可逆なスペクトル音声電子透かしの可変的な埋め込み位置に関する検討

黄 緒平[1,2,a)]　伊藤 彰則[1,b)]

概要：音声合成や機械学習の高速な発展により，音声コンテンツへの精巧な改ざんができるようになった．証拠性の高い音声信号データの真正性を保証するため，電子透かしは有効な手段とされているが，音質の劣化を抑止する必要がある．人間の聴覚システムに基づいて，知覚しにくい高い周波数領域へ改ざん検出用の署名データを埋め込む可逆電子透かし手法が提案されているが，スペクトル解析によって埋め込み箇所の境界線が判別できてしまう．本研究では埋め込み箇所を秘匿するため，様々な可変的な埋め込み位置を検討し，音質評価を行う．

キーワード：音声電子透かし，可逆電子透かし，改ざん検出，埋め込み位置

Exploring Variable Embedding Locations of Reversible Acoustic Watermarking

Xuping Huang[1,2,a)]　Akinori Ito[1,b)]

Abstract: Rapid advancements in machine learning and speech synthesis have made manipulating and tampering audio content feasible. 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. This research explored variable embedding locations while preserving good sound quality.

Keywords: Acoustic watermarking, Reversible watermarking, Tampering detection, Embedding locations

1. はじめに

近年のデジタル技術の発展に伴い，デジタルコンテンツの改ざん防止がますます重要になってきている．改ざんの問題を解決する方法として，音声 [1-7]、画像、動画データ [8-19] などのコンテンツデータ自体に整合性チェック用のペイロード情報を挿入する電子透かしに基づく方法が提案されてきている．電子透かしに基づくアプローチでは，埋め込みのペイロード容量が大きいことと，埋め込みを行ったときのコンテンツの劣化が小さいことが求められる．さらに，医療や法律関係の記録などでは，埋め込みを行う前のデータを利用できなければならず，透かしの埋め込みに伴う品質の劣化が供用されない場合が多い [16]．これらの理由から、このような目的には可逆電子透かし (Reversible watermarking) が利用される．

歪みの低減を実現するためには，波形領域と周波数領域のどちらに埋め込みを行うかが重要となる．オーディオ分野では、時間領域線形予測符号化 [3,4] や，蝸牛遅延特性に基づく埋め込み方式 [5] が提案されている．一方，画像の

[1] 東北大学工学研究科
Graduate School of Engineering, Tohoku University
[2] 島根大学総合理工学部
Interdisciplinary Faculty of Science and Engineering, Shimane University
[a)] huang@cis.shimane-u.ac.jp
[b)] aito.spcom@tohoku.ac.jp

分野では，可逆電子透かしにおいて整数離散コサイン変換 (intDCT) が用いられることが多く，高容量と低歪みを達成している [14-18]．これらを参考に，我々は intDCT ベースの可逆的な音声電子透かし手法を開発してきた．

画像と音声は統計的特徴と知覚的性質が異なるため，intDCT による可逆音声電子透かしの設計にあたっては，ペイロードを DCT 係数に埋め込む方法を音声に合わせて調整する必要がある。画像の場合，ペイロードをより高周波の DCT 係数に埋め込むと復元力が低下し，より低周波の DCT 係数に統合すると不可視性が低下する．そのため、DCT 係数の中間領域が埋め込みに利用される．一方，音声においては，人間は高周波領域の違いにより鈍感である．そのため，我々の以前の研究 [7] では，ペイロードデータを高周波帯域の DCT 係数に埋め込んだ．

しかし，以前の研究 [7] では，高周波 DCT 領域の半分に埋め込み領域が集中しているため、DCT 係数の拡大によって生成された境界線がスペクトル上に表示され，攻撃者が透かしの方法を推測するためのヒントになる可能性がある．そのため，敵対的な攻撃を回避するには、埋め込み領域が目立たないようにする必要がある．これを実現するには，ステゴデータの品質を維持するだけでなく、埋め込み領域自体をわかりにくくするアルゴリズムが必要になる．Li らは，ステゴデータの品質と複雑な隠蔽場所を改善するために，透かしを画像に適応的に埋め込むことを提案した [16]．PSNR による評価によれば，適応的な埋め込み手法は適応なしの手法よりも優れていた．

本研究では、埋め込み位置のランダム化を含むさまざまな埋め込み位置と、歪み推定による DCT 係数へのペイロードの適応埋め込みについて調査し，良好な音質を維持しながら埋め込み位置を隠ぺいすることを目指す．

本稿では，第 2 節でこれまでの研究の概要を述べ，第 3 節で提案法を紹介する．第 4 節でオーディオ品質に関する実験的評価について述べ，第 5 節で結論と今後の課題について述べる．

2. intDCT に基づく可逆音声電子透かしの概要

我々の以前の研究 [7] では，改ざん検出のために，修正された整数 DCT 係数拡張に基づく可逆電子透かし法を提案した．整数 DCT タイプ IV の一般的な傾向として，高い周波数帯域ほど，係数の大きさが小さくなる。したがって提案法では，知覚的なステゴデータの歪みを制御するために，高周波数領域で DCT 係数を拡張して，その領域をペイロードとした．改ざんを正確に検出するために、まず元の音声データを固定サンプル長 N (通常は 512，1024 など) でフレームに分割する．フレーム内のオーディオ信号を x_n $(1 \le n \le N)$ とする．フレーム内の時間領域サンプルは，整数 DCT を使用して DCT（周波数）領域に変換される．

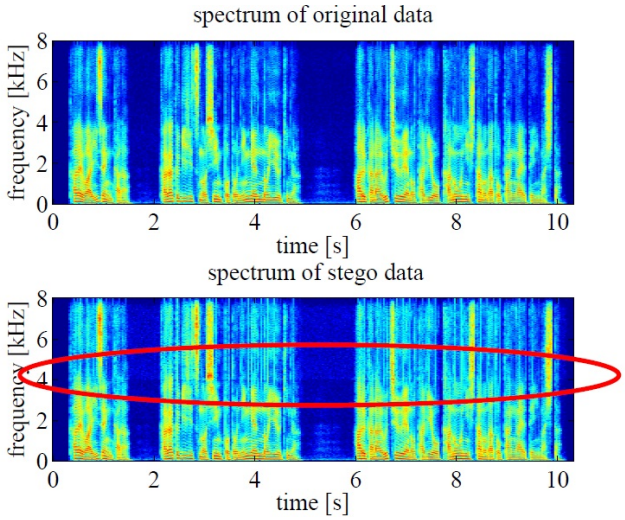

図 1 元のデータとステゴデータ DCT 係数の対数．拡張と埋め込みによって生じた境界を赤で囲っている．Ja_m5.wav (N = 2048)

まず，

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

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

はそれぞれ N 点フレームにおける時間領域信号とその DCT 係数である. 連続的なケースでは、時間領域信号 h から DCT-IV 行列によって DCT 係数 H を得ることができる．

$$\mathbf{H} = C_N^{IV}\mathbf{h} \tag{3}$$

ここで DCT-IV 行列 C_N^{IV} の i 行 t 列の要素は次のようにあらわされる．

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

提案法では，DCT 係数を拡大することで透かしを埋め込む．DCT 係数を X_n $(1 \le n \le N)$ とする．1 ビットのメッセージ $B \in \{0,1\}$ を X_n に埋め込むとき，次のように係数を拡大する．

$$X'_n = 2X_n + B. \tag{5}$$

埋め込み後，逆整数 DCT で係数を時間領域に変換する．整数 DCT の性質により x_n と X_n の可逆性が保証される．我々の研究 [7] においては、人間は高周波係数の違いにそれほど敏感ではないため、高周波数帯域の DCT 係数を埋め込みに使用した．ペイロードが DCT 係数の半分である場合，すなわち $(N/2+1 \le n \le N)$ となる X_n に埋め込みを行った場合，サンプルあたり 0.5 ビットの容量が達成される．ただし，拡張用の DCT 係数は高周波領域に集中しているため，攻撃者がスペクトルを観測することによって

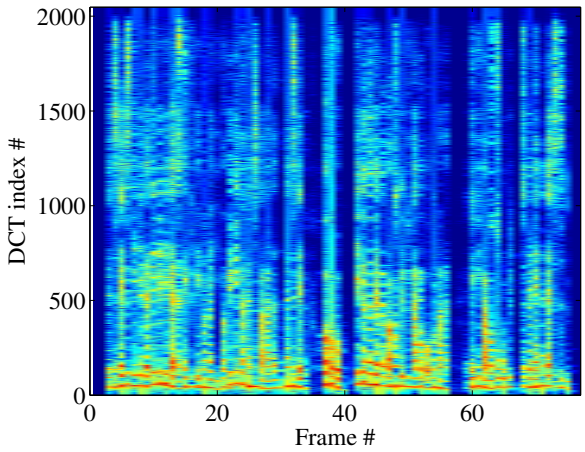

図 2 元の音声データの DCT 係数の振幅の対数

埋め込み位置が知られてしまう可能性がある．

埋め込まれた信号が目視で確認できるかどうかを確かめるために，透かしを埋め込んだ信号のスペクトログラムを観察した．図 1 は、Ja_m5.wav の元の音とステゴ信号のスペクトログラムである．横軸は時間，縦軸は DCT 係数に対応する周波数を表す．上半分の周波数の係数が埋め込みのために拡張されている場合，ステゴデータと元のデータを比較すると、拡張と埋め込みの境界が見える可能性があることがわかる．

3. 提案法：可変的な位置への透かしの埋め込み

境界が見えてしまう問題を解決し，ステガナリシスに対する頑健性を高めるために，DCT 係数拡張に基づく電子透かしの埋め込み位置について検討した．以前の研究 [6] では、拡張のランダムキーが歪みを引き起こし，元のデータに雑音が発生する可能性があることを示した．したがって、透かしを埋め込む適切な場所を選択し、拡張後の品質を改善する必要がある．

3.1 方法 1: 埋め込み位置のランダムな変更

目的は境界を見えにくくことなので，係数を拡張する位置をランダムに追加することが有効となる可能性がある．図 2 は，音声データ (Da_f2.wav、ドイツ語の女性音声データ) の DCT 係数の対数振幅の例である．Y 軸はフレーム長 $N = 2048$ の DCT 係数である．従来法では，透かしを埋め込むために， 1025 番目から 2048 番目の係数を拡張していた。今回は，境界線を隠すために，追加の係数を拡張する方法を提案する．

データを抽出して元のデータを再構築するには，埋め込み位置を記録するためのロケーションマップが必要である．まず，N 個の DCT 係数をインデックス i $(1 \le i \le M)$ を持つ M 個のブロック $(M < N)$ に分割する．M が大きいほど、ロケーションマップを埋め込むために必要な容量が大きくる．N 個の係数を M 個のブロックに分割すると、ロケーションマップのサイズを N から M に短縮できる．

予備的な検討として，$N = 2048, M = 16$ として，高い周波数の 4 ブロックまたは 5 ブロックに情報を埋め込む方法を試した．埋め込みの容量を変えるために，8 ブロックまたは 9 ブロックに埋め込むことも可能である．

3.2 方法 2: 歪み推定による適応的な埋め込み

次に，歪の推定に基づいて適応的に埋め込み位置を決定する方法を検討した．まず，歪みを推定する考え方について説明する．透かしを埋め込むために選択および拡張される DCT 係数は $X'_n = 2X_n + B, (1 \le n \le N)$ で、埋め込みデータ $B \in \{0, 1\}$ が含まれるものとする．歪みがより小さい適切な DCT 係数を効果的に見つけるために，統計的手法を使用して，各拡張 DCT 係数に埋め込まれた透かしによって発生する可能性のある歪みを推定する．各 DCT 係数 X_n が定数であると仮定すると，ステゴデータと元のデータの平均二乗誤差を計算することで歪みを推定できる．埋め込みによる信号の劣化が独立に発生すると仮定すれば，平均二乗誤差によって昇順に埋め込み位置を選択することで，ステゴデータの歪みを小さくできるはずである．この方法は，以前の研究 [6] の拡張であり，さまざまな埋め込み場所を調査し，各方法の品質を比較した．

DCT 係数に透かし B を埋め込んだ後の推定歪みを求める．B が一様分布すると仮定すれば，歪み E_n は

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

となる．式 (6) によれば，歪みの値は係数自体の値の大きさによって上下することが明らかである．

各ブロックに含まれる係数の個数は $S = \frac{N}{M}$ であるから，各ブロックの平均的な歪みは次のように計算できる．

$$\overline{E_i} = \frac{1}{S} \sum_{n=(i-1)S+1}^{iS} E_n \tag{7}$$

次に、各ブロックの推定歪み $\overline{E_i}$ を昇順に並べ替えて、ブロック番号インデックス i を示す並べ替え結果 $ind, (1 \le ind \le M)$ を取得する．そして， $\overline{E_i}$ が小さいブロック，たとえば $ind \le 8$ の DCT 係数ブロックで DCT 係数を展開する．埋め込み手順は次のようになる．

Step 1 元のデータを長さ N のフレームに分割し，1 フレーム内の時間領域データ x_n, $(1 \le n \le N)$ を整数 DCT-IV によって DCT 係数 X_n に変換する．

Step 2 元のデータから推定歪み E_n を算出する．

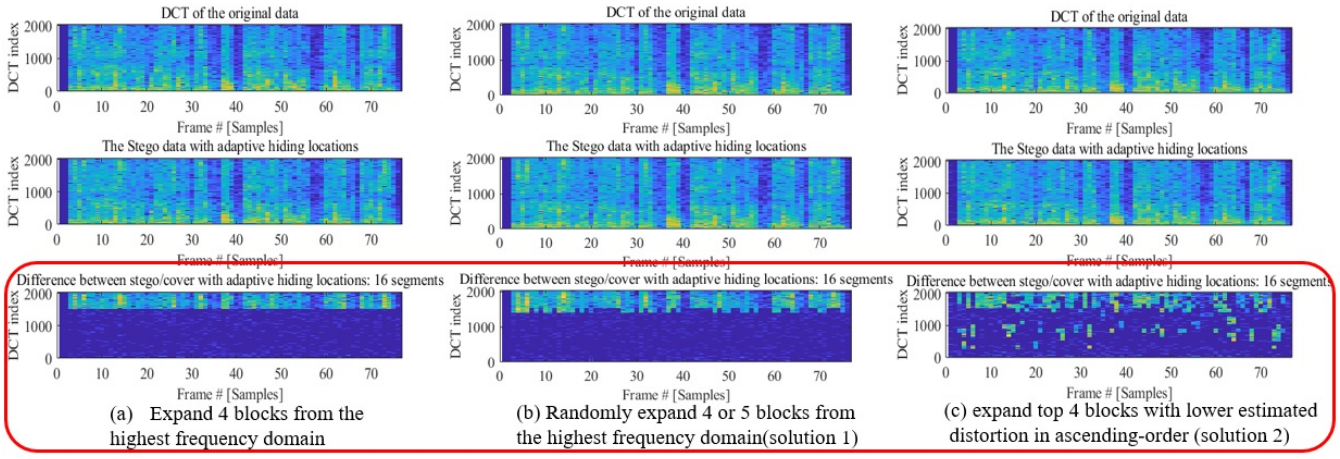

図 3 ステゴデータの比較 (DCT 係数の例) (a) 従来法，(b) ランダム拡張 (方法 1)，適応的な拡張 (方法 2)

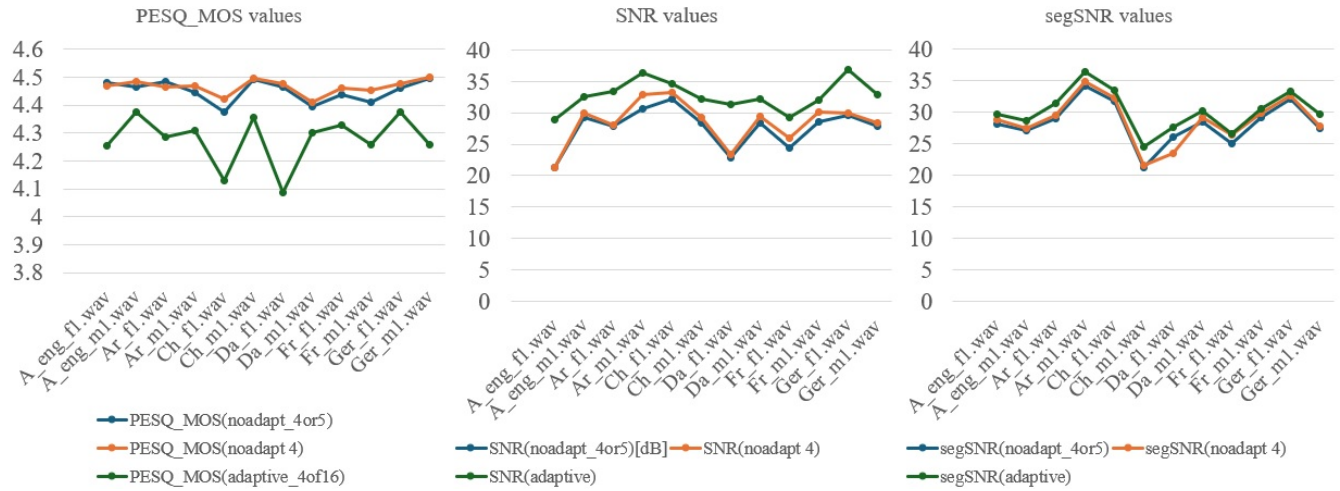

図 4 ステゴデータの音質比較 (a) 従来法，(b) ランダム拡張 (方法 1)，適応的な拡張 (方法 2)

Step 3 N 個の DCT 係数を M ブロックに分割し，平均推定歪み $\overline{E_i}$ を求める．これを昇順にソートし，ind を求める．それを使って埋め込むブロックを決定し，埋め込みテーブル T を生成する．

Step 4 ind の小さい方から一定数のブロックの DCT 係数を拡張し，透かしを埋め込む．

Step 5 $N-M+1 \leq n \leq N$ となる DCT 係数 X'_n に T を埋め込む．

Step 6 逆整数 $DCT-IV$ によって X'_n を時間領域信号 x'_n に変換する．

同様に，透かしの抽出とデータの復元は次のように行うことができる．

Step 1 ステゴデータ x'_n を DCT 係数 X'_n に変換する．

Step 2 $N-M+1 \leq n \leq N$ について $T_i \leftarrow X'_n \bmod 2$ として埋め込みテーブル T を復元する．

Step 3 $T_i = 1$ となるブロックに透かしが埋め込まれているので，それらのブロックに対して以下の抽出を行う．

Step 4 透かしが埋め込まれたブロックに対して，

$B_n \leftarrow X'_n \bmod 2$

$X_n \leftarrow \lfloor X'_n/2 \rfloor$

によって透かしと元のデータを復元する．

Step 5 X_n を時間領域信号 x_n に変換する．

4. 実験結果と評価

4.1 音質評価

音質評価には、従来の研究で音質を客観的に評価するために広く使用されているセグメンタル SNR (segSNR) と，知覚的音声品質評価手法である PESQ を使用した．PESQ による評価値として， ITU-T P.862.1[20] で定義されている MOS-LQO スコアを利用した．SNR と segSNR スコアの計算には AFsp パッケージ version 9.0 を使用し，MOS-LQO スコアの計算には PESQ version 1.2 を使用した．

テストデータとして，ITU-T P.50 Appendix I のデータセットを使用した．12 の音声データ (6 つの言語による 12 人の話者: アメリカ英語、アラビア語、中国語 (北京語)、デンマーク語、フランス語、ドイツ語、女性と男性の話者の最初のトラック) を使用した．データセットは、16 kHz

表 1 従来法、提案法 1、提案法 2 の音質比較

Track	PESQ_MOS			SNR (dB)			segSNR (dB)		
	方法 1	従来法	方法 2	方法 1	従来法	方法 2	方法 1	従来法	方法 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

サンプリング・16 ビット量子化のモノラル音声である．各フレームの DCT 係数の数は $N = 2048$ とした．ブロック数は $M = 16$ とした．

実験として，次の 3 つの方法を比較した．

(a) 最高係数から 4 ブロック拡張（従来法）

(b) 最高係数からランダムに 4 または 5 ブロック拡張（方法 1）

(c) 推定歪みが低い上位 4 ブロックを昇順で拡張（方法 2）

それぞれの方法により透かしを埋め込んだステゴデータの例を図 3 に示す．上段はオリジナルデータの DCT 係数の振幅の対数値，中段はステゴデータ，下段はオリジナルデータとステゴデータの差分である．図 3(a) と (b) を比較すると、(b) では境界線が直線でないので，より協会が分かりにくくなっている可能性がある．図 3(c) では各フレームの位置マップが異なり、ステガナリシス攻撃が難しくなっていると考えられる．

各音声サンプルの MOS-LQO，SNR，および segSNR 値の値を図 4 に示す．これらの結果によると、従来法（オレンジ）と提案法 1（青）はほぼ同じ MOS-LQO 値を持ち，提案法 2（緑）の MOS-LQO は低かった．ただし、すべての値は 4.0 より良く，最小の MOS-LQO 値は 4.085 であるため，提案法のすべてが「歪みが知覚できない」品質であった． SNR 値については提案法 2 が最良であり，従来法よりも高い値を示した．提案法 1 は、従来法とほぼ同じ SNR 値であった．segSNR 値については提案法 2 の結果が提案法 1 よりも優れており， 2 つの提案法はいずれも従来用よりも高い segSNR を示した．表 1 に音声品質の結果の詳細を示す．

4.2 拡張の境界線に関する考察

提案手法の本来の目的は，埋め込み位置の検出の脆弱性を改善することであった．以前の研究 [7] では，埋め込み位置の集中により，図 1 に示すように 4kHz 付近の境界線が目立っていた．境界線を目立たなくするために，ランダムに追加拡張ブロックを使用する方法 1 と，適応拡張アルゴリズムを使用する方法 2 を提案した．提案法による埋め込み位置の境界線付近の例を図に示す．これは，ブロックサイズ $M = 64$ の場合に、歪みを昇順にソートして適応的に埋め込んだ結果（方法 2）であり，図 5 のように境界線が見えなくなっていることが分かる．

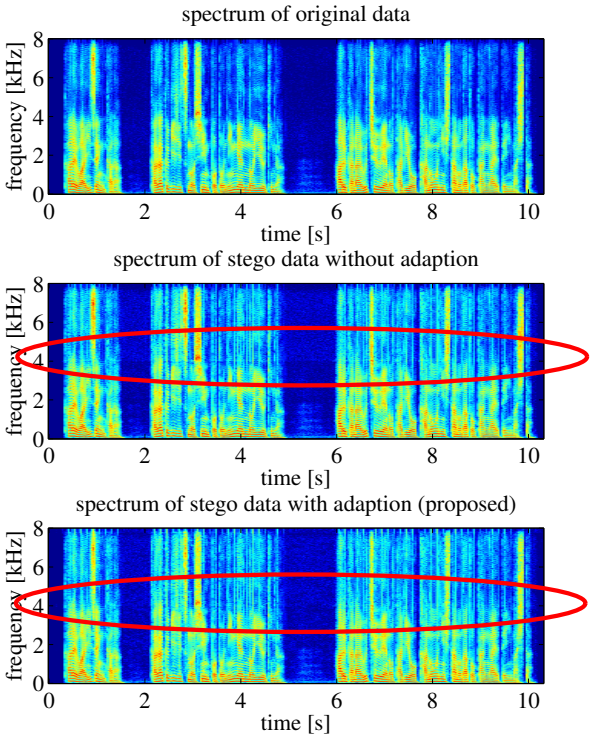

図 5 適応なしと適応ありの方法で生成されたステゴデータのスペクトルの比較（M=64）：Ja_m5.wav

5. むすび

埋め込み位置を確保するための拡張の境界線を隠すために，適切な拡張位置を探すための解決策を提案した．拡張ブロックのランダム追加（方法１）と，歪みを最小化することによる適応拡張（方法２）の２つの方法を提案した．これらの手法による埋め込み結果をを評価したところ，従来法と同程度の音声品質が達成され，さらに悪意のある攻撃を避けるために隠蔽位置が目立たないことが示された．今後の課題として，グローバルステガナリシスによる統計的スペクトル解析に対して拡張位置を隠すという課題が残っている．

謝辞　この研究の一部は，科研費若手研究 18K18052 の支援を受けた．

参考文献

[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