel. The overall level of each speech signal was normalized to –26 dBov. The files were converted to 8-kHz sampling and 8-bit G.711 ($\mu$-law) files.

The embedding parameters were as follows: $\alpha$=1.1, 1.2, and 1.3; the order of prediction coefficients $n = 4$; and the length of frame $N$=160 and 320(20, 40 ms), which is identical to the assumed frame size of VoIP. If the host speech signal is assumed to be in a G.711 file, a prediction coefficient set can be derived from an entire speech signal in the individual file. This condition is called the 'individual coefficient condition'. If the host signal is assumed to be a real-time VoIP data stream, a coefficient set cannot be derived prior to embedding. Thus, a fixed coefficient set should be applied to all speech signals for embedding. This condition is called the 'fixed coefficient condition'. These two conditions are tested to simulate actual situations. In the fixed coefficient condition, the median of the coefficients calculated from half of the files was applied to the other half of the files.

The signature bits were not embedded into the stego data because the frame synchronization was considered to be already established in the above two conditions. The prediction coefficients $a(i)$ were not embedded into the stego data because it is practical to use them as a secret key for embedding and extraction. To model the payload data, random-bit data were embedded into the host signal.

Perceptual evaluation for speech quality (PESQ) is a perceptual-based method of objective sound quality evaluation for speech codecs. PESQ compares an original signal with a signal that has been degraded by passing through a communications system using the psychophysical representation of audio signals. The transformed output of PESQ (ITU-T P.862.1) is called the Mean Opinion Score Listening Quality Objective (MOS-LQO) and corresponds to the results of the Mean Opinion Score Listening Quality Subjective (MOS-LQS) obtained from human listeners through subjective experiments. The PESQ software distributed by ITU-T was used to evaluate the objective quality degradation of the decoded speech signals. A MOS-LQO of 4 corresponds to the subjective evaluation of 'Good', 3 corresponds to 'Fair', 2 corresponds to 'Poor', and 1 corresponds to 'Bad'. On average, the MOS-LQO of the coded speech signals by the typical low-bitrate (4.75 kbps) adaptive multi-rate (AMR) speech coder is 3.5. [14]

### 5.4 Results

Figure 8 plots the MOS-LQO and bit-rate of each speech file as a dot for $N$=320. The types of the dots represent experimental parameters, $\alpha$ and 'Fixed' or 'Individual' coefficient conditions. Table 3 presents the mean MOS-LQO and mean bit-rate of the payload for each condition for $N$=320. The results obtained from the conditions $N$=160 exhibited quite small differences in both mean MOS-LQOs (less than 0.1) and mean bit-rates of payload (less than 7 bits/s) compared

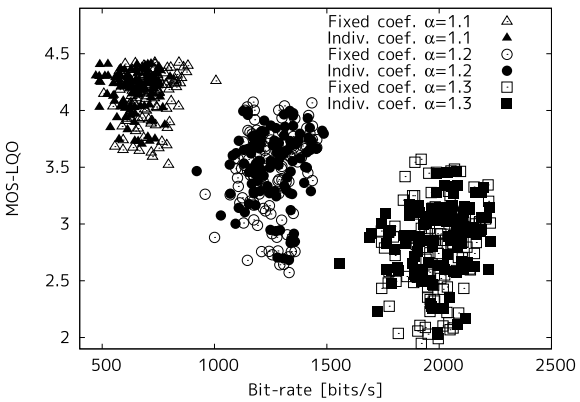


**Fig. 8** MOS-LQO and bit-rate of payload for the two prediction coefficient conditions and expansion factor $\alpha$, $N$=320.

**Table 3** Mean bit-rate of payload and MOS-LQO for G.711 reversible hiding, $N$=320.

| Expansion | Fixed coefficient set | | Individual coefficient set. | |
|---|---|---|---|---|
| factor $\alpha$ | MOS-LQO | bit-rate [bits/s] | MOS-LQO | bit-rate [bits/s] |
| 1.1 | 4.13 | 711.2 | 4.18 | 645.8 |
| 1.2 | 3.44 | 1253.3 | 3.52 | 1262.0 |
| 1.3 | 2.78 | 1986.9 | 2.84 | 1995.2 |

with those obtained from the $N$=320 conditions.

On average, the individual coefficient conditions exhibited slightly smaller payload sizes than the fixed coefficient conditions for an expansion factor of $\alpha = 1.1$. This result is caused by the small prediction error in the linear prediction using the individual coefficient set. If $\alpha = 1.1$ and the prediction error $d(t)$ ranges –4 to 4, no payload can be embedded into $d(t)$. If $\alpha$ becomes larger, the number of samples that overflows or underflows gradually increases, which results in a slightly smaller payload size in the fixed coefficient conditions than that in the individual coefficient conditions, as shown in Table 3. The MOS-LQO was slightly better for the individual coefficient conditions than the fixed conditions because the prediction error $d(t)$ is generally smaller in the individual coefficient conditions.

In summary, the difference between the two conditions is relatively small in terms of the amount of payload and MOS-LQO. The advantage of the fixed coefficient condition is that a single fixed prediction coefficient set can be applied to any kind of speech signals in both VoIP and file. Therefore, it is practical to use the fixed prediction coefficient set for G.711.

## 6. Discussion

### 6.1 Reversible Hiding for Music Signal

Steganography has been considered useful for the recording of meta-data [15], covert communications [16], and quality enhancements, such as bandwidth extension [17]. The most useful advantage of reversible data hiding compared with conventional steganography technology is that the host signal can be recovered from the stego signal. Reversible data hiding for music signals can be applied not only to these applications but also to applications that require high-quality audio, including commercial music and music recording.

The audio content used in recording is saved as waveform data or as losslessly compressed waveform data because audio quality is indispensable. Such audio data require metadata, e.g., time stamp, recording targets, copyright information, editing processes and their parameters, and their histories. These metadata are generally recorded in the header area of the waveform file or other files that accompany the waveform file. Reversibly embedding these metadata into the originally recorded waveform file results in non-destructive editing, which can be used to trace the editing history, copyright management, and transmission, regardless of the waveform file and lossless compression formats. Audio editing software that supports the reversible data hiding format can be used to read, edit, and rewrite the metadata as payload only for authenticated users. Therefore, reversible audio data hiding technology will be useful in high-quality, non-destructive, secured audio editing and recording environments [4].

The re-embedding feature that embeds and removes the payload repeatedly using reversible data hiding is useful in conjunction with the feature that the payload is inseparable with the host signal. Recently, user-generated music and its derivative works through the Internet is widely spreading. Such audio contents do not require traditional copyright protection but are expected to provide declaration of copyright for the honor of the creators and editors, though the declaration is not mandatory. Reversible audio data hiding supplies audio contents that are inseparable from the copyright declaration as the payload to distribute on the Internet. Authenticated users can edit and listen to the high-quality audio that removed payload from the distributed audio contents. They can also redistribute the modified audio contents, which are embedded metadata, including editing histories of the audio contents and additive copyright declaration. This framework requires additional tools and software to handle reversible data hiding. Although it is invalidated by applying perceptual codecs to audio contents for distribution, the framework may enhance the experience and motivation for creating music.

The conventional method [2] controls sound quality degradation through a threshold embedding value $T$. Specifically, embedding is canceled if the prediction error exceeds $T$. Although such a limit on embedding is effective for improving SNRs, it is not effective for improving ODG because frequent and abrupt changes are induced in the residual noise components as a result of the prediction-error expansion. In contrast, the proposed expansion method constantly expands the prediction error for almost all samples, and therefore, the resulting residual components are smooth and difficult to distinguish.

### 6.2 Reversible Data Hiding for Speech Signals

Reversible data hiding for speech signals is useful for pro-

bative recording during investigations, where editing and modifying are strictly prohibited. Additionally, bandwidth extension without quality degradation in the low-frequency region is promising. Kataoka et al. developed a steganographic bandwidth extension for G.711 using side information of 600 bits/s [18]. The proposed reversible speech hiding achieves the payload size of 600 bits/s on average without severe quality degradation. However, a technique for embedding at a constant rate of payload has not been developed. It is a future problem to be solved.

Aoki proposed reversible data hiding for G.711 $\mu$-law speech, which embeds a payload bit into a sample whose absolute amplitude is zero [19]. G.711 $\mu$-law coding expresses zero as +0 and -0, depending on the bit that represents its sign. Therefore, the amount of payload is proportional to the number of zero-amplitude data of the host. The proposed method can be used in conjunction with Aoki's method to increase the amount of payload while preserving the reversible feature.

## 7. Conclusions

This paper proposed a variable expansion method for reversible data hiding. This method is able to control the payload size by varying the expansion factor. Furthermore, this method is combined with an error expansion of linear prediction [6]. Reversible embedding, payload detection, and recovery of the host signal are achieved for framed audio signals. A smaller expansion factor results in a smaller payload size and less degradation in the quality of stego audio. Computer simulations revealed that embedding a random-bit payload of less than 0.4 bits per sample into CD-format music signals realized an allowable objective quality of stego audio. The method was also applied to G.711 $\mu$-law-coded speech signals. Computer simulations revealed that embedding a random-bit payload of less than 0.1 bits per sample into speech signals realized good objective quality of stego speech.

## Acknowledgments

Part of this work was performed under the Cooperative Research Project Program of the RIEC, Tohoku University. This work was also supported by a Grant-in-Aid for Scientific Research C (KAKENHI 24500128), 2014.

### References

[1] M. van der Veen, A. van Leest, and F. Bruekers, "Reversible audio watermarking," Proc. 114th AES Convention, no.5818, p.10, 2003.
[2] D. Yan and R. Wang, "Reversible data hiding for audio based on prediction error expansion," Proc. of IIHMSP2008, pp.249–252, 2008.
[3] X. Huang, A. Nishimura, and I. Echizen, "A reversible acoustic steganography for integrity verification," Digital Watermarking LNCS 6526, pp.305–316, 2011.
[4] A. Nishimura, "Reversible audio data hiding in spectral and time domains," in Multimedia Information Hiding Technologies and Methodologies for Controlling Data, ed. K. Kondo, pp.19–41, IGI Global, 2012.
[5] L. Liu, M. Li, Q. Li, and Y. Liang, "Perceptually transparent information hiding in G.729 bitstream," Proc. of IIHMSP2008, pp.406–409, 2008.
[6] A. Nishimura, "Reversible audio data hiding using linear prediction and error expansion," Proc. of IIHMSP2011, pp.318–321, 2011.
[7] A. Nishimura, "Controlling quality and payload in reversible data hiding based on modified error expansion for segmental audio waveforms," Proc. of IIHMSP2012, pp.110–113, 2012.
[8] J.P. Burg, "Maximum entropy spectral analysis," 1975. http://sepwww.stanford.edu/data/media/public/oldreports/sep06/06_01.pdf.
[9] ITU-T, "ITU-T recommendation G.711: Pulse code modulation (PCM) of voice frequencies," 1972.
[10] ITU-T, "ITU-T recommendation G.711.0: Lossless compression of G.711 pulse code modulation," 2009.
[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," Proc. 4th International Conference on Music Information Retrieval (ISMIR 2003), pp.229–230, 2003.
[12] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," Telecommunication Signal Process. Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University, pp.1–92, 2003.
[13] ITU-T, "ITU-T recommendation P.50 Appendix I: Artificial voices; test signals," 1998.
[14] 3rd Generation Partnership Project, "Performance characterization of the adaptive multi-rate (AMR) speech codec (release 6)," vol.26.975, 2004.
[15] A. Kunisa, "Host-cooperative metadata embedding framework," Proc. IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp.33–36, IEEE, 2007.
[16] M. Mason, S. Sridharan, and R. Prandolini, "Digital coding of covert audio for monitoring and storage," Proc. Fifth International Symposium on Signal Processing and its Applications, pp.475–478, IEEE, 1999.
[17] N. Aoki, "A band extension technique for G.711 speech using steganography," IEICE Trans. Commun., vol.E89-B, pp.1896–1898, 2006.
[18] A. Kataoka, T. Mori, and S. Hayashi, "Bandwidth extension of G.711 using side information," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J91-D, no.4, pp.1069–1081, April 2008.
[19] N. Aoki, "A technique of lossless steganography for G.711," IEICE Trans. Comm., vol.E90-B, pp.3271–3273, 2007.

**Akira Nishimura** received B. Eng. and M. Eng. degrees in acoustics from Kyushu Institute of Design in 1990, 1992 respectively. He received Ph. D. degree in audio information hiding from Kyushu University in 2011. Since 1996 he is a faculty member of Tokyo University of Information Sciences. He is a professor in the Department of Informatics. His current research interests are auditory modeling, audio information hiding, musical acoustics, and psychology of music. He is a member of Acoustical Society of Japan, Audio Engineering Society, IEEE, and Japanese Society of Music and Cognition. He got the Sato Prize from Acoustical Society of Japan in 2012.