A Study on Variable Embedding Locations of Reversible Spectral Speech Watermarking

Xuping Huang[†*] and Akinori Ito[†]
[*] Interdisciplinary Faculty of Science and Engineering, Shimane University, Matsue 690-8504, Japan
E-mail: huang@cis.shimane-u.ac.jp
[†] Department of Communications Engineering, Graduate School of Engineering, Tohoku University, Sendai 980-8577, Japan
E-mail: akinori.ito.a2@tohoku.ac.jp

***Abstract*—Rapid advancements in machine learning and speech synthesis have made it feasible to manipulate and tempering audio content. Although watermarking is a useful technique for guaranteeing the integrity of evidential audio data, distortion of sound quality must be prevented. A reversible digital watermarking technique based on the human auditory system was conventionally proposed to embed feature data for tamper detection in the high frequency coefficients. However, this technique has the drawback that spectral analysis can reveal the embedding locations through the borderline. This research explored variable embedding locations while preserving good sound quality. The experimental results show the effectiveness of the proposed to conceal the embedding location, and the distortion are controlled to be imperceptible.**

I. Introduction

With the speed at which digital technologies are developing these days, it is becoming increasingly important to prevent tampering with digital materials. Tampering issues can be effectively resolved by using the information concealing technique, which allows us to insert the payload information for integrity check to the content data itself, including audio [1-7], image, and video data [8-19]. For the information concealment approach, high capacity and minimal distortion are often preferred. Furthermore, original data must be available for a number of uses, including military, medical, and investigative records, as permanent degradation is unacceptable [16]. For these reasons, reversible information hiding has been proposed for such purposes.

To achieve reduced distortion, the domain in which the payload is integrated is an important consideration. In the audio field, time domain linear predictive coding [3,4] and inaudible domain based on cochlear delay property [5] have been presented. In the imaging area, the integer discrete cosine transform (intDCT) is well-known for its ability to produce high capacity and low distortion for reversible information concealing [14-18]. Motivated by them, we looked on intDCT-based reversible information hiding.

Because picture and audio have different statistical features and perceptual qualities, the method of embedding payload into DCT coefficients should have been adjusted differently. When it comes to images, embedding the payload into higher DCT coefficients leads in poor resilience, while integrating it into lower DCT coefficients results in poor invisibility. Thus, the intermediate area of DCT coefficients has been expanded for embedding. Humans are less sensitive to differences in high frequency coefficients in audio than they are in images. As a result, in our previous work [7], we embedded the payload data into higher DCT coefficients.

However, due to the concentration of concealing locations, which is a half of high DCT regions in the previous work [7], the border line generated by the expansion of DCT coefficients would be visible on the spectrum, which could be a hint for attackers to estimate a watermarking method. To avoid hostile attacks, hiding places should be unobtrusive. As a result, a sophisticated algorithm with complex concealing places is necessary, as well as maintaining high-quality stego data.

Lee recommended that the watermark be adaptively implanted into the image [16] for improved stego data quality and complex concealing locations. The findings of image quality with the Peak Signal-to-Noise Ratio (PSNR) criterion are superior to the no-adaptation scenario. In this research, we study various embedding locations, including random addition of expansion locations, and the efficiency of adaptive embedding of payload to specified DCT coefficients via distortion estimation in order to keep good sound quality as well as conceal the embedding locations.

The overview of our previous work is summarized in Section II. The proposed algorithm is presented in Section III. The experimental evaluation on audio quality is described in Section IV, and the conclusion and future work are described in Section V.

II. Overview of previous work: intDCT-based reversible audio information hiding

In our previous work [7], a modified integer DCT coefficients expansion based reversible digital watermarking method is proposed and implemented for tampering detection. Suitable coefficients are explored according to the audio feature. As a general trend of integer DCT Type IV, the higher the frequency domain is, the lower the amplitude of the audio data will be. Thus, in our previous work, to control the distortion of stego data to be imperceptible, we expanded the DCT coefficients in the high frequency domain to obtain the hiding locations for payload embedding. In order to detect tampering precisely, a given original audio data is first divided into frames with a fixed sample-length N (typically, 512, 1024, etc.) for tampering localization. Let x_n $(1 \leq n \leq N)$ be an audio signal

within a frame. It is transformed from time domain to DCT domain with integer DCT by our previous work[7]. The integer discrete cosine transform (intDCT) is a transform similar to the discrete cosine transform (DCT). Let

$$\mathbf{h} = (h(1)\ h(2)\ \ldots\ h(N))^T \quad (1)$$
$$\mathbf{H} = (H(1)\ H(2)\ \ldots\ H(N))^T \quad (2)$$

be a time-domain signal at an N-point frame and its DCT coefficients, respectively. In a continuous case, we can obtain DCT coefficients $\mathbf{H}$ from a time-domain signal $\mathbf{h}$ by DCT-IV matrix as

$$\mathbf{H} = C_N^{IV}\mathbf{h} \quad (3)$$

where the (i,t)-th ($1 \le i \le N$, $1 \le t \le N$) elements of the DCT matrix C^{IV} are represented as

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

In this paper, for DCT coefficients expansion based digital watermarking, Let X_n ($1 \le n \le N$) be DCT coefficients. When one bit of payload B is embedded to the n-th DCT coefficient X_n, the expansion technique is applied such as

$$X'_n = 2X_n + B. \quad (5)$$

After embedding, X'_n is transformed to time domain with the inverse integer DCT. The reversibility between x_n and X_n is guaranteed by the nature of the integer DCT. In our previous work [7], higher DCT coefficients are used because human is not so sensitive to the difference in high frequency coefficients. When payload is embedded into the upper half of DCT coefficients X_n ($N/2+1 \le n \le N$), the capacity about 0.5 bit per sample is achieved. However, since the DCT coefficients for expansion are concentrated on the high frequency domain, the borderline between the high frequency and the low frequency might be conspicuous to reveal the existence of the embedded payload by the spectral analysis, as Fig. 1 shown. We tried global steganalysis to check whether there is any signal embedded in another data. Figure 1 plots the a graphic about the original sound and stego signal of Ja_m5.wav, of some samples to compare how they change, where the horizontal axis represents the frame index and the vertical axis represents the order of DCT coefficients. It is obvious that when [9th, 16th] coefficients are expanded for embedding, by comparing the stego data and original data, the boundary for expansion and embedding might be visible.

III. Proposed Method: Variable Embedding Locations Exploration

On purpose to solve the boundary problem, and enhance the robustness towards to global steganalysis, we tried to explore variable embedding locations based on DCT coefficients expansion based digital watermarking. Our previous work [6] indicated that the random key of expansion may cause distortion to make the original data sounds noisy. Thus, distortion should be controlled to be imperceptible, since the original data is with probative importance, such as telephone recording used as evidence, and last will and testament, et al. Thus, we have to select appropriate locations for hiding the digest information for tampering detection, as well as considering the audio quality after expansion.

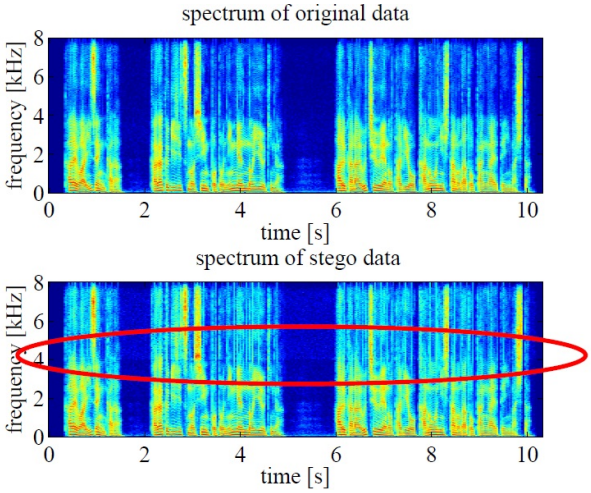

Fig. 1. Log of DCT coefficients difference between original data and stego data with a red mark on the borderline caused by expansion and embedding: Ja_m5.wav ($N = 2048$).

A. Solution 1: random addition of expansion blocks

Since the purpose is to keep the boundaries confidential, a random addition of expansion coefficients is supposed to be an efficient alternative solution. Figure 2 is an example of log amplitudes of DCT coefficients of the original speech data (Da_f2.wav, a female Germany speech data). The Y-axis is the DCT coefficients with a frame length N=2048. According to the feature of integer DCT type IV, amplitude tends to decrease as frequency increases. By utilizing this feature after transforming audio data from the time domain to the frequency domain, we could expand the amplitudes in the higher frequency domain, to reserve adequate hiding locations for payload to achieve imperceptibility. Our previous work [7] expanded the DCT coefficients from 1025-th to 2048-th coefficients to reverse embedding locations for embedding payload. In order to conceal the borderline, we hereby propose a solution to expand additional coefficients for hiding.

In order to extracted the data and reconstructed the original data, a location map is necessary here to mark the coefficients with expansion. We first segment N DCT coefficients to M blocks with an index i, ($1 \le i \le M$), indicating each block in order. The larger M is, the larger capacity for embedding the location map is required. Dividing N coefficients into M blocks can shorten the location map from N to M.

As the preliminary work, we tried to expand 512 coefficients, counting from the highest coefficients, out of 2048 coefficients in each frame, and we divided each 2048 coefficients into $M = 16$ blocks. This means 1/4 of coefficients (4 blocks) from the highest coefficients are expanded. With the

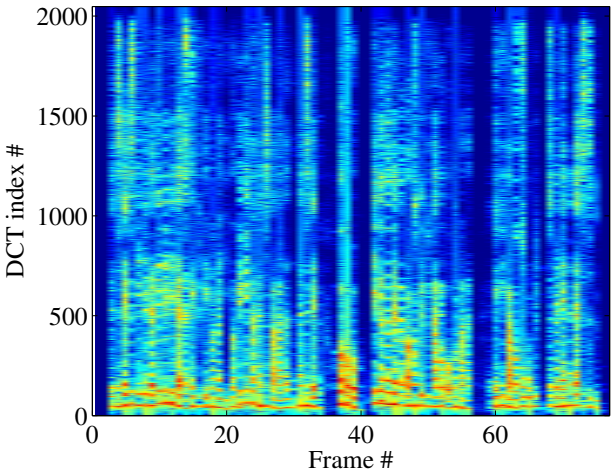

Fig. 2. The log amplitudes of DCT coefficients of the original speech data

motivation to conceal the boundary line, we hereby generated a random number with [4,5] as elements to decide each frame to expand 4 blocks or 5 blocks from the highest coefficients. We could also generated a random number with [8,9] for different capacity for hiding.

B. Solution 2: adaptive hiding with distortion estimation

The purpose of distortion estimation is to select the DCT coefficients with smaller distortion, and the policy to estimate distortion is discussed here. Stego DCT coefficients which are to be selected and expanded to hide payload are suppose to be: $X'_n = 2X_n + B, (1 \le n \le N)$, with a bit of embed data $B, B \in \{0, 1\}$ in it. In order to find the proper DCT coefficients with smaller distortion in an effective way, we use a statistics way to estimate the possible distortion with payload embedded into each expanded DCT coefficients. Suppose each DCT coefficient X_n is a constant, then it is possible to estimate the distortion by calculate the error of mean square between stego data and original data. Theoretically, the distortion of stego data is supposed to be smaller if the hiding locations are adaptively selected by the error of mean square in ascending order. This work is an extension work of our previous work [6], and we explored variable embedding locations and compare the audio quality of each solution.

Let X'_n be the n-th DCT coefficient of stego data, X_n be the n-th DCT coefficient of original data, and E_n be the estimated distortion (error of mean square of difference between X'_n and $X_n, 1 \le n \le N$) after hiding a bit of data B in each DCT coefficient. Since $B \in \{0, 1\}$, the probability of B is $P(B = 0) = 0.5$, and $P(B = 1) = 0.5$. Distortion estimation of each DCT coefficient can be represented by:

$$\begin{aligned} E\left[(X'_n - X_n)^2\right] &= X_n^2 + 2X_nB \sum_{B\in\{0,1\}} P(B) \\ &+ B^2 \sum_{B\in\{0,1\}} P(B^2) \\ &= X_n^2 + X_n + \frac{1}{2} \end{aligned} \quad (6)$$

According to formula (6), it is obvious that that the distortion of each coefficients has an up-and-down trend as the same as X_n value.

Length of each block is $\frac{N}{M}$, and the average distortion of DCT coefficients estimated in each block could be calculated by

$$\overline{E_i} = \frac{\sum_{n=\frac{(i-1)N}{M}+1}^{\frac{iN}{M}} E_n}{\frac{N}{M}}, \quad (1 \le i \le M, 1 \le n \le N) \quad (7)$$

Then the distortion estimated of each block $\overline{E_i}$ is sorted in ascending order to get the sorting result $ind, (1 \le ind \le M)$, indicating blocks number index i. We expand DCT coefficients in blocks where $\overline{E_i}$ is smaller, for example DCT coefficients blocks with $ind \le 8$.

The embedding steps could be summarized as follows:

Step 1 Divide original data into frames with length N and transform data $x_n, (1 \le n \le N)$ from time domain to DCT coefficients X_n using integer DCT IV algorithm;

Step 2 Calculate the estimated distortion E_n of difference between DCT coefficients of stego data and DCT coefficients of original data.

Step 3 Segment N DCT coefficients into M blocks, and calculate the average estimated distortion $\overline{E_i}$ and then sort them in ascending order to have the blocks index number i where $(1 \le ind \le M)$ to generate the location table T.

Step 4 Expand DCT coefficients X_n, where $(i-1)\frac{N}{M}+1 \le n \le i\frac{N}{M}$ and embed data to generate stego DCT coefficients of stego data $X'_n = 2X_n + B$ where $1 \le ind \le 8$.

Step 5 Hide table $T_i, (1 \le i \le M)$ into stego DCT coefficient X'_n where $N - M + 1 \le n \le N$

Step 6 Apply inverse integer DCT IV algorithm to X'_n to get stego data x'_n in time domain.

The extraction and reconstruction steps are summarized as follows:

Step 1 Transform stego data x'_n into DCT coefficients X'_n

Step 2 Calculate $\lfloor X'_n/2 \rfloor$ where $N - M + 1 \le n \le N$ and table information $T_i, (1 \le i \le M)$ equals to $X'_n - 2\lfloor X'_n/2 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function.

Step 3 Read table T_i to get block index i where $T_i = 1$.

Step 4 Calculate $\lfloor X'_n/2 \rfloor$ where $\frac{N(i-1)}{M} + 1 \le n \le i\frac{N}{M}$, then original data after embedded data is removed equals to $\lfloor X'_n/2 \rfloor$ and embedded data equals to $X'_n - 2\lfloor X'_n/2 \rfloor$.

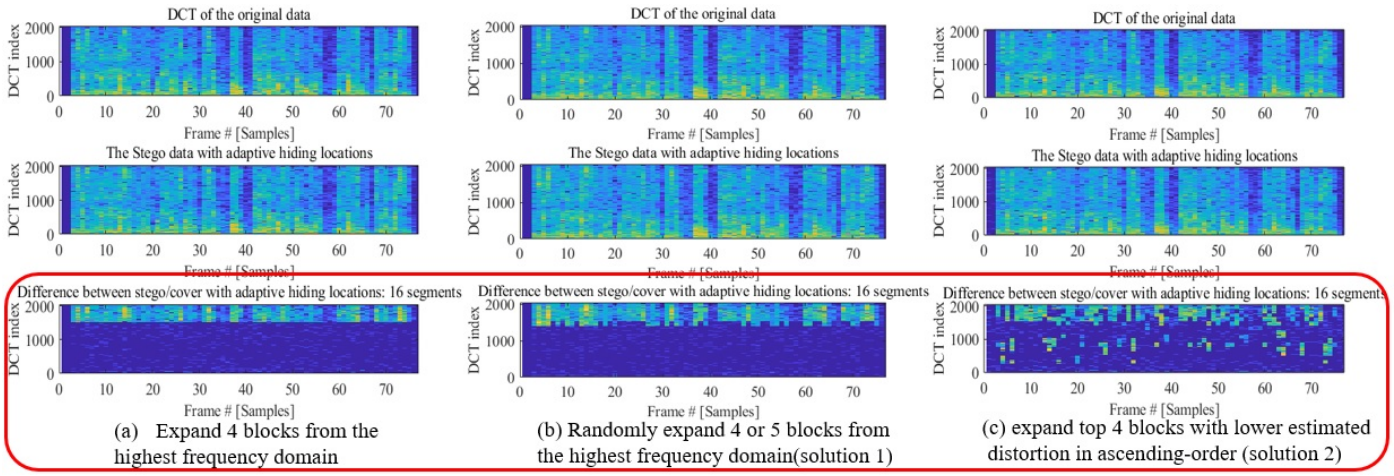

Fig. 3. Comparison on stego data generated of our previous work, randomly expansion (solution 1) and adaptive expansion (solution 2) with different expansion locations

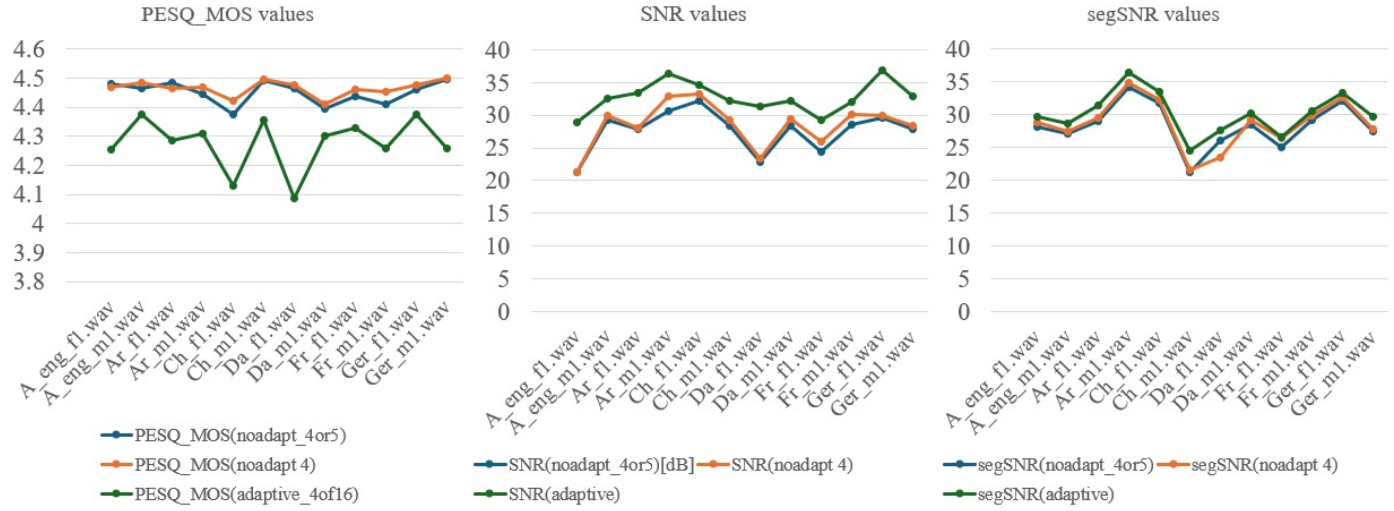

Fig. 4. Comparison on audio quality of our previous work, randomly expansion (solution 1) and adaptive expansion (solution 2) with different expansion locations

Step 5 Reconstruct original DCT coefficients $X_n, 1 \leq n \leq N$

Step 6 Transform X_n into time domain to get original data $x_n, 1 \leq n \leq N$.

IV. Experimental results and evaluation

A. Audio quality evaluation

When evaluating audio quality, we used segmental Signal-to-Noise Ratio (segSNR) and Perceptual Evaluation of Speech Quality (PESQ), which have been extensively used for evaluating the sound quality objectively in conventional works. In order to evaluate the listening quality of speech data, we used Mean Opinion Score Listening Quality Objective (MOS-LQO) score, which is defined by the ITU-T Recommendation P.862.1 [20]. We used AFsp packages version 9.0 to calculate SNR, segSNR scores and PESQ version 1.2 to calculate MOSLQO scores.

We evaluated our method using the dataset from ITU-T Test Signals for Telecommunication Systems–Test Vectors Associated to Rec. ITU-T P.50 Appendix I. We used 12 speech data (12 speakers by 6 languages: American English, Arabic, Mandarin Chinese, Danish, French, Germany, the first track of female and male speakers). The datasets are with a monophonic waveform with 16-kHz sampling and 16-bit quantization. The number of DCT coefficients in each frame is set to be 2048.

As the experimental results, we compared three alternative solutions for expansion: (1) expand 4 blocks from the highest coefficients (our previous work); (2) randomly expand 4 or 5 blocks from the highest coefficients (solution 1); and (3) expand top 4 blocks with lower estimated distortion in ascending-order (solution 2). Figure .3 plots the different expansion locations of (a) our previous work with no adaption, the proposed method of (b) solution 1 randomly 4 or 5 blocks

TABLE I
COMPARISON ON AUDIO QUALITY OF THE CONVENTIONAL WORK, PROPOSED SOLUTION 1, AND PROPOSED SOLUTION 2

Track	PESQ_MOS			SNR (dB)			segSNR (dB)		
	solution 1	conventional [7]	solution 2	solution 1	conventional [7]	solution 2	solution 1	conventional [7]	solution 2
A_eng_f1.wav	4.48	4.468	4.253	21.244	21.363	28.936	28.088	28.806	29.68
A_eng_m1.wav	4.466	4.483	4.374	29.208	29.926	32.568	27.151	27.497	28.733
Ar_f1.wav	4.484	4.466	4.285	27.806	27.991	33.447	28.934	29.465	31.404
Ar_m1.wav	4.447	4.468	4.307	30.699	32.847	36.289	34.201	34.845	36.346
Ch_f1.wav	4.374	4.423	4.131	32.16	33.173	34.622	31.792	32.305	33.461
Ch_m1.wav	4.494	4.495	4.354	28.353	29.272	32.171	21.153	21.639	24.43
Da_f1.wav	4.464	4.476	4.085	22.925	23.424	31.35	26.056	23.424	27.595
Da_m1.wav	4.393	4.412	4.302	28.423	29.362	32.15	28.516	29.226	30.283
Fr_f1.wav	4.438	4.462	4.327	24.476	26.005	29.222	24.957	26.36	26.572
Fr_m1.wav	4.411	4.454	4.26	28.6	30.094	32.005	29.11	29.869	30.474
Ger_f1.wav	4.461	4.475	4.377	29.532	29.925	36.903	32.048	32.581	33.402
Ger_m1.wav	4.496	4.499	4.258	27.886	28.39	32.963	27.36	27.866	29.686

expansion and (c) solution 2 adaptive expansion. By comparing Figure 3(a) and (b), the boundary line might be concealed, and Figure 3(c) has a complex key for expansion, and the location map for each frame is different, which enhance the complexity for global steganalysis attacks.

The audio quality of MOS values, SNR[dB] values, and segSNR values are plotted in Fig. 4. According to the results, the previous work and the proposed randomly expansion with 4 or 5 blocks (solutions) has considerable MOS values, and the adaptive expansion (solution 2) has generally lower MOSLQO values. However, all the data are better than 4.0, with the minimum MOSLQO value 4.085, which means all of the proposed method results an imperceptible distortion. For SNR values, the adaptive expansion (solution 2) has the best values, with is better than the conventional work with no adaption, and the randomly 4 or 5 blocks expansion (solution 1) have considerable SNR values with the conventional work with no adaption. For segSNR values, results of solution 2 is better than solution 1, and both of the proposed works have better segSNR values than the previous work. Since segSNR values indicate the differential in time domain, the difference between the stego data and the original data of the proposed methods are supposed to be lower. Table I lists up the audio quality results in details.

B. Discussion on borderline for expansion

The original purpose of the proposed method is to improve the vulnerability of detection on embedding locations. In our previous work [7], due to the concentration of hiding locations, the border line around 4 kHz is conspicuous as shown in Figure 1. To make the border line inconspicuous, we proposed solution 1 with random additional expansion blocks, and solution 2 with an adaptive expansion algorithm. The result to conceal the borderline of embedding locations is as shown in Figure 5. It is the result of an adaptive hiding with distortion sorting in ascending-order when block size $M = 64$, where the borderline is concealed as Figure 5 shows.

Though concealing the borderline is proposed as a countermeasure against spectral statistical analysis to hide the suspected watermarked data, it is still challenging to consider potential attacks techniques by global steganalysis. Global

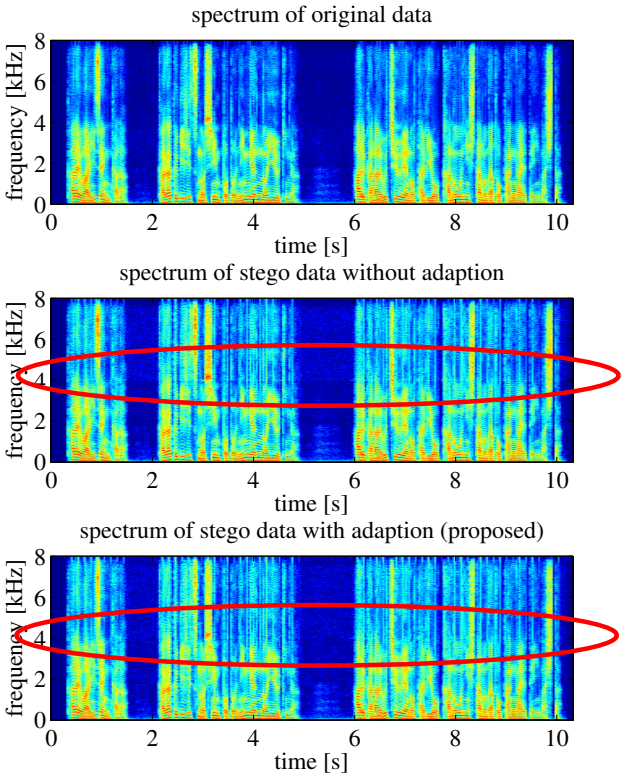

Fig. 5. Comparison of spectrum between stego data generated by method without adaption and with adaption (64 segments): Ja_m5.wav

steganalysis employs various techniques to detect the presence of embedded payload by analyzing the statistical properties and structural characteristics of audio data without cover data. The frequency components, temporal patterns, and perceptual attributes can be indicative of watermarked payload by analyzing the stego data. Analyzing the robustness and resistance and evaluate the effectiveness of the proposed methods against global steganalysis is listed as a future challenge of this work.

V. Conclusions

To conceal the borderline of expansion for reserving embedding locations, we proposed solutions to explore the appropriate expansion locations. A random addition of expansion blocks solution, and an adaptive expansion by sorting the expected distortion are proposed. We evaluated the audio quality and the effectiveness to conceal the borderline, and the results show that comparably good quality of stego data as our previous work is achieved, besides, the hiding locations are inconspicuous to avoid malicious attack. Even the vulnerability of borderline is fixed, we still have the challenging to conceal the expansion locations against statistical spectral analysis by global steganalysis by objective evaluation as a future work.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP18K18052.

This work was supported by "Shimane University Grants for Joint Research Project led by Female Researchers" under the MEXT "Initiative for Realizing Diversity in the Research Environment (Collaboration Type).

References

[1] X. Zhang, et al.: Robust Reversible Audio Watermarking Scheme for Telemedicine and Privacy Protection. Computers, Materials & Continua 71.2 2022.

[2] M. Charfeddine, et al.: Audio watermarking for security and non-security applications, IEEE Access 10, pp: 12654-12677, 2022

[3] D.Q.Yan, and R.D.Wang.: Reversible Data Hiding for Audio Based on Prediction Error Expansion, Proc. of Intelligent Information Hiding and Multimedia Signal Processing, 249–252, 2008

[4] A. Nishimura, Reversible Audio Data Hiding Using Linear Prediction and Error Expansion, Proc. of Intelligent Information Hiding and Multimedia Signal Processing, 318-321, 2011

[5] Unoki, M.: Construction of auditory media signal processing infrastructure to prevent media clone attacks. Impact, vol. 2020(2), 21-23, 2020.

[6] X. Huang, et al.: Reversible Audio Information Hiding Based on Integer DCT Coefficients with Adaptive Hiding Locations. In: Shi, Y., Kim, HJ., Pérez-González, F. (eds) Digital-Forensics and Watermarking. IWDW 2013. Lecture Notes in Computer Science(), vol 8389. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-43886-2_27, 2014

[7] X. Huang, A. Ito.: Imperceptible and Reversible Acoustic Watermarking Based on Modified Integer Discrete Cosine Transform Coefficient Expansion. Applied Sciences. 2024; 14(7):2757. https://doi.org/10.3390/app14072757, 2024

[8] A. Anand, et al.: An improved DWT-SVD domain watermarking for medical information security. Computer Communications, vol.152, pp: 72-80, 2020

[9] Y. Yu, J. Gao, X. Mu, et al.: Adaptive LSB quantum image watermarking algorithm based on Haar wavelet transforms, Quantum Inf Process vol.22(180) , https://doi.org/10.1007/s11128-023-03926-1, 2023

[10] P. Garg, P.:A robust technique for biometric image authentication using invisible watermarking, Multimedia Tools and Applications, vol. 82(2), pp: 2237-2253, 2023

[11] Hernández-Joaquín, et al.: A secure DWT-based dual watermarking scheme for image authentication and copyright protection, Multimedia Tools and Applications, vol.82(27), pp: 42739-42761, 2023

[12] M. Roy, et al.: A perceptual hash based blind-watermarking scheme for image authentication, Expert Systems with Applications, vol. 227, https://doi.org/10.1016/j.eswa.2023.120237, 2023

[13] Sharma, Sunpreet, et al.: A review of image watermarking for identity protection and verification, Multimedia Tools and Applications, vol.83(11), pp.31829-31891, 2024

[14] H. R. Chennamma, et al.: A comprehensive survey on image authentication for tamper detection with localization, Multimedia Tools and Applications, vol.82(2), pp: 1873-1904, 2023

[15] D. Singh, et al.: An efficient self-embedding fragile watermarking scheme for image authentication with two chances for recovery capability, Multimedia Tools and Applications, vol. 82(1), pp: 1045-1066, 2023

[16] L. De, et al.:A reversible watermarking for image content authentication based on wavelet transform, Signal, Image and Video Processing, pp. 1-11, 2024

[17] G. Gao, M. Wang and B. Wu.:Efficient Robust Reversible Watermarking Based on ZMs and Integer Wavelet Transform, IEEE Transactions on Industrial Informatics, vol. 20(3), pp. 4115-4123, doi: 10.1109/TII.2023.3321101, 2024

[18] C. Zhan, et al.:Reversible Image Fragile Watermarking with Dual Tampering Detection, Electronics, vol.13(10), 1884. https://doi.org/10.3390/electronics13101884, 2024

[19] L. Tanwar, Lavi, et al.: Hybrid reversible watermarking algorithm using histogram shifting and pairwise prediction error expansion, Multimedia Tools and Applications, vol.83(8), pp: 22075-22097, 2024

[20] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, ITU-T Recommendation P.862.1, International Telecommunication Union, 2001