A Robust Recoverable Algorithm Used for Digital Speech Forensics Based on DCT

Zhenghui Liu[1], Yanli Li[1], Fang Sun[1], Junjie He[1], Chuanda Qi[1],
and Da Luo[2](✉)

[1] College of Computer and Information Technology,
Xinyang Normal University, Xinyang 464000, China
[2] School of Computer Science and Network Security,
Dongguan University of Technology, Dongguan 523000, China
luoda@dgut.edu.cn

Abstract. Recoverable speech forensics algorithm not only can locate the attacked frames, but can reconstruct the attacked signals. Meanwhile, the method can provide useful information for the prediction of attacker and attacker's intent. We proposed a robust recoverable algorithm used for digital speech forensics in this paper. We analyze and conclude that large amplitude DCT coefficients play a more significant role for speech reconstruction. Inspired by this, we regard the large amplitude coefficients as compressed signal, used for the reconstruction of attacked frames. For embedding, we scramble samples of each frame, and embed frame number and compressed signal into less amplitude DCT coefficients of scrambled signal by substitution. Frame number is used for tamper location of watermarked speech, and compressed signal is used for the reconstruction of attacked signals. Experimental results demonstrate that the algorithm is inaudible and robustness to signal processing operations, has ability of tamper recovery and improves the security of watermarking system.

Keywords: Watermarking · Speech forensics · Speech compression

1 Introduction

So far, many audio watermarking methods have been developed [1], including time domain methods and transform domain methods. The time domain methods include time aligned methods [2] and echo-based methods [3]. The transform domain methods include spread spectrum methods [4], quantization index modulation methods [5], patchwork methods [6, 7]. Transform domain methods are generally more robust because they can take advantage of signal characteristics and auditory properties [8]. So, most robust audio watermarking schemes for copyright protection are based on transform domain [9, 10].

For digital speech, as a carrier to transmit information, the semantic should be intact and authentic. If audiences regard the semantic of attacked speech as authentic and act according to the wrong instructions, it can cause serious consequences. So, apart from copyright protection, speech forensics is indispensable, which is one of the applications of digital audio watermarking [11].

X. Sun et al. (Eds.): ICCCS 2018, LNCS 11068, pp. 300–311, 2018.
https://doi.org/10.1007/978-3-030-00021-9_28

There have been some schemes for speech forensics [12, 13]. In [12], as to compressed speech by using codebook-excite linear prediction, authors proposed the scheme used for compressed speech forensics, and embedded watermark into the least significant bits [14]. Least significant bits are fragile, and will be changed after signal processing operations. For the scheme, it will regard common signal processing operation as hostile attack. So, the scheme is unsuitable for the speech subjected to signal processing. In [13], authors proposed audio amplitude cooccurrence vector features used for verifying whether audio signal is subjected to post process. The features can exploit cooccurrence patterns of audio signals, and have the ability of distinguish between the original audio and the postprocessed audio with an average accuracy of above 95%.

The semantic of hostile attacked speech is different to the original one. For the speech, expressing emergency tasks and important directives, the greatest wish of audience maybe to acquire the semantic of the original signal. In this case, reconstructing the attacked speech is the users to pray. There have been a considerable amount of recovery schemes for digital images [15–17]. While, there are comparatively few recovery schemes for digital speech [18].

Considering the background and motivation above, we proposed a robust recoverable algorithm used for digital speech forensics. By the statistical distribution rule of Discrete Cosine Transform (DCT) coefficients, we analyze and conclude that large amplitude coefficients play a more significant role for speech reconstruction. Inspired by this, we give the digital speech signal compression method, and get the conclusion that original speech can be reconstructed by using the compressed signal, under the condition of keep semantic. We scramble samples of each frame, and substitute the less amplitude DCT coefficients to embed frame number and compressed signal into host speech. Use frame number to locate the attacked frame. And then extract compressed signal and reconstruct the attacked frame to perform tamper recovery. Theoretical analysis and experimental results demonstrate that the proposed scheme is inaudible and robustness to signal processing operations, and can recover the attacked signals.

2 Speech Compression Based on DCT

2.1 Distribution of Discrete Cosine Transform Coefficients

Select one speech signal randomly, and we perform DCT on the signal. Figure 1 gives the DCT coefficients. It can be seen that the amplitudes of low-band DCT coefficients is great than high-band coefficient, and the energy is mainly distributed in the low frequency domain (results shown in Fig. 1 also validate the conclusion). It means that low-band coefficients, or large amplitude coefficients play a more significant role than small amplitude coefficients for speech reconstruction.

We select 500 speech signals recorded in four different environments, including quiet room, seminar, park and railway station. The length of each signal is about 10 s, and sampled at 8 kHz. In Fig. 2, we show the statistical result of the number of DCT coefficients taking different values, for the 500 speech signals. And the horizontal coordinate is the DCT coefficients value, and the vertical coordinate represents the

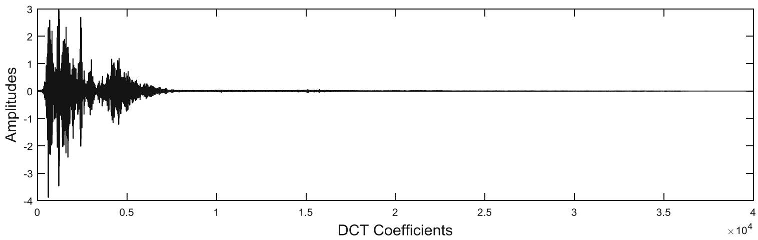

Fig. 1. DCT coefficients of one speech signal

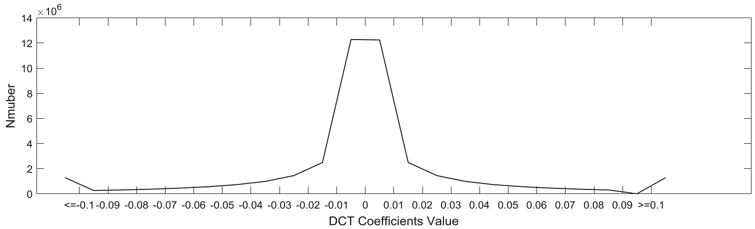

Fig. 2. The number of DCT coefficients for different value

number of DCT coefficients. Figure 2 indicates that most of the DCT coefficients are close to 0, except for the coefficients greater than 0.1 and less than −0.1.

Inspired by the results shown in Figs. 1 and 2, we analyze and obtain the conclusion. (1) The amplitudes of low-band DCT coefficients are great than high-band coefficients, and the energy is mainly distributed in the low frequency domain. And for speech reconstruction by inverse DCT, large amplitude coefficients play a more significant role than small amplitude coefficients. (2) Even if we set the small coefficients (greater than −0.1 and less than 0.1) to 0, we can obtain the signal having the same semantic to original one by inverse DCT.

From the distribution of DCT coefficients shown in Fig. 2, we can see that the number of larger amplitude coefficients is about 4% of the total coefficients. Simultaneously, based on the conclusion obtained by Fig. 1 that large amplitude coefficients are almost all the low-band DCT coefficients, we record the 4% low-band DCT coefficients as the compressed signal in this paper.

2.2 Speech Compression and Reconstruction

2.2.1 Speech Compression

We denote $A = \{a(l), 1 \leq l \leq L\}$ as the L length speech signal, and $a(l)$ represents the l-th sample.

(1) We cut A into P frames. Denote N as the length of each frame and A_i as the i-th frame.

(2) Perform DCT on A_i, for the large amplitudes are almost all low-band DCT coefficients, we select the 4% low-band DCT coefficients and denote as $C_i = \{c_l, 1 \leq l \leq N/25\}$.

(3) Narrowing the amplitudes of $C_i = \{c_l, 1 \leq l \leq N/25\}$, by using Eq. (1).

$$\bar{c}_l = c_l / MC_i \tag{1}$$

where $MC_i = \max\{|c_l|, 1 \leq l \leq N/25\}$, and $|c_l|$ represents the amplitude of c_l. We denote the $\bar{C}_i = \{\bar{c}_l, 1 \leq l \leq N/25\}$ as the compressed signal.

2.2.2 Speech Reconstruction

We can reconstruct A_i under the condition of keeping semantic by using the compressed signal $\bar{C}_i$.

(1) Except for the 4% low-band DCT coefficients, we set other coefficients to 0, to construct N length DCT coefficients.

(2) Perform inverse DCT on the N length constructed DCT coefficients. Then we normalize the signal, and can obtain the reconstructed signal.

We take the speech signal shown in Fig. 3 as an example. Cut the signal into 3 frames, each frame is 1000 length. Figure 4 shows DCT coefficients of the 3 frames. By using the above compression and reconstruction method, we can get the compressed signal (large amplitudes coefficients) of the 3 frames, shown in Fig. 5. Based on the compressed signal, we construct N length DCT coefficients of each frame, and perform inverse DCT on the coefficients to reconstruct the signal approximatively, shown in Fig. 6.

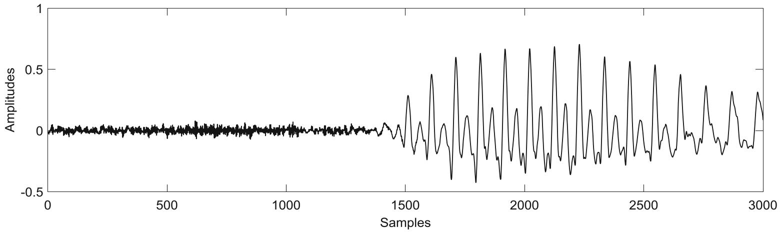

Fig. 3. 3000 length speech signal

For speech signals, they have mute parts and non-silence parts generally. The non-silence parts play an important role for semantics expression, while the mute parts have no semantics. By comparing with the waveform of speech signal (shown in Fig. 3) and the reconstructed one (shown in Fig. 6), we can conclude that the reconstruction method can recover the non-silence parts of speech signal approximately, and the reconstructed mute parts are obviously different to the original ones. For mute parts have no semantics, the reconstructed mute parts can be considered as acceptable for semantics expression.

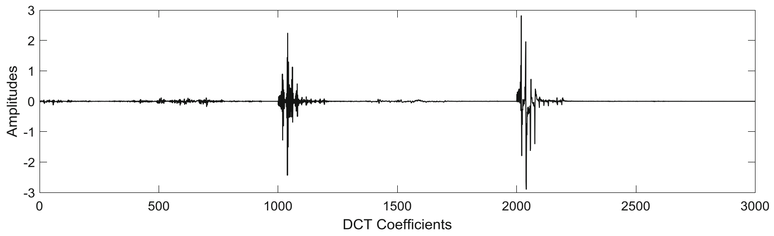

Fig. 4. DCT coefficients of the 3 frames

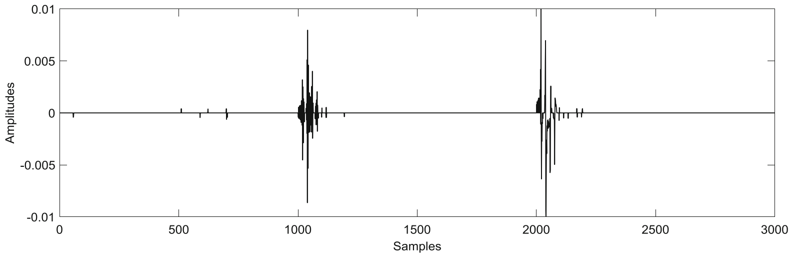

Fig. 5. Constructed DCT coefficients based on compressed signal

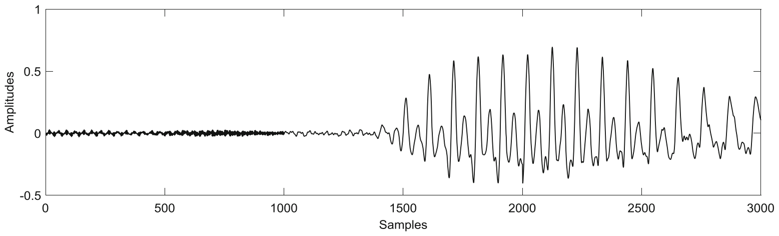

Fig. 6. Reconstructed signal

3 The Scheme

3.1 Preprocessing

Step 1: Divide the signal $A = \{a_l | 1 \leq l \leq L\}$ into divide the signal into N-sample frames. $A_i = \left\{a_{i,j} | 1 \leq j \leq N\right\}$ represents the i-th frame, $1 \leq i \leq P$, and $P = L/N$.
Step 2: We compress A_i with the method in Sect. 2.2, and obtain the compressed signal $\bar{C}_i$.

The compressed signal should not be embedded into the current frame, because the reconstruction will fail in the case that the current frame itself is malicious attacked. In order to reconstruct A_i, the compressed signal $\bar{C}_i$ should be embedded into other frame except for A_i. In this paper, we adopt the method in [18] to scramble the compressed signal $\bar{C}_i$, and denote the i-th compressed signal after being scrambled as SC_i, $1 \le i \le P$. The initial value of Logistic chaotic mapping k_0 and the parameter μ are regarded as the key.

Step 3: We map frame number i to $F_i = (i/(10)^{n_1})^{|n|}$. In this paper, we define $Y^{|n|}$ as the Eq. (2), and $P < 10^{n_1}$.

$$Y^{|n|} = \underbrace{Y \cup Y \cup \cdots \cup Y}_{n} \tag{2}$$

Step 4: Denote $W_i = (SC_i)^{|n_1|}$, $F_i \cup W_i$ as watermark embedded into A_i.

3.2 Embedding

Step 1: We scramble the sample of A_i by using the method [18], and then perform DCT on the scrambled signal. Denote the DCT coefficients as $C_i = \{c_1, c_2, \cdots, c_N\}$.
Step 2: We use F_i to substitute the n coefficient $c_1, c_2, \ldots, c_n$, and use W_i to substitute the $3N/25$ coefficients, amplitudes less than 0.1.
Step 3: Perform inverse DCT on the DCT coefficients after being substituted, and anti-scrambling on the signal obtained to generate the watermarked signal.

By using the method, we can embed F_i and W_i into A_i, $1 \le i \le N$.

3.3 Forensics and Tamper Recovery

Suppose $A' = \{a'_l | 1 \le l \le L\}$ represents the watermarked signal. The steps of forensics and tamper recovery are described in following.

Step 1: Cut A' into N length frames, and denote the i-th frame as A'_i, $1 \le i \le L/N$.
Step 2: We scramble the samples of the signal A'_i, and perform DCT on the scrambled signal. Denote the DCT coefficients as $C'_i = \{c'_1, c'_2, \cdots, c'_N\}$.
Step 3: Extract frame number from the first n DCT coefficients, $c'_1, c'_2, \cdots, c'_n$.
By using Eq. (3), we calculate $F1_i = f\left(c'_1, c'_2, \cdots, c'_{n/2}\right)$. Similarly, we can get $F2_i = f\left(c'_{n/2+1}, c'_{n/2+2}, \cdots, c'_n\right)$.

$$f\left(c'_1, c'_2, \cdots, c'_{n/2}\right) = \sum_{l=1}^{n/2} \left\lfloor \left(100 \times c'_l\right) + \frac{1}{2} \right\rfloor / n/2 \tag{3}$$

If $F1_i = F2_i$, it indicates that the signal A'_i is authentic. Otherwise, it indicates that the i-th frame has been tampered, and the tamper location and tamper recovery steps are described in following.

Step 4: Tamper location and tamper recovery. Suppose the 1st to i-1th frame are all intact, and the next N samples cannot pass the authentication.

(1) Search the next N samples that can be authenticated, and denote the frame as $A'_{i'}$. We can extract frame number from $A'_{i'}$ $(F1_{i'} = F2_{i'})$, and denote the frame number as i'.
(2) We regard the frame between i-1th to i' th as the attacked signal.
(3) Based on the scrambling method, suppose the compressed signal used for reconstructing one attacked frame is embedded into the $\bar{i}$-th frame $A'_{\bar{i}}$. We scramble the sample of $A'_{\bar{i}}$, and perform DCT. Then, by using the principle of the minority subordinating to the majority, we extract compressed signal from the $3N/25$ coefficients, amplitudes less than 0.1. Then we can reconstruct the attacked frame using the method in Sect. 2.2.

4 Performance Analysis and Experimental Results

In this section, the comprehensive performance of the scheme is analyzed and tested, including inaudibility, security, robustness, tamper location and tamper recovery. We select 500 speech signals recorded in four different environments as the test signal, denoted by T1, T2, T3 and T4. They represent the signals recorded in quiet room, seminar, park and railway station, respectively. The signals are sampled at 8 kHz, WAVE format 16-bit quantified mono signals. The parameters are set as follows, $L = 80000$, $P = 80$, $N = 1000$, $n = 5$, $n_1 = 3$, $n_2 = 3$, $k_0 = 0.82$, $\mu = 3.9875$.

4.1 Inaudibility

Inaudibility means that watermark embedding is inaudible, and reflects the change degree of original speech after watermarking. We use signal to noise ratio (SNR) and subjective difference grades (SDG) to test the inaudibility of the scheme proposed. The definition of SNR is in [19], and the meaning of the scores of SDG is in the references [18].

The mean values of SNR and SDG of the four types watermarked speech signal are listed in Table 1. SDG values are acquired from 30 listeners. The test results listed in Table 1 indicate that the scheme proposed is inaudibility.

Table 1. The SDG and SNR values of watermarked signals

Speech type	SDG	SNR(dB)
T1	−0.78	26.64
T2	−0.72	26.73
T3	−0.66	27.41
T4	−0.53	28.26

4.2 Robustness

For the convenience of storage and playing, and many other reasons, speech signal may be subjected to some signal processing operations. If watermark embedding is fragile, it will extract false watermark, and regard common signal processing operation as hostile attack. Thereby the authentication schemes should be robust against signal processing operations. We use bit error rate (BER) [20] to test the robustness of the proposed scheme. BER is defined by Eq. (4), and less BER value implies stronger robustness to signal processing operations.

$$BER = \frac{W_e}{W_t} \tag{4}$$

where W_e represents the number of watermark erroneously extracted, and W_t represents the number of watermark embedded.

We list the average BER value of the 800 test signals, after being subjected to different signal processing operations, containing MP3 compression, re-sampling and low pass filtering. And compare the results with the schemes proposed in [21, 22], which are shown in Table 2. The results shown in Table 2 indicate that the scheme proposed has the ability to tolerate common signal processing operations.

Table 2. BER values after being subjected to common signal processing

Signal processing		BER		
		Ref. [21]	Ref. [22]	Proposed
MP3 compression	64 kbps	0.0748	0.0919	0.0612
	96 kbps	0.0516	0.0697	0.0452
	128 kbps	0.0482	0.0465	0.0364
Low pass filtering	6 kHz	0.0685	0.0662	0.0209
Gauss noise	30 dB	0.0553	0.0684	0.0436
Echo	40% and 100 ms delay	0.0470	0.0446	0.0335

4.3 Tamper Location and Tamper Recovery

We select one speech from the 800 test signals, and show the signal in Fig. 7. Authors in [23] say that all attack channels can be viewed as deletion, insertion and substitution channel for watermarking. We perform the 3 types attack on the signal shown in Fig. 7, and then give the tamper location and tamper recovery results.

Because of space cause, we only give the detailed steps, tamper location and tamper recovery results of deletion attack. For other attacks, the tamper recovery results are similar.

(1) Delete 8000 samples from the watermarked speech shown in Fig. 7, and show the attacked signal in Fig. 8.
(2) Cut the attacked signal into N length frames, and scramble the first frame. Then we perform DCT on first frame, and extract frame number $F1_1$ and $F2_1$.

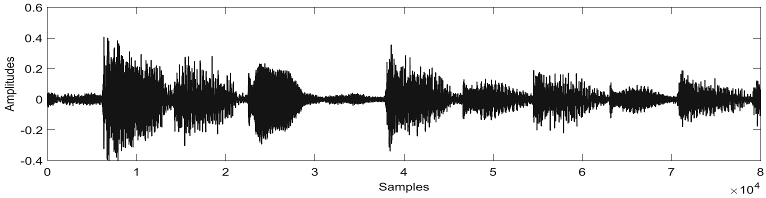

Fig. 7. Watermarked speech

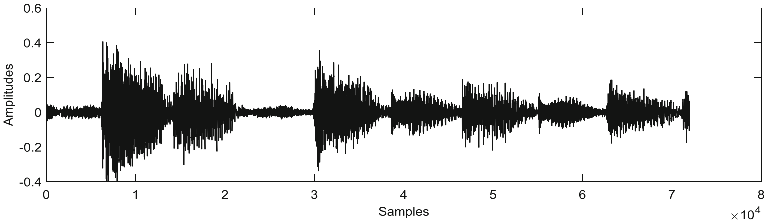

Fig. 8. Watermarked speech of 8000 samples deleted

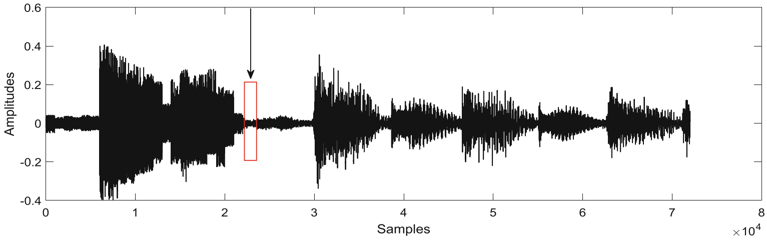

Fig. 9. Finding the *N* successive samples that cannot pass authentication

If $F1_1 = F2_1$, we regard the first frame is intact, and record $F1_1$(or $F2_1$) as the frame of the first frame. By using the method, we can verify the authenticity of next frames, until finding the *N* successive samples that cannot pass authentication. The result is shown in Fig. 9.

(3) Move and verify the next N successive samples, until that the N samples can pass the authentication. We show the result in Fig. 10, and reconstruct the frame number.

(4) The difference between the two reconstructed frame numbers is the frame attacked, as the tamper location result shown in Fig. 11, in which $TL = 1$ represents the frame is attacked, and $TL = 0$ represents the frame is intact.

Based on scrambling method, we find and extract the compressed signal of attacked frames, 23rd to 30-th frame, from 76th, 69th, 21st, 45th, 13th, 31th, 61th and 65th frame.

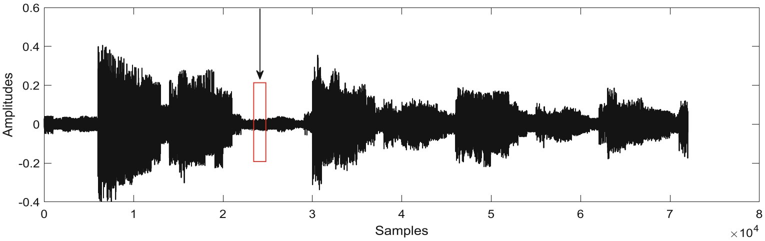

Fig. 10. Finding next *N* successive samples can pass the authentication

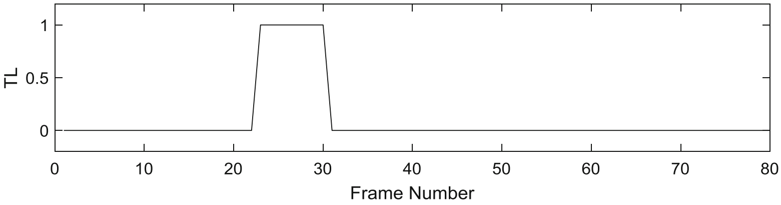

Fig. 11. Tamper location result

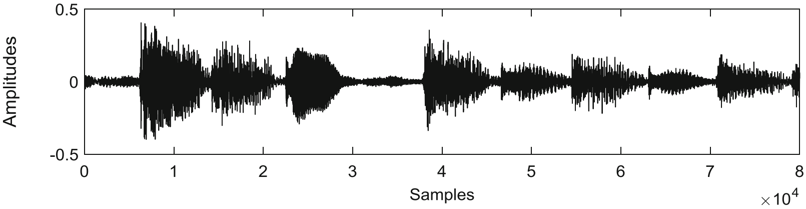

Fig. 12. Tamper recovery result of deletion attack

(5) By using the reconstruction method, we reconstruct the attacked frames based on the compressed signal to perform tamper recovery, and show the result in Fig. 12.

From the tamper recovery results above, we get the conclusion that scheme proposed can locate the attacked frames precisely, and has a good ability of tamper recovery.

5 Conclusion

In this paper, by using the proposed speech compression method, we obtained the compressed speech, used for the reconstruction of attacked frames. We embedded frame number and compressed signal into host speech, by the substitution of less amplitude DCT coefficients. Using frame number to locate the attacked frame, after that we extracted compressed signal and reconstructed the attacked frame to perform tamper recovery. Simulation results demonstrate that the proposed recoverable algorithm is inaudible and robustness to signal processing operations. Not only can locate the attack frames, but can recover the attacked signals.

Acknowledgments. This paper is supported by the National Natural Science Foundation of China (Grant No. 61502409, 61602318, 61602318, 61631016), and Nanhu Scholars Program for Young Scholars of XYNU. We would like to thank the anonymous reviewers for their constructive suggestions.

References

1. Hua, G., Huang, J.W., et al.: Twenty years of digital audio watermarking—a comprehensive review. Signal Process. **128**(11), 222–242 (2016)
2. Nishimura, A.: Audio watermarking based on subband amplitude modulation. Acoust. Sci. Technol. **32**(5), 328–336 (2010)
3. Hu, P., Peng, D., Yi, Z., Xiang, Y.: Robust time-spread echo watermarking using characteristics of host signals. Electron. Lett. **52**(1), 5–6 (2016)
4. Li, R.K., Yu, S.Z., Yang, H.Z.: Spread spectrum audio watermarking based on perceptual characteristic aware extraction. IET Signal Process. **10**(3), 266–273 (2016)
5. Chen, B., Wornell, G.W.: Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. IEEE Trans. Inf. Theory **47**(4), 1423–1443 (2001)
6. Natgunanathan, I., Xiang, Y., et al.: Robust patchwork-based embedding and decoding scheme for digital audio watermarking. IEEE Trans. Audio Speech Lang. Process. **20**(8), 2232–2239 (2012). IEEE Signal Processing Society
7. Xiang, Y., Natgunanathan, I., et al.: Patchwork-based audio watermarking method robust to de-synchronization attacks. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(9), 1413–1423 (2014)
8. Hu, H.T., Hsu, L.Y.: Robust, transparent and high-capacity audio watermarking in DCT domain. Signal Process. **109**(3), 226–235 (2015)
9. Kang, X., Yang, R., Huang, J.: Geometric invariant audio watermarking based on an LCM feature. IEEE Trans. Multimed. **13**(2), 181–190 (2011)
10. Erfani, Y., Pichevar, R., Rouat, J.: Audio watermarking using spikegram and a two-dictionary approach. IEEE Trans. Inf. Forensics Secur. **12**(4), 840–852 (2017)
11. Liu, Z.H., Wang, H.X.: A novel speech content authentication algorithm based on Bessel-Fourier moments. Digit. Signal Process. **24**(1), 197–208 (2014)
12. Chen, O.T.C., Chia, H.L.: Content-dependent watermarking scheme in compressed speech with identifying manner and location of attacks. IEEE Trans. Audio Speech Lang. Process. **15**(5), 1605–1616 (2007)
13. Luo, D., Sun, M.M., Huang, J.W.: Audio postprocessing detection based on amplitude cooccurrence vector feature. IEEE Signal Process. Lett. **23**(5), 688–692 (2016)
14. Xia, Z., Wang, X., Sun, X., Liu, Q., Xiong, N.: Steganalysis of LSB matching using differences between nonadjacent pixels. Multimed. Tools Appl. **75**(4), 1947–1962 (2016)
15. Chamlawi, R., Khan, A., Usman, I.: Authentication and recovery of images using multiple watermarks. Comput. Electr. Eng. **36**(3), 578–584 (2010)
16. Li, C.L., Wang, Y.H., Ma, B., Zhang, Z.X.: Tamper detection and self-recovery of biometric images using salient region-based authentication watermarking scheme. Comput. Stand. Interfaces **34**(4), 367–379 (2012)

17. Roldan, L.R., Hernandez, M.C., Miyatake, M.N., Meana, H.P., Kurkoski, B.: Watermarking-based image authentication with recovery capability using halftoning technique. Signal Process. Image Commun. **28**(1), 69–83 (2013)
18. Liu, Z.H., Zhang, F., Wang, J., Wang, H.X., Huang, J.W.: Authentication and recovery algorithm for speech signal based on digital watermarking. Signal Process. **123**(1), 157–166 (2016)
19. Hu, H.T., Chang, J.R., Lin, S.J.: Synchronous blind audio watermarking via shape configuration of sorted LWT coefficient magnitudes. Signal Process. **147**(1), 190–202 (2018)
20. Hu, H.T., Hsu, L.Y., Chou, H.H.: Perceptual-based DWPT-DCT framework for selective blind audio watermarking. Signal Process. **105**(12), 316–327 (2014)
21. Liu, Z.H., Luo, D., Huang, J.W., Wang, J., Qi, C.D.: Tamper recovery algorithm for digital speech signal based on DWT and DCT. Multimed. Tools Appl. **76**(10), 12481–12504 (2017)
22. Ali, A.H.: An imperceptible and robust audio watermarking algorithm. EURASIP J. Audio Speech Music Process. **37**(1), 1–12 (2014)
23. Wang, Y., Wu, S.Q., Huang, J.W.: Audio watermarking scheme robust against desynchronization based on the dyadic wavelet transform. J. Adv. Signal Process. **13**(1), 1–17 (2010)