Regular Paper

Reversible Audio Information Hiding for Tampering Detection and Localization Using Sample Scanning Method

Xuping Huang[1,†1,a)] Nobutaka Ono[2,3,b)] Akira Nishimura[4,c)] Isao Echizen[2,3,d)]

Received: August 2, 2016, Accepted: April 10, 2017

Abstract: Reversible audio information hiding and sample-scanning methods are proposed for digital audio content to achieve detailed detection and localization of tampered positions in each frame. The method proposed in this study allows detecting multiple tampering and reusing reliable content as well as avoiding false detection which were impossible for other methods to simultaneously achieve. In the proposed method, the original signal is partitioned into fixed-length frames and then transformed into discrete cosine transform (DCT) coefficients by the integer modified DCT (intDCT). Expansion of the DCT coefficients is applied to embed a content-based hash as a payload. The integer DCT algorithm ensures the reversibility of the transform so that the original data and embedded payload can be perfectly restored to enable blind verification of the data integrity. The perceptual evaluation of speech quality (PESQ) with the listening quality objective mean opinion (MOSLQO), the segmental signal to noise ratio (segSNR), and subjective evaluation results show that the proposed algorithm provides good sound quality (MOSLQO and segSNR are respectively 4.41 and 23.31 dB on average for a capacity of 8,000 bps). Detection and localization are accurate in terms of correctly localizing tampered frames in case of insertion or deletion.

Keywords: reversible audio information hiding, tampering detection and localization, DCT coefficient expansion

1. Introduction

The rapid development of multimedia technologies has made it easier to create, replicate, transmit, and distribute digital content. In most application scenarios, it is crucial to guarantee the integrity of important data [1], [2], [3]. Examples include medical records and surgery videos [1], police investigations and witness depositions, interviews and telephone conversations which must be reliable in order to prevent fraud [2], [3]. These materials must be reliable and their integrity should be guaranteed. The proposed method is intended to serve as an alternate recording solution that provides authentication. If tampering with original data is suspected then stego data can be applied to detect it. The proposed method utilizes a reversible algorithm so embedded information can be cleanly removed and the the original data restored for use as evidence.

Tampered content submitted as evidence to a court may cause incorrect judgments and false accusations. To prevent or detect tampering, digital signatures [4], information hiding [1], [2], [3], [5], [6], [7], [8], [9], [10], [11], [12], and other alternatives based on noise, device, and environment identification [13], [14], [15] are widely used. In digital signature systems, a digital signature is appended to the header of the content and once it is removed, the content can no longer be verified. Technologies [13], [14], [15] can be used to detect whether material has been spliced with recordings from different acoustic environments or recording devices. In information hiding methods, the information for verification is embedded within the content itself. Watermarking-based methods for verification have been proposed for image and video data [5], [6], [7]. Audio authentication has also been a subject of interest [9], [12].

However, there are some drawbacks in the conventional methods of authentication [2], [7], [12]. In Ref. [2], the fingerprint of an entire file was embedded within the file. In Ref. [7], a time stamp was embedded into frames of a video. In Ref. [12], the frame number was embedded into audio frames.

In this paper, we propose a sample-scanning method for the detection and localization of tampered positions by relocalizing the starting point of the sampling by indexing to the next non-tampered frame. The starting point of sampling is shifted by a sample each time to carry out extraction, reconstruction, and verification until the tampered positions are localized correctly, and assigned to each frame, which makes it possible to reconstruct and reuse the remaining non-tampered frames even with multiple tampering, and the verification of short parts of the stego data is possible.

Additionally, the reversibility of the original data with little distortion is essential owing to the probative or authenticating importance of the data. If an algorithm is lossy, the original data cannot be reconstructed from stego data. This means that the algorithms used to hide the payload, including information for verification and positional data, extract the payload, and recon-

[1] The Kyoto College of Graduate Studies for Informatics (KCGI), Kyoto 606–8225, Japan
[2] National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[3] SOKENDAI (School of Multidisciplinary Sciences, The Graduate University for Advanced Studies), Chiyoda, Tokyo 101–8430, Japan
[4] Department of Information Systems, Tokyo University of Information Sciences, Chiba 265–8501, Japan
[†1] Presently with Organization for the Strategic Coordination of Research and Intellectual Properties, Meiji University
[a)] huang_xp@meiji.ac.jp
[b)] onono@nii.ac.jp
[c)] akira@rsch.tuis.ac.jp
[d)] iechizen@nii.ac.jp

struct the original data must be lossless. Reversible information hiding for the authentication of images and videos has been proposed [1], [2], [7], [10], [11]. Hiding based on a modified discrete cosine transform (DCT) [16] in the audio field is a viable alternative method used for a reference in this study. Embedding a payload into audio data by expanding the intDCT coefficient to reserve hiding capacity was preliminarily investigated in Ref. [17].

This paper contributes to resolving the above problems by both accurately detecting and localizing tampering positions which has not been achieved up to now. We propose a reversible audio information hiding method based on intDCT with framewise partition and a sample-scanning method to localize tampered positions. DCT coefficients corresponding to high frequencies are expanded to hide the payload including information for verification and positional data. Detection and verification are achievable using the positional data, and the remaining reliable data can be reconstructed and reused. Detection effectiveness experiments involving the insertion and deletion of up to 112 speech data show that the proposed method is valid with an average false alarm rate (FAR) of approximately 50%.

Moreover, audio distortion due to the framewise partition and hiding capacity are evaluated objectively with the PESQ and AFsp packages. Comparison of the quality of data with an existing method based on linear predictive coding (LPC) shows that the proposed algorithm provides comparable quality with an average listening quality objective mean opinion (MOSLQO) scores of 4.41 and a better average segmental signal-to-noise ratio (segSNR) scores of 23.31 dB for 112 speech data and a capacity of 8,000 bps. According to the results, the amount of distortion is below a perceptible level and the stego data are comprehensible.

This paper is organized as follows. Section 2 introduces the proposed sample-scanning algorithm. Section 3 describes the proposed method based on intDCT. Section 4 summarizes the experimental results for detection effectiveness, reversibility, computational cost, and audio quality. Session 5 concludes this paper.

2. Current Tampering Detection Methods

2.1 Definition of Tampering Detection and Localization

Let $c(j)$ be the j-th sample of the original data and $c'(j)$ be the j-th sample of the signal extracted from the stego data $s(j)$ in the time domain, where j ($j \in 1, 2, 3, \ldots,$) is an integer. If there is no tampering, $c'(j) == c(j)$. If $c'(j) \neq c(j)$, then tampering has occurred in the j-th sample of the stego data. We detect location with $c'(j) \neq c(j)$ as a detected tampering position and localize the range of data (including location of j) where tampering occurred as well as localizing the next start sampling of the next non-tampered frames as tampering localization.

2.2 Current Tampering Detection Methods Involving Reversible Information Hiding

The simpler conventional tampering detection method involving information hiding was proposed in Ref. [2]. A fingerprint of a file was embedded as a payload to generate stego data. Detection was achieved by comparing the extracted fingerprint with the embedded fingerprint. However, if the data at any location had been tampered with, the fingerprint of the file was broken and it was impossible to detect which part of the content was suspicious.

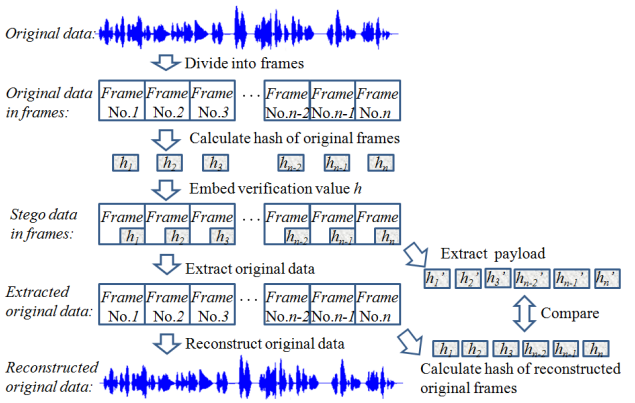


Fig. 1 Reversible audio information hiding for detecting tampering within each frame.

2.3 Principle of Framewise Tampering Detection and Problems

To localize tampering in detail, in Refs. [7], [8], [12], [17] the original data were divided into frames. Details of framewise tampering detection were introduced in Ref. [17] as follows. The audio data are first segmented into frames with a fixed length N to detect and localize tampering in a framewise unit. On the encoder side, information for verification is embedded into the frame itself in a reversible manner. On the decoder side, both the embedded information and the original data are extracted and reconstructed. The integrity of each frame can then be independently checked by blindly comparing the extracted information for verification and that of the reconstructed original data. If they are not equal, tampering has probably occurred. This principle has been discussed in a previous work [17] and is illustrated in **Fig. 1**. The advantage of framewise partition is that it is able to localize the modified positions in detail in each frame. The disadvantages are that (1) once tampering has occurred and been detected, the remaining part cannot be used for verification since the index of the next reliable frame cannot be relocalized, which leads to false detection; (2) the payload is not relevant to the content in that once the payload is modified rather than the content, it causes false detection.

2.4 False Detection Problem of Current Framewise Tampering Detection Methods

To detect and localize tampering precisely, there is another problem to solve, which is false detection. Once insertion or deletion occurs, the length of the data is changed, which means that a part of the frames of stego data are repartitioned with a shifted sampling start point. Thus, the extracted hash value is different from that of the reconstructed original data even though the content has not been tampered with. This causes false detection. **Figure 2** shows a case of false detection caused by deletion. Schemes [7], [8], [12], [17] divided data into frames and investigate whether the extracted watermarks equal to the original watermark, however, these works cannot avoid the false detection problem plotted in Fig. 2.

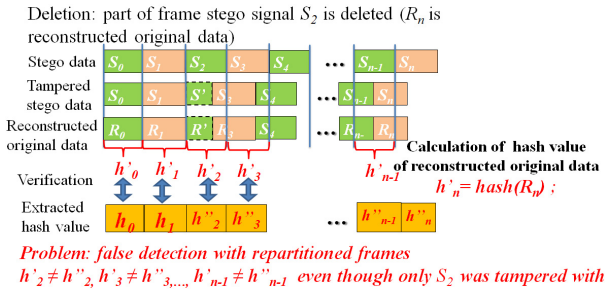


Fig. 2 Example of false detection caused by deletion.

2.5 Discussion on Payload

In conventional methods, fingerprint information [2], time sequence information [7], hash values [1], [10], and binary numbers [11] are widely used as the payload to verify the integrity of original data. Bitmap image is used as watermark in Ref. [8]. Scheme [17] used a content-based hash value as the verification data.

Independent payload [1], [2], [7], [8], [10], [11] and content-based payload [17] have disadvantages and advantages. Independent payload is convenient to use and no additional calculation is necessary if the payload is shared in advance. However, these kinds of payload are vulnerable in case that the content is tampered with while the payload is kept as the same in case the hiding algorithm is disclosed. In this case, tampering cannot be detected. Content-based payload can avoid this problem since once content is tampered with, the authentication payload differs according to the tampering that was detected. However content-based payload has the drawback that it cannot detect kinds of tampering that deletion and frame order exchanging occurs from the exact start sampling point to the exact end sampling point of frames.

A combination of independent payload and content-based payload maybe a solution. For example, a content-based hash value combined with frame numbers may solve the drawbacks discussed above. In this paper, since the main purpose is to verify the effectiveness of sample-scanning method to detect and localize multiple tampering, content-based hash value is used as the payload. An extension for payload exploring will be a subject for future work.

3. Proposed Sample-scanning Algorithm for Tampering Detection and Localization

3.1 Principle of Proposed Sample-scanning Algorithm for Tampering Detection and Localization

Relocalization of the correct sampling by indexing it to the start sampling of next non-tampered frame is necessary to avoid false detection. In this paper, we propose a sample-scanning method to achieve the detection and localization of tampered positions. It involves localizing the starting point for sampling by repeatedly shifting a sampling point in the time domain for reconstruction and verification between the hash value of the reconstructed original data and the extracted hash value until they equal. By doing this, even if local tampering is found, the remainder of the audio data is available and reusable.

Suppose the fixed frame length is N and the length of the original data is Len, then the number of frames in the original data is $\frac{Len}{N}$. Input data is samplings of tampered stego data $s'(m)$; output is the frame number with comparison results of matched and unmatched and reconstructed reliable WAV data $s_r(m)$ with tampered signal removed. Here, m ($1 \leq m \leq \frac{Len}{N}$) is the frame number of signals during verification. Frame start index $start$ ($1 \leq start \leq Len - N + 1$) is initialized to be 1, and then frame end index end ($N \leq end \leq Len$) is $start + N - 1$ and $m = \lceil \frac{start}{N} \rceil$, where $\lceil \cdot \rceil$ denotes the ceil function. The algorithm for verification and reconstruction is as follows:

- STEP 1) Extract embedded hash data h'_m and reconstruct original data $c'(m)$ from $s'(m)$. Then hash value of $c'(m)$ is calculated as h''_m.
- STEP 2) Compare h'_m and h''_m. In case $h'_m == h''_m$, set $start = end + 1$ and $end = start + N - 1$ and comparison result is set to be matched with frame number m memorized; In case $h'_m \neq h''_m$, set $start = start + 1$ and $end = end + 1$ and comparison result is set to be unmatched with frame number m memorized. In this case, STEP 1) is repeated following with the remaining $s'(m)$ until $h'_m == h''_m$.
- STEP 3) Output the comparing result of matched frame number and unmatched number and generate $s_r(m)$.

For both of insertion and deletion, the process of sampling scanning stopped when $start$ indicates the first sampling of non-tampered frame correctly.

The proposed sample scanning method makes localization of tampering possible in detail. The index of the start sampling of frame is shifted to localize the non-tampered frame and this algorithm is processed in a loop once the hash value is unmatched. Ordinary verification by comparison of hash values is run for the non-tampered frames. Thus, multiple tampering detection is possible to detect tampering towards a data in multiple places at the same time. This multiple tampering detection has not been achieved up to now in conventional methods. **Figure 3** gives an example of detecting multiple detection of insertion and deletion by the proposed scanning algorithm.

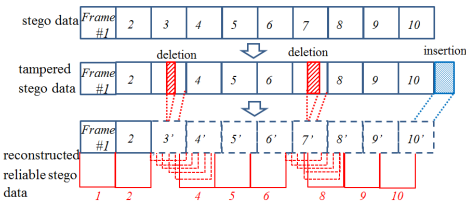


Fig. 3 Illustration of sample-scanning for detecting multiple tampering.

3.2 Implementation with Integer DCT for Reversible Information Hiding

The hiding algorithm is the preliminary process used for detecting tampering. Many information hiding methods can detect tampering. In order to achieve reversible information hiding with greater robustness against attack, we used integer modified DCT IV in this method.

To meet the requirement of reversibility in audio information hiding, we studied the conventional methods [18], [19], [20], [21]. In these methods, different domains for embedding data have been proposed which affect the quality of stego data and the hiding capacity. Aoki proposed a method of hiding data in

sign bits [18]. Yan and Wang proposed a method of expanding the residual between the predicted signal and the original signal based on LPC [19]. Nishimura [20] extended the algorithm of Yan and Wang's work. Unoki and Miyauchi proposed a method of hiding data in phase information [21]. These studies hide data in the time domain, which may directly distort audio data. Another alternative method based on intDCT uses the DCT domain to hide information, which has been commonly used in image fields [22], [23], [24] to achieve high capacity and low distortion by expanding the DCT coefficients. In the audio field, the modified DCT is also used [16]. As an overall trend of audio data, the amplitude of DCT coefficients decreases when the frequency increases. Thus, expanding the amplitude of DCT coefficients corresponding to higher frequencies causes lower distortion, which makes the DCT coefficients a viable domain for audio information hiding.

3.2.1 Reversible Transform with intDCT

A modified DCT type IV is used for data transform. Let

$$\boldsymbol{c} = (c(1)\ c(2)\ \ldots\ c(N))^T \tag{1}$$

$$\boldsymbol{H} = (H(1)\ H(2)\ \ldots\ H(N))^T \tag{2}$$

be a time-domain signal at an N-point frame and its DCT coefficients, respectively. In a continuous case, we can obtain the DCT coefficients $\boldsymbol{H}$ from the time-domain signal $\boldsymbol{c}$ using the modified DCT matrix as

$$\boldsymbol{H} = C_N^{DCT-IV}\boldsymbol{c}, \tag{3}$$

where the (i,t)-th element $(1 \le i \le N, 1 \le t \le N)$ of the modified DCT matrix.

C^{DCT-IV} is represented as

$$C_N^{DCT-IV}(i,t) = \sqrt{\frac{2}{N}}\left[\left(\cos\frac{(t+\frac{1}{2})(i+\frac{1}{2})\pi}{N}\right)\right]. \tag{4}$$

The reversibility of intDCT is based on factorization [25].

$$C_N^{DCT-IV} = R_1R_2R_3T_1T_2T_3 \tag{5}$$

Here, R_1, R_2, R_3, T_1, T_2, and T_3 are the block triangular matrices defined in Ref. [16].

Appendix supplies an example of multiplying a triangular matrix followed by rounding and shifting operations. As shown, the result is reversible even though elements of the triangular matrix for multiplying are not integers.

3.2.2 Information Hiding Using DCT Expansion

Suppose a bit of information b is embedded in the i-th DCT coefficient $H(i)$ of the original data to generate stego data of the i-th DCT coefficient $S(i)$ by the expansion: $S(i) = 2H(i) + b$, where $S(i)$ is the i-th DCT coefficient of the stego data.

For extraction and reconstruction, the embedded payload $b(i) = S(i) - 2H(i)$ and the original data can then be reconstructed using $H(i) = \lfloor S(i)/2 \rfloor$, where $\lfloor\cdot\rfloor$ denotes the floor function.

The original data are segmented into frames with length N to hide content-based hash values for detection. Inverse intDCT is then applied to transform the reconstructed original data in frames from the DCT domain $H(i)$ $(1 \le i \le N)$ to the time domain $c(i)$ $(1 \le i \le N)$ to reconstruct the original data $c'(i)$. Details are given in Ref. [17]. Hash values of the reconstructed $c'(i)$ are then compared with the extracted hash values for verification.

4. Experimental Evaluation

4.1 Data Used for Experiments

We focus on speech data with probative importance for evaluation. We used a dataset from ITU-T Test Signals for Telecommunication Systems–Test Vectors Associated to Rec. ITU-T P.50 Appendix I [26]. This dataset includes 16-kHz-sampled and 16-bit quantized waveforms. We used 112 speech signals (16 speakers in seven languages: American English, Arabic, Mandarin Chinese, Danish, French, German, and Japanese). The average length of each track was approximately 10 s. Each piece of data was normalized by a data length that was an integer multiple of the frame length.

4.2 Effectiveness of Tampering Detection

4.2.1 Purpose of Experiments

The purpose of these experiments was to detect tampering by insertion and deletion and to localize the tampered positions. The effectiveness of detection was calculated. As output files, audible original data generated from the reliable parts of the tampered stego data were extracted and reconstructed.

4.2.2 Experimental Method

We performed experiments on tampering by insertion into and deletion from the 112 speech data. For insertion, a piece of white noise with a random length (shorter than 2,000 samples) was inserted into the stego data generated by the proposed method from a random starting point (among the first 10,000 samples). For deletion, a random length (shorter than 2,000 samples) of stego data was deleted from a random starting point (among the first 10,000 samples).

4.2.3 Results for Effectiveness of Tampering Detection

The effectiveness of tampering detection was evaluated by analyzing the *Miss Rate* and *FAR*. In Ref. [12], three experiments were carried out in which data were replaced in the $50{,}000^{th}$-$60{,}000^{th}$ samples, data were deleted in the $100{,}000^{th}$-$130{,}000^{th}$ samples, and 30,000 samples were inserted from the $600{,}000^{th}$ sample, and the results of detection using a conventional method were located in the 6^{th}-8^{th}, 12^{th}-16^{th}, and 72^{th}-75^{th} frames, without a clear definition of the effectiveness. In Ref. [27], the detection effectiveness was calculated to be 100% if tampered samples (discontinuity) were located in the range of detected tampered samples.

We assume that T samples are inserted into (deleted from) stego data from the K^{th} sample, n_t (as T) is the number of tampered samples, and n_d is the number of detected tampered-samples. The symbols used are listed in **Table 1**.

Since the detection is based on the hash value of each frame, the detected tampered samples start from sample $N * (m-1) + 1$

Table 1 Symbols used in Figs. 4 and 5.

Symbols	Meaning
$P0$	end sample of non-tampered frame
$P1$	starting sample of tampered frame
$P2$	starting sample of non-tampered frame
K	starting sample from which tampering occurred
$P3$	starting sample of non-tampered frame
T	number of tampered samples

Table 2 Examples of results of tampering detection test (insertion).

Data .wav	*Inserted from: K*	*Inserted length* n_t (as T)	*Detected from: P1*	*Detected to: P2-1*	*Detected range:* n_d(as P2-P1)	*Verify:* $N+T$	*Miss Rate* (%)	*FAR* (%): $1-\frac{n_t}{n_d}$
Da_f2	8,479	41	8,193	9,257	1,065	1,065	0	96.2
Ger_m2	1,822	825	1,025	2,873	1,849	1,849	0	55.4
Ch_m5	954	1,969	1	2,993	2,993	2,993	0	34.2

Table 3 Examples of results of tampering detection test (deletion).

Data .wav	*Deleted from: K*	*Deleted length* n_t (as T)	T mod N or $N-(T \bmod N)$ (if $T \bmod N > n_d$)	*Detected from: P1*	*Detected* to: $P2-1$	*Detected range:* n_d (as $P2-P1$)	*Verify:* $N*num-T$	*num*	*Miss Rate* (%)	*FAR* (%)
Ar_m5	5,282	14	14	5,121	6,130	1,010	1,010	1	0	98.6
Ch_f4	4,914	688	688	4,097	5,456	1,360	1,360	2	0	49.4
A_eng_f5	5,216	1,948	100	5,121	5,220	100	100	2	0	0

once when $h'_m \neq h''_m$ and end at the final tampered samples. Thus, if tampering localization is correctly achieved, the detected tampered samples will include and be larger in number than the tampered samples.

Miss Rate is used to investigate the proportion of samples that are tampered with but cannot be detected, which indicates tampered samples located out side the range of detected tampered samples. If *Miss Rate* is 0%, the detection and location are always correct. Moreover, if $P2-P1 = N+T$ for insertion and $P2-P1 = N*num-T$ for deletion, the localization has detected the tampered positions correctly. Here, the variable *num* ($num \in 1,2,3,\ldots$) indicates the number of frames deleted.

FAR is used to investigate the proportion of samples that are not tampered with but are counted as detected tampered samples since the localization can be only detailed to frames instead of samples. The closer *FAR* is to 0%, the more precise the tampering localization and the greater the quantity of non-tampered data can be reconstructed. The definition of *FAR* is as follows:

$$FAR = \begin{cases} 1-\frac{n_t}{n_d} & \text{(insertion)} \\ 1-\frac{n_t \bmod N}{n_d} & \text{(deletion, if } \frac{n_t \bmod N}{n_d} \leq 1\text{)} \\ 1-\frac{N-(n_t \bmod N)}{n_d} & \text{(deletion, if } \frac{n_t \bmod N}{n_d} > 1\text{)} \end{cases}$$

Here, adjustment of $N-n_t \bmod N$ is necessary to keep the value of *FAR* ranges between [0,1]. Otherwise, the *FAR* may be larger than 1. To verify whether the proposed method is effective in detecting and locating the tampered locations, there are two requirements to satisfy: 1) $K+1$ and $K+T$ must be located in the ranges of [P1, P2-1] for insertion, and $K+1$ must be located in the ranges of [P1, P2-1] for deletion, which means that the detected tampered locations has includes the actual tampering locations within its range and *Miss Rate* = 0%; otherwise, *Miss Rate* = 1; and 2) $n_d = N+n_t$ for insertion and $n_d = N*num-n_t$ for deletion, which means the tampered locations are detected exactly. Precise *num* is not necessary to calculate *FAR* and it is used to verify the effectiveness of localization. However, *num* can be accurately calculable by $num = \frac{n_d+n_t}{N}$.

Experimental results show that the tampering of the 112 data was detected and localized correctly by the proposed sample-scanning method. **Tables 2** and **3** respectively list three examples of results of tampering detection tests for insertion and deletion, whose detection effectiveness is closest to values of the minimum, average and maximum.

Even when *FAR* becomes large, if *Miss Rate* is 0, the localization correctly includes tampered samples in the range of detected tampered samples. A shorter frame length N can achieve a lower *FAR*; however, a longer calculation time is required for sample scanning.

Furthermore, although non-tampered reliable frames can be reconstructed successfully, the proposed method cannot distinguish whether the tampering involves deletion or insertion.

Illustrations of the detection of insertion and deletion (N = 1,024) are shown in **Figs. 4** and **5**, respectively.

The experimental results of average detection *FAR* are 55.6% for insertion and 49.5% for deletion. The best value of *FAR* is 0% (deletion). In this case, 900 samples were deleted from the 1,107th sample and the detected range of tampering was from the

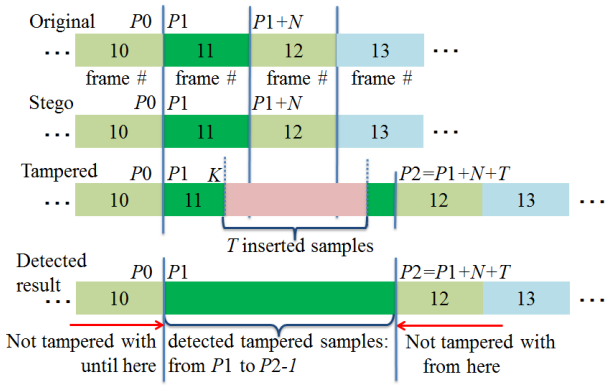


Fig. 4 Illustration of detection and localization of tampering by insertion: T samples are inserted from the $K+1^{th}$ sample in the 11^{th} frame.

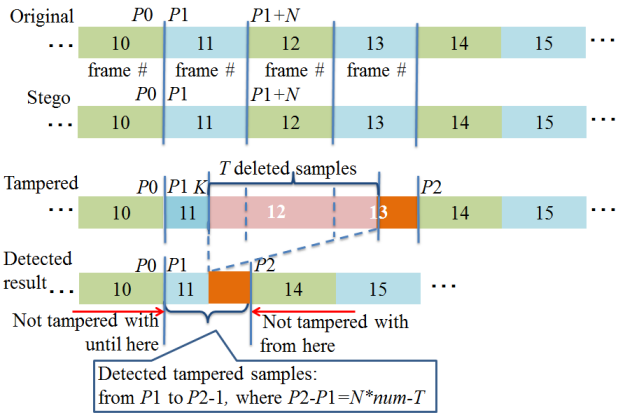


Fig. 5 Illustration of detection and localization of tampering by deletion: T samples are deleted from the $K+1^{th}$ sample in the 11^{th} frame.

Table 4 Comparison of tampering detection effectiveness.

Works	Tampering Localization	Multiple Detection	Blind Detection	Partial Detection
Gomez et al. [2]	NO	NO	NO	NO
Echizen et al. [7]	YES	NO	NO	YES
Unoki et al. [8]	YES	NO	YES	YES
Han et al. [12]	YES	NO	NO	YES
Huang et al. [17]	YES	NO	YES	NO
Gartner et al. [27]	YES	NO	NO	NO
Proposed	YES	YES	YES	YES

$1{,}025^{th}$ sample to the $1{,}148^{th}$ sample, with a length of 124 as T. As $1 - \frac{T \mod N}{N*num-T} = 1 - \frac{124 \mod 1{,}024}{1{,}024*1-900} = 0\%$, the number of detected tampered samples was exactly equal to the number of tampered samples.

There are also other types of attacks, for example, the replacement of stego data, changing the frame order, and replacement of the original data if the hiding algorithm is disclosed. With a more complex payload, it is possible to detect complex tampering. However, we do not discuss the relationship between the payload and detectable tampering in this paper.

4.2.4 Comparison of Tampering Detection Effectiveness with Conventional Methods

We compare the proposed method with conventional methods of tampering detection in terms of a number of issues as shown in **Table 4**. The proposed sample-scanning method has the following achievements that have not been simultaneously achieved in previous methods.

1) Relocalization of the starting sample of the non-tampered frame has been achieved to avoid false detection and the remaining reliable data can be reconstructed and reused.

2) Part of the data can be clipped as a target for detection.

3) Detection of multiple tampering of the data is theoretically achievable.

4.3 Evaluation of Audio Quality

To evaluate the audio quality of stego data, we mainly use PESQ and segSNR in this paper, which have been extensively used to objectively evaluate the sound quality [19], [20]. As a time-domain-based measure, segSNR is a method for checking the distortion caused by differences in the time domain, by comparing original data and stego data sample by sample. To evaluate the listening quality of the speech data, we used MOSLQO, which is an objective technique defined by ITU-T Recommendation P.862.1. MOSLQO scores are obtained by a mapping from MOS scores and range from 1.02 (lowest quality) to 4.56 (highest quality). For an objective evaluation using MOSLQO scores, we used PESQ version 1.2 [28]. For evaluation of the segSNR, which is defined as the average of SNR value over segments, we used AFsp package version 9.0.

The frame length is considered to affect both the precision of localizing tampered positions and the audio quality. To determine the effect of the frame length on the audio quality, signals were segmented into frames with lengths of 2,048, 1,024, and 512 for evaluation.

MOSLQO and segSNR are used to evaluate the quality of stego data generated by the proposed method and conventional method, with original data as the references. The results of the comparison of the effectiveness of the frame length on the audio quality for the proposed method and an LPC-based method [20] are listed in **Table 5**. As shown, when the frame length is shorter, MOSLQO and segSNR deteriorate. However, even when the frame length is set to 512, the average MOSLQO is 4.27, which falls in the range between "imperceptible" and "perceptible but not annoying" and the average segSNR is 22.23 dB, which means a clear audio quality. The proposed method generally has a better segSNR but worse MOSLQO than the LPC-based method [20].

Table 5 Comparison with conventional method of quality (average MOSLQO and segSNR) of stego data for different frame lengths using 112 signals (capacity ≈ 8,000 bps).

Method	MOSLQO	segSNR (dB)	Frame length
LPC [20]	4.50	16.22	2,048
	4.48	16.11	1,024
	4.45	16.04	512
Proposed	4.41	23.31	2,048
	4.34	22.99	1,024
	4.27	22.23	512

4.4 Evaluation of Reversibility

Reversibility is assured by the integer modified DCT [25] theoretically as shown in the Appendix. We performed experiment to verify reversibility of the proposed method by computing the differences between the data. Data are read in time domain as matrix in MATLAB and subtraction is applied to extracted embedded data and the embedded data, reconstructed original data and original data to calculate the differences. All of the data result no differences, which means the reversibility of all dataset are verified.

5. Conclusion

We propose a reversible information hiding method for detecting and localizing tampering positions, and we evaluated its effectiveness for the cases of insertion and deletion. We expanded the DCT coefficients in higher DCT blocks to achieve good audio quality. Using the proposed sample-scanning method, the remaining reliable data can be reconstructed and reused.

Audio quality with imperceptible distortion was also achieved. However, since the expansion region is concentrated in higher DCT blocks, the hiding positions can be easily estimated by referring to a histogram of the stego data. A topic for future work is therefore developing a sophisticated algorithm to the surrounding context so that they are more difficult to predict.

References

[1] Kaur, M. and Kaur, R.: *Reversible watermarking of medical images: Authentication and recovery-A survey, Journal of Information and Operations Management*, Vol.3, No.1, pp.241–244 (2012).

[2] Gomez, E., Cano, P., Gomes, L.D., Batlle, E. and Bonnet, M.: Mixed watermarking fingerprinting approach for integrity verification of audio recordings, *International Telecommunications Symposium* (*ITS 2002*), Natal, Brazil (2002).

[3] Kalker, T., Haitsma, J.A. and Oostveen, J.C.: Robust audio hashing for content identification, *Content Based Multimedia Indexing* (*CBMI*), pp.2091–2094, Italy (2001).

[4] Pramatejakis, M., Oelbaum, T. and Diepold, K.: Authentication of MPEG-4-based surveillance video, *Proc. International Conference on Image Processing*, Vol.1, pp.33–37 (2004).

[5] Celik, M.U., Sharma, G. and Tekalp, A.M.: Lossless watermarking for image authentication: A new framework and an implementation,

IEEE Trans. Image Processing, Vol.15, No.4, pp.1042–1049 (2006).
[6] Hwang, J.H., Kim, J.W. and Choi, J.U.: A reversible watermarking based on histogram shifting, *Proc. International Workshop on Digital Watermarking*, *LNCS 4283*, pp.348–361 (2006).
[7] Echizen, I., Yamada, T., Tezuka, S., Singh, S. and Yoshiura, H.: Improved video verification method using digital watermarking, *Proc. International Conference of Intelligent Information Hiding and Multimedia Signal Processing*, pp.445–448 (2006).
[8] Unoki, M. and Miyauchi, R.: Detection of tampering in speech signals with inaudible watermarking technique, *Proc. International Conference of Intelligent Information Hiding and Multimedia Signal Processing*, pp.118–121 (2012).
[9] Chen, O.T.C. and Liu, C.H.: Content-dependent watermarking scheme in compressed audio with identifying manner and location of attacks, *IEEE Trans. Audio, Audio, and Language Processing*, Vol.15, No.5, pp.1605–1616 (2007).
[10] Caldelli, R., Filippini, F. and Becarelli, R.: Reversible watermarking techniques: An overview and a classification, *EURASIP Journal on Information Security*, Vol.2010, pp.1–19 (2010).
[11] Zeng, X., Chen, Z.Y., Chen, M. and Xiong, Z.: Reversible video watermarking using motion estimation and prediction error expansion, *Journal of Information Science and Engineering*, Vol.27, pp.465–479 (2011).
[12] Han, B. and Gou, E.J.: A robust speech content authentication algorithm against desynchronization attacks, *Journal of Communications*, Vol.9, No.9, pp.723–728 (2014).
[13] Pan, X., Zhang, X. and Lyu, S.: Detecting splicing in digital audios using local noise level estimation, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp.1841–1844 (2012).
[14] Malik, H.: Acoustic environment identification and its applications to audio forensics, *IEEE Trans. Information Forensics and Security*, Vol.8, No.8, pp.1827–1837 (2013).
[15] Cuccovillo, L., Mann, S., Tagliasacchi, M. and Aichroth, P.: Audio tampering detection via microphone classification, *Proc. IEEE 15th Workshop on Multimedia Signal Processing* (*MMSP*), pp.177–182 (2013).
[16] Geiger, R., Yokotani, Y. and Schuller, G.: Audio data hiding with high data rates based on IntMDCT, *Proc. Acoustics, Audio, and Signal Processing*, pp.205–208 (2006).
[17] Huang, X.P., Echizen, I. and Nishimura, A.: A reversible acoustic steganography scheme to authenticate use, *digital watermarking, Lecture Notes in Computer Science LNCS 6526*, Kim, H.-J., Shi, Y. and Barni, M. (Eds.), Vol.2011, pp.305–316, Springer-Verlag, Berlin Heidelberg (2011).
[18] Aoki, N.: A technique of lossless steganography for G.711, *IEICE Trans. Commun.*, Vol.E90-B, No.11, pp.3271–3273 (2007).
[19] Yan, D.Q. and Wang, R.D.: Reversible data hiding for audio based on prediction error expansion, *Proc. International Conference of Intelligent Information Hiding and Multimedia Signal Processing*, pp.249–252 (2008).
[20] Nishimura, A.: Reversible audio data hiding using linear prediction and error expansion, *Proc. International Conference of Intelligent Information Hiding and Multimedia Signal Processing*, pp.318–321 (2011).
[21] Unoki, M. and Miyauchi, R.: Reversible watermarking for digital audio based on cochlear delay characteristics, *Proc. International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp.314–317 (2011).
[22] Yang, B., Schmucker, M., Niu, X.M., Busch, C. and Sun, S.G.: Reversible image watermarking by histogram modification for Integer DCT coefficients, *Proc. Workshop on Multimedia Signal Processing*, pp.143–146 (2004).
[23] Lin, C.C. and Shiu, P.F.: High capacity data hiding scheme for DCT-based images, *Journal of Information Hiding and Multimedia Signal Processing*, Vol.1, No.3, pp.220–240 (2010).
[24] Chang, C.C., Chen, T.S. and Chung, L.Z.: A Steganographic Method Based upon JPEG and Quantization Table Modification, *Information Sciences*, Vol.141, pp.123–138 (2002).
[25] Haibin, H., Susanto, R. and Rongshan, Y.: A fast algorithm of integer MDCT for lossless audio coding, *Proc. IEEE International Conference on Acoustics, Audio and Signal Processing* (*ICASSP*), pp.177–180 (2004).
[26] ITU-T Test Signals for Telecommunication Systems – Test Vectors Associated to Rec. ITUT P.50 Appendix I, available from ⟨http://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm⟩.
[27] Gartner, D., Ditmar, C., Schroth, P., Cuccovillo, L., Mann, S. and Schuller, G.: Efficient cross-codec framing grid analysis for audio tampering detection, *Proc. 136th Audio Engineering Society Convention*, pp.306–316 (2014).
[28] Perceptual evaluation of audio quality (PESQ): An objective method for end-to-end audio quality assessment of narrow-band telephone networks and audio codecs, ITU-T Recommendation P.862, International Telecommunication Union (2001).

Appendix
Reversibility of Integer Modified DCT-IV

Let c be a time-domain signal at an N-point frame and H represent its DCT coefficients, which are represented by Eqs. (A.1) and (A.2).

$$\boldsymbol{c} = (c(1)\ c(2)\ \ldots\ c(N))^T \tag{A.1}$$

$$\boldsymbol{H} = (H(1)\ H(2)\ \ldots\ H(N))^T \tag{A.2}$$

c_n means the n-th sampling point of original data, and H_n means the n-th DCT coefficient. Suppose we have $H = Kc$, $I_N = diag(1, \ldots, 1)$, $A_N = a$, (a is an arbitrary rational number, including integer and non-integer), and $Z_N = 0$. As a triangular matrix, $K = \begin{pmatrix} I_{N/2} & Z_{N/2} \\ A_{N/2} & I_{N/2} \end{pmatrix}$, then $\begin{pmatrix} H_1 \\ H_2 \\ \vdots \\ H_N \end{pmatrix} = \begin{pmatrix} I_{N/2} & Z_{N/2} \\ A_{N/2} & I_{N/2} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{pmatrix}$. Then, we may have

$$H_1 = c_1 \tag{A.3}$$

$$H_2 = ac_1 + c_2$$

$$H_N = a(c_1 + c_2 + \ldots + c_{N-1}) + c_N \tag{A.4}$$

There are rounding calculations after expanding each nontrivial sub-matrix, and then we have

$$H_1 = \mathrm{round}(c_1) = c_1 \tag{A.5}$$

$$\begin{aligned} H_N &= \mathrm{round}(a(c_1 + c_2 + \ldots + c_{N-1})) + \mathrm{round}(c_N) \\ &= \mathrm{round}(a(c_1 + c_2 + \ldots + c_{N-1})) + c_N \end{aligned} \tag{A.6}$$

We have $c' = K^T H$ because of the lifting calculation: $\begin{pmatrix} c'_1 \\ c'_2 \\ \vdots \\ c'_N \end{pmatrix} = \begin{pmatrix} I_{N/2} & Z_{N/2} \\ -A_{N/2} & I_{N/2} \end{pmatrix} \begin{pmatrix} H_1 \\ H_2 \\ \vdots \\ H_N \end{pmatrix}$. After applying rounding, we have

$$c'_1 = \mathrm{round}(H_1) \tag{A.7}$$

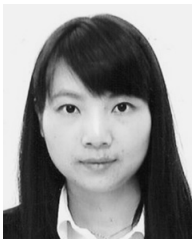
$$\begin{aligned} c'_N &= \mathrm{round}(-a(H_1 + H_2 + \ldots + H_{N-1})) \\ &\quad + \mathrm{round}(H_N). \end{aligned} \tag{A.8}$$

Put Eqs. (A.3) and (A.4) into Eqs. (A.7) and (A.8), and then we have

$$c'_1 = \mathrm{round}(H_1) = \mathrm{round}(c_1) = c_1 \tag{A.9}$$

$$\begin{aligned} c'_N &= \mathrm{round}(-a(H_1 + H_2 + \ldots + H_{N-1})) + \mathrm{round}(H_N) \\ &= \mathrm{round}(-a(H_1 + H_2 + \ldots + H_{N-1})) \\ &\quad + \mathrm{round}(a(H_1 + H_2 + \ldots + H_{N-1})) + c_N \\ &= c_N \end{aligned} \tag{A.10}$$

Then reversibility of the transform has been verified. In case $K = \begin{pmatrix} I_{N/2} & B_{N/2} \\ Z_{N/2} & I_{N/2} \end{pmatrix}$, where $B_N = b$, (b is an arbitrary rational number), the result is the same. According to Eq. (A.10), no matter what value a and b are, $c'_N = c_N$, which promises reversibility.



Xuping Huang received B.S. and B.A. degrees from the Department of Software Science, Dalian JiaoTong University, China in 2007 and an M.S. degree from the Department of Information Science, Graduate School of Iwate Prefectural University, Japan in 2009. Since 2009, she has been a Ph.D. candidate at the Graduate University for Advanced Studies (SOKENDAI), a research assistant at the National Institute of Informatics (NII) during May, 2009–Sep., 2013 and a technical staff at the National Institute of Advanced Industrial Science and Technology (AIST), Japan during Oct., 2013–Mar., 2014. After working as an Assistant Professor in The Kyoto College of Graduate Studies for Informatics (KCGI) during Apr., 2014–Mar., 2017, she is currently a researcher in Meiji University. Her research interests include information hiding and audio signal processing. She received the Best Paper Award from CCIT in 2014. She is a member of IEICE and IPSJ.

Nobutaka Ono received B.E., M.S., and Ph.D. degrees in Mathematical Engineering and Information Physics from the University of Tokyo, Japan, in 1996, 1998, and 2001, respectively. He joined the Graduate School of Information Science and Technology, the University of Tokyo, Japan, in 2001 as a Research Associate and became a Lecturer in 2005. He moved to the National Institute of Informatics, Japan, as an Associate Professor in 2011. His research interests include microphone array processing, source localization and separation, and optimization algorithms for them. He is the author or co-author of more than 180 papers published in international journals and peer-reviewed conference proceedings. He has been a member of the IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee since 2014. He received the Measurement Division Best Paper Award from SICE in 2013, the Best Paper Award from IEEE IS3C in 2014, the Excellent Paper Award from IIHMSP in 2014, and the Unsupervised Learning ICA Pioneer Award from SPIE.DSS in 2015.

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 was awarded the Sato Prize by the Acoustical Society of Japan in 2012.

Isao Echizen received B.S., M.S., and D.E. degrees from the Tokyo institute of technology, Japan, in 1995, 1997, and 2003. He joined Hitachi, Ltd. in 1997 and until 2007 was research engineer in company's systems development laboratory. He is currently a Professor of the National institute of informatics (NII) and a Professor of the Graduate University for Advanced Studies (SOKENDAI). He was a visiting professor at University of Freiburg in 2010 and a Visiting Professor at the University of Halle-Wittenberg in 2011. He has been engaged in research on information security and content security and privacy. He received the Best Paper Award from IPSJ in 2005 and 2014, the Fujio Frontier Award and the Image Electronics Technology Award in 2010, the One of the Best Papers on IFIP SEC and the IPSJ Nagao Special Researcher Award in 2011, the Docomo Mobile Science Award in 2014, and the Information Security Cultural Award in 2016.