

Tamper recovery algorithm for digital speech signal based on DWT and DCT

Z. H. Liu^{1,2} · D. Luo¹ · J. W. Huang¹ ·
J. Wang³ · C. D. Qi³

Received: 18 November 2015 / Revised: 15 April 2016 / Accepted: 3 June 2016 /
Published online: 14 June 2016
© Springer Science+Business Media New York 2016

Abstract It is a challenging work to design tamper recovery schemes for digital speech signal. Briefly, there are two problems need to be solved. One is that the signals used to tamper recovery are difficult to generate and embed, and the second is that it's hard to tamper location precisely for attacked speech signal. In this paper, compression and reconstruction method based on discrete wavelet transform (DWT) and discrete cosine transform (DCT) is given, to obtain the compressed signals used to tamper recovery. And then frame number and compressed signals are embedded based on block-based method. Attacked signal can be located by frame number, and compressed signals are extracted and used to reconstruct the attacked signal. Theory analysis and experimental results indicate that the scheme proposed not only improves the accuracy of tamper localization, but also can reconstruct the attacked signals.

✉ J. W. Huang
jwhuang@szu.edu.cn

Z. H. Liu
zhenghui.liu@163.com

D. Luo
luoda@szu.edu.cn

J. Wang
wangjing_cosy@163.com

C. D. Qi
qichuanda@sina.com

¹ College of Information Engineering and Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China

² College of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China

³ College of Mathematics and Information Science, Xinyang Normal University, Xinyang 464000, China

Keywords Digital watermark · Tamper recovery · Speech compression · Tamper localization

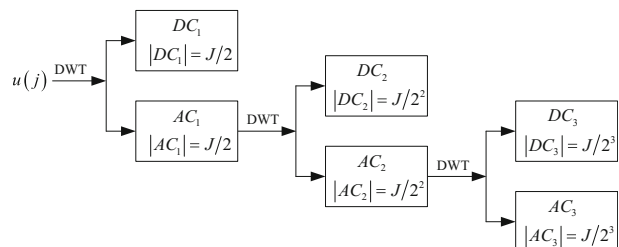
1 Introduction

With the development of internet and multimedia technology, the authentication of digital media becomes more and more important. And much research has been done in this field [10, 14]. In this paper, we limit our attention to digital speech forensics. On the one hand, speech signal often implies some instructions. If users consider the attacked signal is the original one and act according to the instructions of the attacked signal, it may cause serious consequences. On the other hand, speech signals are more likely to cause attacker's interest and be maliciously attacked. So, as to recipient, verifying the integrity and authenticity of received speech signal is the first and essential step. Fortunately, the forensic technology based on digital watermarking [1, 9, 17, 18] gives a method to verify the authenticity of digital speech.

For speech signals, most of the results are focused on speaker recognition and identification [5, 6, 15, 20]. For the existed content authentication schemes, there are shortcomings more or less. In [16], a speech authentication scheme based on digital watermarking and pattern recovery is proposed. By using the cyclic pattern embedding method, the scheme has the ability of tamper location. While the method increases the load of the watermarked signal. In [3], an authentication scheme based on compression technique and codebook-excited linear prediction for compressed speech is proposed. In the paper, watermark bits are generated by the features extracted during compression process based on codebook-excite linear prediction. For speech signals compressed based on other speech codecs, the scheme is ineffective. The watermark bits are embedded based on least significant bits (LSBs), which is fragile to signal processing operations. As to the scheme, signal processing will be regarded as hostile attack. Therefore the scheme is powerless in some applications. In [24], authors introduced an integrity and authenticity mechanic for real-time multimedia communication, and produced a method for real-time speech integrity and authentication incorporating with GSM 610 full-rate coder. And a speech authentication scheme mainly used in the real-time communication process is proposed. For some signals saved on a hard drive, the scheme is unsuitable.

The expressed meaning of hostile attacked speech signal is different to the original one, which may cause serious consequences. If one speech is attacked and the instruction is an urgent task, maybe the greatest wish is to acquire the meaning of the original one. So, reconstructing the attacked signal is the users to pray in this case. Currently, there has been a considerable amount of work on authentication and recovery for digital

Fig. 1 The structure of the DWT



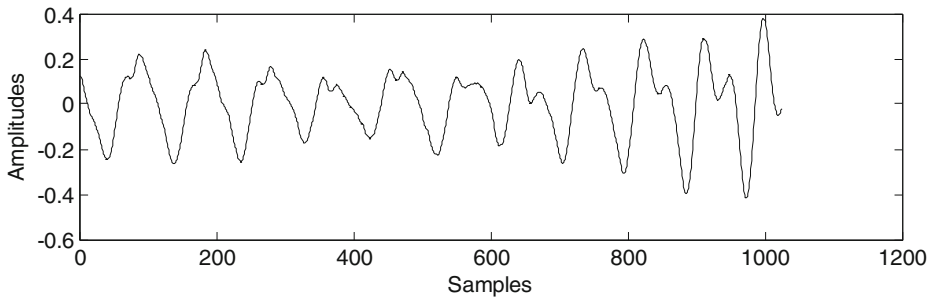


Fig. 2 The original signal after down-sampling

images [2, 7, 11, 19]. It's a pity that, there are comparatively few schemes for digital speech [4, 13].

For tamper recovery schemes, the first step is to locate the attacked signals precisely. But so far the problem has not been solved well. When it comes to the problem of tamper location, the first method thought is based on synchronization code [8, 21, 22]. Indeed the method can locate watermarked signal. However, they have some shortcomings. (1) Synchronization codes embedded based on public features are vulnerable to substitution attack [12]. (2) For the schemes, the content between two neighboring synchronization codes is regarded as the watermarked signal. But the authenticity of the watermarked signal is not verified. So, it can locate the watermarked signal only, but not locate the signal attacked. (3) For the short-time stationarity of speech signal, the synchronization codes can be extracted from some segments, but not only one. Therefore, the schemes can locate watermarked signal roughly not precisely.

Considering the background and motivation above, the tamper recovery algorithm for digital speech is proposed. One compression and reconstruction method for speech signal based on discrete wavelet transform (DWT) and discrete cosine transform (DCT) is given. Compressed signal is as watermark and comprise approximate coefficients from DWT and some DCT coefficients. Approximate coefficients are used to reconstruct the approximate signal, and DCT coefficients are used to reconstruct the residual signal. Then original signal can be obtained by combining the approximate signal and residual signal. Frame number and compressed signal are embedded based on block-based method. Frame number is used for tamper localization, and compressed signals are used to reconstruct the attacked frames. Theoretical analysis and experimental evaluation demonstrate that the scheme has ability of tamper localization, and can recover the attacked signal.

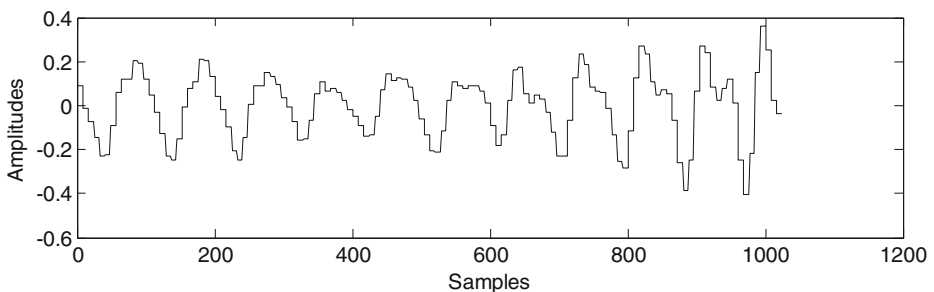


Fig. 3 The signal *R1* by inverse DWT

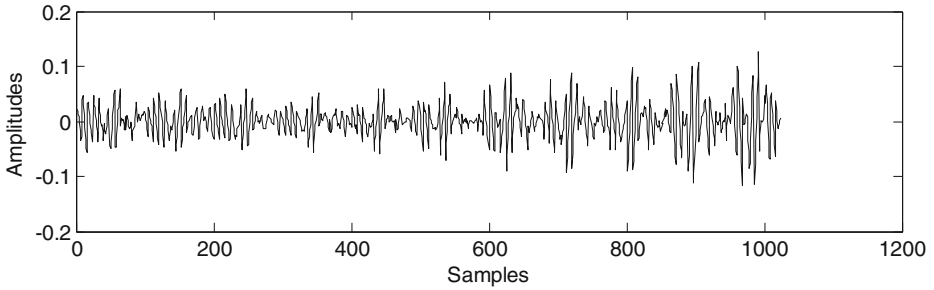


Fig. 4 The signal $R2$ by inverse DCT

The remainder of this paper is organized as follows. Section 2 shows the fundamental theory for the scheme proposed. Section 3 gives the recovery scheme based on digital watermark. Section 4 analyzes the performance of the algorithm theoretically. The simulation results are presented in Section 5. Finally, some conclusions are included in Section 6.

2 Fundamental theory

2.1 Discrete Wavelet Transform

Discrete wavelet transform (DWT) can be viewed as the multiresolution decomposition of a sequence. It takes a length J sequence $u(j)$, and generates an output sequence of length N . The output is the multiresolution representation of $u(j)$. It has $J/2$ values at the highest resolution, $J/4$ values at the next resolution, and so on.

The structure of the DWT is due to the dyadic nature of its time-scale grid, shown in Fig. 1, in which AC_q and DC_q represent the approximate and detail coefficients from q -level DWT, respectively. The length of AC_q is $J/2^q$, equal to DC_q .

2.2 Compressed signal generation

In this paper, speech signal is compressed, and the compressed signal is as watermark and embedded. Denote A as the original speech signal, the method of compressed signal generation is as follows.

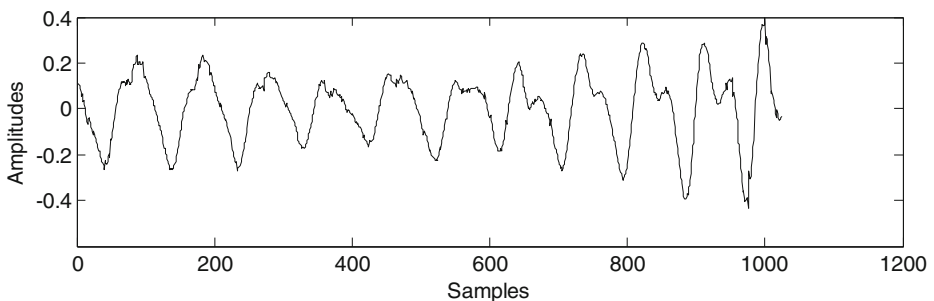
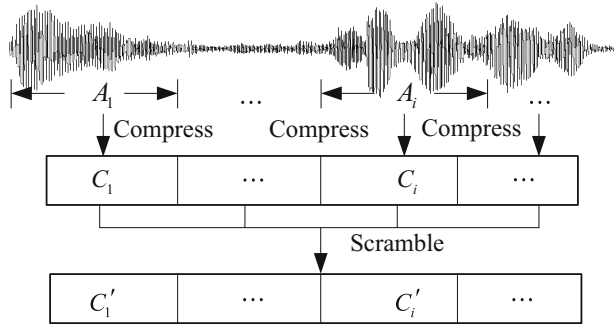


Fig. 5 The signal combined by $R1$ and $R2$

Fig. 6 Scrambling the compressed signal



2.2.1 Down sampling and DWT

- Step 1 Cut A into P frames, and the length of each frame is N . The i -th frame is denoted by A_i .
- Step 2 Re-sampling the speech signal A , and the sampling frequency is dropped from f to f' , where f denotes the sampling frequency of original signal, and f' denotes that after re-sampling. The signal re-sampled is denoted by A' .
- Step 3 A' is cut into P frames. The i -th frame is denoted by A'_i , and the length is N' ($N' = N \cdot f'/f$).
- Step 4 D -level DWT is performed on A'_i , and the approximate coefficient is denoted by $C1_i$, which is as one part of the compressed signal. And the length of $C1_i$ is M , $M = N'/2^D$.

2.2.2 DCT on residual signal

- Step 1 Reset the detail coefficients to 0.
- Step 2 D -level inverse DWT is performed on the approximate coefficient $C1_i$ and the detail coefficients. The signal obtained is denoted by IA'_i .
- 3. Step 3: The difference between A'_i and IA'_i is regarded as the residual signal, denoted by E_i , which can be obtained by Eq. 1.

$$E_i = A'_i - IA'_i \tag{1}$$

- Step 4 DCT is performed on the residual signal E_i , $E_i = \{e_1, e_2, \dots, e_{N'}\}$. Rearrange $|e_i|$ ($1 \leq i \leq N'$) in order of large to small, and denote the $M/2$ -th value as $|e_{M/2}|$.
- Step 5 Record the coefficients (amplitude larger than $|e_{M/2}|$) and the location, as another part of compressed signal, denoted by $C2_i$ and $C3_i$, respectively.
Let $C_i = C1_i \cup C2_i \cup C3_i$ as the compressed signal of A_i .

Fig. 7 Segmentation method

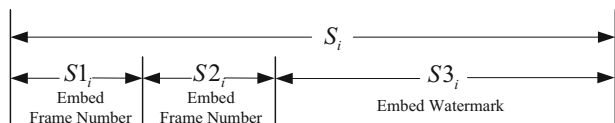


Table 1 An example of quantifying method

Samples		V		Quantified	
$s1_1$	0.5692	9	$y_3 = 2$	$s1_1$	0.5692
$s1_2$	-0.3817			$s1_2$	-0.3817
$s1_3$	0.3271			$s1_3$	0.3371
$s1_4$	-0.4128	2	$y_2 = 7$	$s1_4$	-0.4128
$s1_5$	-0.6534			$s1_5$	-0.6634
$s1_6$	0.2326			$s1_6$	0.2426
$s1_7$	0.1125	1	$y_1 = 2$	$s1_7$	0.1225
$s1_8$	0.3278			$s1_8$	0.3278
$s1_9$	-0.4569			$s1_9$	-0.4569

2.3 Reconstruction

- Step 1 Perform D -level inverse DWT on the approximate coefficient $C1_i$ and the 0 value detail coefficients, to generate the approximate signal $R1$, as the first part of reconstructed signal.
- Step 2 Perform inverse DCT on the coefficients obtained by the Step 5 in Section 2.2.2, to reconstruct the residual signal $R2$, as the second part of reconstructed signal.
- Step 3 Combine the approximate and residual signal by $R1 + R2$.
- Step 4 Resample the signal obtained in Step 3, using the sampling frequency f , to obtain the reconstructed signal.

As an example, one signal is selected, and the reconstructed signal is given. In this section, the length of the signal after down-sampling is $N' = 1024$, $D = 3$, $M = 128$. Figure 2 is the speech signal after down-sampling. Figure 3 is the approximate signal ($R1$), obtained by inverse DWT on approximate coefficients $C1_i$ and 0 value detail coefficients. Figure 4 is the residual signal, obtained by inverse DCT. And Fig. 5 is the signal combined by $R1$ and $R2$.

Based on the results shown above, it can be seen that the signal shown in Fig. 5 is close to that shown Fig. 2, which demonstrates that the original speech can be reconstructed approximately based on the compressed signal.

Table 2 The partition result

6 samples	B_1		B_2		B_3		B_4		B_5		B_6	
-0.3125	1	3			1	1			1	2		
-0.2191	1	2			1	1			1	9		
0.1567	0	1			0	5			0	6		
0.2237			0	2			0	2			0	3
-0.5123			1	5			1	1			1	2
0.6318			0	6			0	3			0	1

Table 3 The embedding method for $c_1 = -1.3551$

6 samples	$B_1 \rightarrow B'_1$ ("–")		$B_3 \rightarrow B'_3$ (c_1^2)		$B_5 \rightarrow B'_5$ (c_1^4)		Samples quantified
-0.3125	3	3	1	1	2	2	-0.3125
-0.2191	2	<u>3</u>	1	1	9	<u>8</u>	-0.3181
0.1567	1	1	5	<u>6</u>	6	<u>5</u>	0.1657
	$B_2 \rightarrow B'_2$ (c_1^1)		$B_4 \rightarrow B'_4$ (c_1^3)		$B_6 \rightarrow B'_6$ (c_1^5)		
0.2237	2	2	2	<u>3</u>	3	3	0.2337
-0.5123	5	5	1	1	2	2	-0.5123
0.6318	6	6	3	3	1	1	0.6318

2.4 Embedding strategy

For watermarked speech, if one frame is attacked, watermark embedded in the frame will be destroyed too. So, the compressed signal of one frame should be embedded into other non-attacked frame.

For the scheme proposed, the compressed signals are scrambled before embedding, aiming to make the compressed signal is embedded into other non-attacked frame. The strategy is shown in Fig. 6.

In Fig. 6, A_i denotes the i -th frame of the speech signal, C_i denotes the compressed signal of i -th frame. C'_i denotes the i -th compressed signal after being scrambled. For the proposed scheme, C'_i is the signal that will be embedded into the i -th frame A_i .

3 The scheme

Denote $A = \{a_l | 1 \leq l \leq L\}$ as original speech signal, where a_l represents the l -th sample. The recovery scheme is described as follows.

Fig. 8 The process of watermark embedding

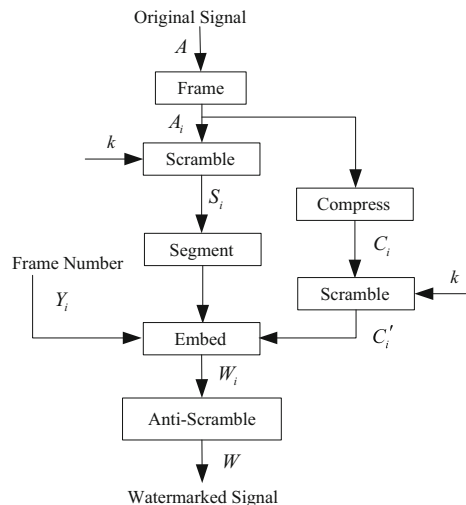
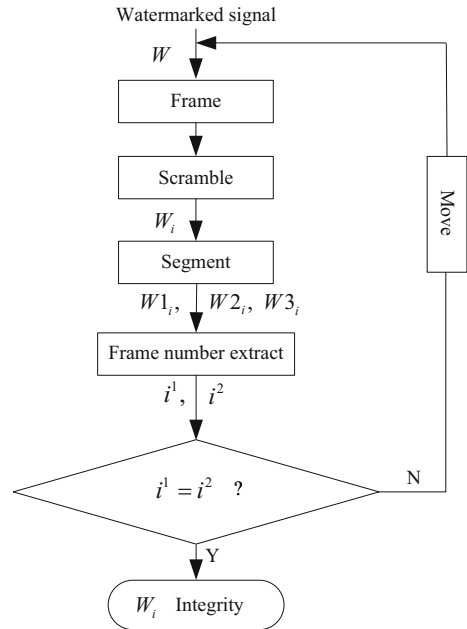


Fig. 9 The process of content authentication



3.1 Preprocessing

Step 1 Cut A into P frames, and N is the length of each frame. Denote $A_i = \{a_{i,j} | 1 \leq j \leq N\}$ as the i -th frame, $1 \leq i \leq P, N = L/P$.

Step 2 By using the compression method in Section 2.2, the compressed signal of A_i can be obtained, denoted by C_i . Then all the compressed signals $C_i (1 \leq i \leq P)$ are scrambled using chaotic address index sequence.

Denote $X = \{x_i | i = 1, 2, \dots, P\}$ as the pseudo-random sequence. The elements of X are generated by the Logistic chaotic mapping shown in Eq. (2).

$$x_{i+1} = \mu x_i(1-x_i), \quad x_0 = k, \quad 3.5699 \leq \mu \leq 4 \tag{2}$$

where k is the initial value and as key of the system. The elements of X are sorted in ascending order shown in Eq. (3), where $h(i)$ is the address index of the sorted chaotic sequence.

$$x_{h(i)} = \text{ascend}(x_i), \quad i = 1, 2, \dots, P \tag{3}$$

Denote C'_i as the compressed signal after being scrambled, $C'_i = C_{h(i)}, 1 \leq i \leq P$.

Fig. 10 The processes of tamper location

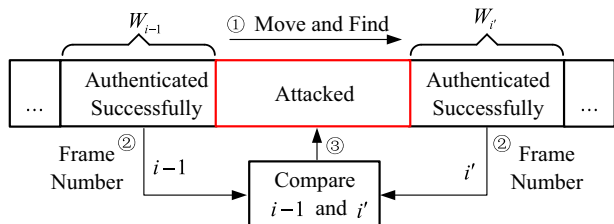
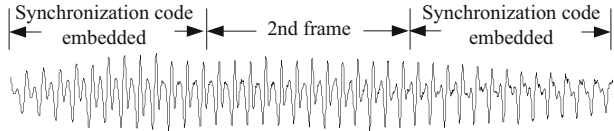


Fig. 11 The signal synchronization code embedded



- Step 3 Scrambling the samples of A_i by using the same method above, denoted by S_i , $S_i = \{s_{i,j}, 1 \leq j \leq N\}$, where $s_{i,j} = a_{i,h(j)}, j = 1, 2, \dots, N$.
- Step 4 Divide S_i into 3 parts, denoted by $S1_i, S2_i$ and $S3_i$, respectively. The length of $S1_i$ and $S2_i$ is N_1 , and the length of $S3_i$ is N_2 .
- Step 5 The frame number i is mapped to the sequence of integers $Y_i = \{y_n, y_{n-1}, \dots, y_1\}$, and the elements can be obtained by Eq. (4).

$$i = y_n \cdot 10^{n-1} + y_{n-1} \cdot 10^{n-2} + \dots + y_1 \tag{4}$$

In this paper, Y_i is as the identifier of i -th frame and embedded into $S1_i$ and $S2_i$, respectively. The compressed signal C'_i is embedded into $S3_i$. The method of segmentation is shown in Fig. 7.

3.2 Watermark Embedding

3.2.1 Embed frame number

$Y_i = \{y_n, y_{n-1}, \dots, y_1\}$ as the identifier of i -th frame is embedded into $S1_i$ and $S2_i$, respectively. Denote $S1_i = \{s1_1, s1_2, \dots, s1_{N_1}\}$, y_n is embedded into the first 3 consecutive samples $s1_1, s1_2$ and $s1_3$. The embedding method is described as follows.

- Step 1 Denote $z_m = \lfloor |100 \cdot s1_m| \rfloor \bmod 10$, where $\lfloor \cdot \rfloor$ returns the largest integer less than the original value, $1 \leq m \leq 3$. $v_m = \text{sign}(s1_m)$, if $s1_m \geq 0$, $v_m = 0$, and if $s1_m < 0$, $v_m = 1$. For $z_m = 0$, set $z_m = 1$. Calculate $V = f(z_1, z_2, z_3, v_1, v_2, v_3)$ according to the Eq. (5).

$$f(z_1, z_2, z_3, v_1, v_2, v_3) = [(v_1 + v_2 + v_3 + z_1) \times 1 + z_2 \times 2 + z_3 \times 3] \bmod 10 \tag{5}$$

If $y_n = V$, $z_m, 1 \leq m \leq 3$, are not need to be quantified.

If $y_n \neq V$, quantify z_1, z_2 or z_3 , to make $y_n = V$, under the condition that the original values are as close as possible to that after being quantified.

The quantitative method is $z_1 \pm 1, z_2 \pm 1$ or $z_3 \pm 1$. Using the quantified values substitute z_1, z_2 and z_3 , to embed y_n . For example, if $s1_1 = 0.5692, s1_2 = -0.3817, s1_3 = 0.3271$ and $y_n = 2$, it's got that $z_1 = 6, z_2 = 8, z_3 = 2, v_1 = 0, v_2 = 1, v_3 = 0$. Based on Eq. (5), $V = 9$. In order to make $V = y_n$, set $z'_3 = z_3 + 1$. z'_3 is the value after being

Fig. 12 The synchronization code embedded signal subjected to deletion attack

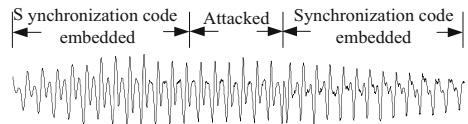
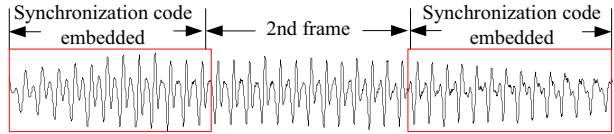


Fig. 13 Searching result for synchronization code



quantified.

Step 2 Using the same method above, y_n, y_{n-1}, \dots, y_1 are embedded into the two segments $S1_i$ and $S2_i$.

The quantified signal is denoted by $W1_i$ and $W2_i$. To clearly show the embedding method, an example is given in Table 1, $i = 272, Y_i = \{2, 7, 2\}, n = 3$.

3.2.2 Embed compressed signal

Denote $C'_i = \{c_t, 1 \leq t \leq 2M\}$ as the compressed signal, which is embedded into $S3_i$ based on the block-based method [13], and $S3_i = \{s3_1, s3_2, \dots, s3_{N_2}\}$.

For each coefficient in C'_i , sign (“+” or “-”) and five numbers are embedded. Let $c_1 \in C'_i$ as the first coefficient, and denote $c_1^1 = \lfloor |c_1| \rfloor \bmod 10, c_1^2 = \lfloor |10 \cdot c_1| \rfloor \bmod 10, c_1^3 = \lfloor |100 \cdot c_1| \rfloor \bmod 10, c_1^4 = \lfloor |1000 \cdot c_1| \rfloor \bmod 10$ and $c_1^5 = \lfloor |10000 \cdot c_1| \rfloor \bmod 10$, which are the five numbers of c_1 . For example, if $c_1 = 1.3628$, sign is “+”, and the five numbers are 1, 3, 6, 2 and 8. They are embedded into the first 6 consecutive samples of $S3_i$.

Let the 6 samples as $s3_1, s3_2, \dots, s3_6$. Partition them into 6 blocks, denoted by B_1, B_2, \dots, B_6 , respectively. $B_1 = \{b_1^1, b_1^2, b_1^3, b_1^4, b_1^5, b_1^6\}$, where $b_1^1 = \text{sign}(s3_1), b_1^2 = \text{sign}(s3_2), b_1^3 = \text{sign}(s3_3), b_1^4 = \lfloor |10 \cdot s3_1| \rfloor, b_1^5 = \lfloor |10 \cdot s3_2| \rfloor, b_1^6 = \lfloor |10 \cdot s3_3| \rfloor$. $B_2 = \{b_2^1, b_2^2, b_2^3, b_2^4, b_2^5, b_2^6\}$, where $b_2^4 = \lfloor |10 \cdot s3_4| \rfloor, b_2^5 = \lfloor |10 \cdot s3_5| \rfloor, b_2^6 = \lfloor |10 \cdot s3_6| \rfloor$. By using the same method, other blocks can be obtained.

As an example, 6 samples are selected, and the partition result is shown in Table 2. Based on the partition, the embedding method is as follows.

Step 1 Embed the sign of c_1 , “+” or “-” into B_1 .

Calculate the sum of 6 values in B_1 , which is denoted by T . If $1 \leq c_1$, and $T \bmod 2 = 1$, quantify b_1^4, b_1^5 or b_1^6 , to make $T \bmod 2 = 0$ and embed the sign “+”, such as $b_1^4 + 1$, or $b_1^4 - 1$. If $c_1 < 0$, and $T \bmod 2 = 0$, quantify b_1^4, b_1^5 or b_1^6 , to make $T \bmod 2 = 1$ and embed the sign “-”. For other conditions, the values in B_1 remain the same without modification.

Step 2 Embed the five numbers of c_1 ($c_1^1, c_1^2, c_1^3, c_1^4, c_1^5$) into B_2, B_3, \dots, B_6 , respectively.

Table 4 Comparison of abilities for different schemes

Scheme	Features Used	AI	All
[18]	Feature points (public)	No	Yes
[8, 21, 22]	Energy (public)	No	No
[23]	DYWT (public)	No	No
Proposed	Samples scrambled (secret)	Yes	Yes

Table 5 Subjective and objective difference grades

SDG/ODG	Description of impairments	Quality
0.0	Imperceptible	Excellent
-1.0	Perceptible, but not annoying	Good
-2.0	Slightly annoying	Fair
-3.0	Annoying	Poor
-4.0	Very annoying	Bad

- (1) Based on the values in B_2 , V can be obtained by using the Eq. (5). Using the same method of frame number embedding in Section 3.2.1, c_1^1 can be embedded by quantifying b_2^4, b_2^5 and b_2^6 .
 - (2) Similarly, c_1^2, c_1^3, c_1^4 and c_1^5 are embedded into the blocks of B_3, B_4, B_5 and B_6 . Suppose $c_1 = -1.3551$, the quantitative method is show in Table 3, $B'_t, 1 \leq t \leq 6$, represent the corresponding values after being quantified.
- Step 3 By using the steps 1–2 above, the compressed signal $C'_i = \{c_t, 1 \leq t \leq 2M\}$ can be embedded into $S3_i$, and signal after embedding is denoted by $W3_i$.
 Concatenate $W1_i, W2_i$ and $W3_i$, then inverse scrambling is performed to obtain the watermarked signal of the i -th frame. The embedding process is shown in Fig. 8.

3.2.3 Content authentication and tamper recovery

Denote W as the watermarked speech signal.

- (1) Cut W into P frames, and the samples of each frame are scrambled. The i -th frame after being scrambled is denoted by W_i . $W1_i, W2_i$ and $W3_i$ are the 3 segments of W_i . The length of $W1_i$ and $W2_i$ is N_1 , the length of $W3_i$ is N_2 .
- (2) Extract the sequence of integers from $W1_i = \{w1_1, w1_2, \dots, w1_{N_1}\}$ and $W2_i = \{w2_1, w2_2, \dots, w2_{N_1}\}$, denoted by $Y_i^1 = \{y_n^1, y_{n-1}^1, \dots, y_1^1\}$ and $Y_i^2 = \{y_n^2, y_{n-1}^2, \dots, y_1^2\}$, respectively.

- Step 1 According to the embedding method in Section 3.2.1, the integers in Y_i^1 and Y_i^2 can be calculated by using Eq. (5).
- Step 2 Reconstruct the frame number based on Y_i^1 and Y_i^2 , denoted by $i^1 = y_n^1 \times 10^{n-1} + y_{n-1}^1 \times 10^{n-2} + \dots + y_1^1$ and $i^2 = y_n^2 \times 10^{n-1} + y_{n-1}^2 \times 10^{n-2} + \dots + y_1^2$, which are the frame number extracted.

Table 6 The SDG and ODG values of different type watermarked signals

Watermarked signal	SDG	ODG
Type 1	-0.4954	-0.617
Type 2	-0.4527	-0.641
Type 3	-0.5081	-0.583
Type 4	-0.4232	-0.735
Type 5	-0.5376	-0.684

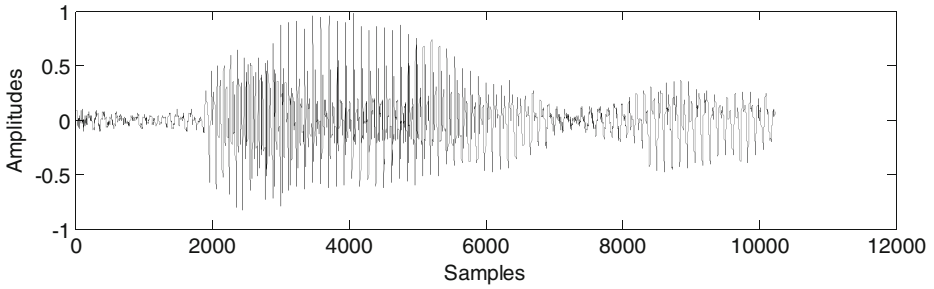


Fig. 14 Speech signal selected from the library randomly

(3) Content authentication

If $i^1 = i^2$, it indicates that the i -th frame is intact. Otherwise, it indicates that the i -th frame has been tampered. The process of content authentication is shown in Fig. 9.

(4)* Tamper recovery

Suppose that the frames of 1st to i -1th are all intact and the next N successive samples are subjected to malicious attack, the tamper recovery method is as follows.

Step 1 Move and authenticate the next N successive samples, until the samples can pass the authentication. Then reconstruct the frame number, denoted by i' . The signal between i -1th and i' th is regarded as the attacked frame. Tamper location process is shown in Fig. 10.

Step 2 According to the address index of the sorted chaotic sequence, find the frame, from which the compressed signal of attacked frame can be extracted. Let W'_i as the frame, and $CS'_i = \{cs_1, cs_2, \dots, cs_{2M}\}$ as compressed signal. The extraction method is described as follows.

- ① Cut W'_i into 3 segments, and the 3rd segment is denoted by $W_{3i'} = \{w_1, w_2, \dots, w_{N_2}\}$.
- ② Take the extraction of cs_1 as an example, which is extracted from the first 6 consecutive samples of $W_{3i'}$. Based on the block-based method (shown in Tables 2 and 3), the sign of cs_1 can be extracted from the values in B_1 , by using Eq. (6).

$$U = T' \bmod 2 \tag{6}$$

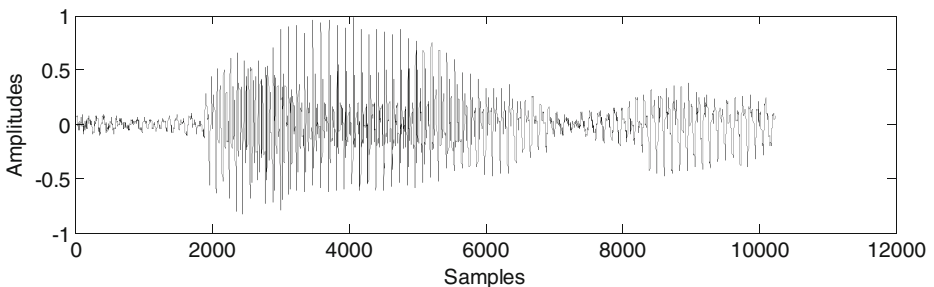


Fig. 15 Reconstructed signal based on the method in this paper

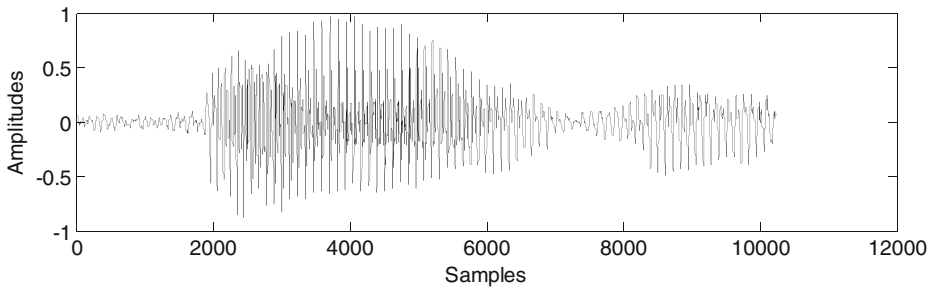


Fig. 16 Reconstructed signal based on the method in [13]

where T represents the sum of the values in B_1 . And $U=0$ indicates that the sign of cs_1 is “+”, $U=1$ indicates that the sign of cs_1 is “-”.

- ③ According to the values in B_2 , $cs_1^1 = \lfloor |cs_1| \rfloor$ can be calculated by using the Eq. (5). Similarly, $cs_1^2, cs_1^3, cs_1^4, cs_1^5$ can be extracted. Then cs_1 can be constructed by the Eq. (7).

$$cs_1 = \text{sign}(cs_1) \times \left(cs_1^1 + \frac{cs_1^2}{10} + \frac{cs_1^3}{100} + \frac{cs_1^4}{1000} + \frac{cs_1^5}{10000} \right) \quad (7)$$

- ④ By using the same method, the compressed signal $CS'_i = \{cs_1, cs_2, \dots, cs_{2M}\}$ can be extracted. Then the attacked frame can be reconstructed approximately using the method in Section 2.3

4 Performance analysis

4.1 Compression ratio

In this paper, speech is compressed firstly, which is as watermark and embedded. Embedding inevitably degrade the quality of speech signal. So, the signal as watermark should be compressed as possible. The compression ratio (CR) is defined by $CR = N_c/N_o$, where N_c

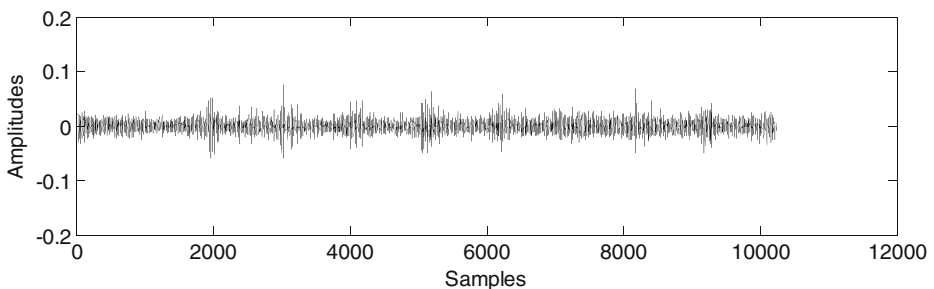


Fig. 17 The residual signal caused by the method in this paper

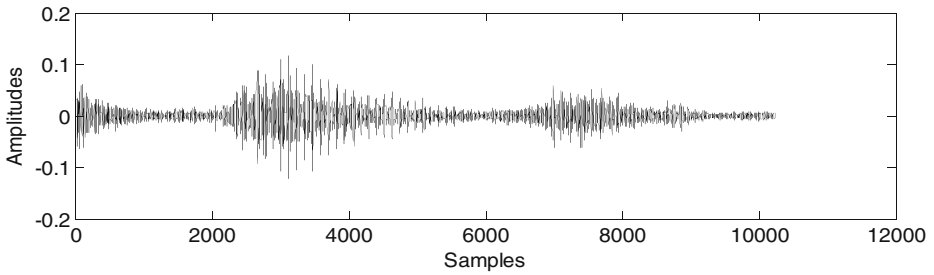


Fig. 18 The residual signal caused by the method in [13]

represents the length of compressed signal, and N_o represents the length of original one. For the compression method proposed, CR can be calculated by

$$CR = \frac{f'}{2^{D-1} \cdot f} \tag{8}$$

where D denotes the level of DWT, and the meaning of f and f' are same to that in section 2.2.1.

4.2 Ability of tamper recovery

Compressed signal is as watermark and embedded. If watermarked signal is attacked, the watermark embedded will be attacked too. So, the compressed signal of one frame should not be embedded into itself, but be embedded into other non-attacked frame. In this paper, the compressed signals are scrambled before embedding, in order to ensure that the compressed signal of the attacked frame can be extracted correctly from other frame. If compressed signal of one attacked frame can be extracted correctly, the attacked signal can be reconstructed. So, the tamper recovery ability is related to the performance of scrambling method.

Denote AC as the ability of tamper recovery, and it's defined by the Eq. (9).

$$AC = \frac{F_a}{F} \tag{9}$$

where F_a represents the number of frames can be recovered, and F represents frame number of watermarked signal. Supposed that, half of frames of watermarked signal are attacked, and the compressed signals of attacked frames are embedded into other half frames and all intact, the attacked frames can be recovered. The tamper recovery can reach the maximum capacity in

Table 7 SDG and ODG values of some reconstructed signals by different methods

Different types	SDG		ODG	
	Ref. [13]	Proposed	Ref. [13]	Proposed
Type 1	-1.0807	-0.9153	-1.175	-0.925
Type 2	-0.9663	-0.7542	-1.064	-0.917
Type 3	-1.0741	-0.8965	-1.081	-0.883
Type 4	-0.8934	-0.8067	-0.962	-0.824
Type 5	-0.9756	-0.8718	-1.153	-0.876

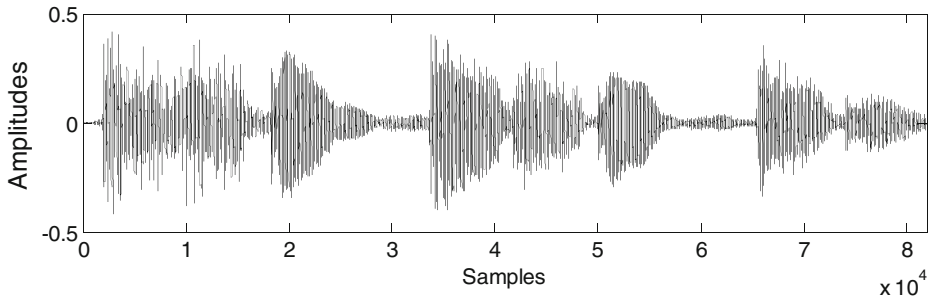


Fig. 19 Watermarked speech

this case. That is the maximum value of F_a is $F/2$. So, the maximum recovery ability of the scheme is $1/2$.

4.3 Security

For some watermark schemes, watermark embedded is based on public features. From the analysis in [12], we can get that the schemes are vulnerable subjected to substitution attack.

In this paper, samples of each frame are scrambled before embedding, and frame number and watermark are embedded into scrambled samples. After that, anti-scrambling is performed to get watermarked signal. So, watermarked samples are randomly throughout the whole frame, and the features used to embed watermark are secret for attackers. It's difficult to get the embedding position to perform attack without the key. If one frame is attacked, the attacked frame can be detected with high probability $1/2 \cdot 10^n$, where n is the number of integers in Y_i (generated by Eq. (4)). The ability of the scheme to resist attack is

$$R_A = 1 - 1/2 \cdot 10^n \quad (10)$$

4.4 Performance of tamper location

4.4.1 Shortcomings of the scheme based on synchronization codes

Desynchronization attacks can disrupt the location of watermark, and cause the watermark bits extracted incorrect. So, desynchronization attack is considered to be one of the most difficult attacks to resist [22]. Most of the schemes robust against

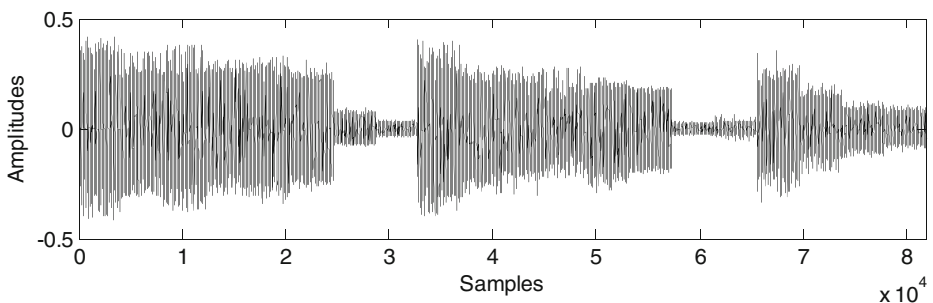


Fig. 20 Watermarked speech after each frame being scrambled

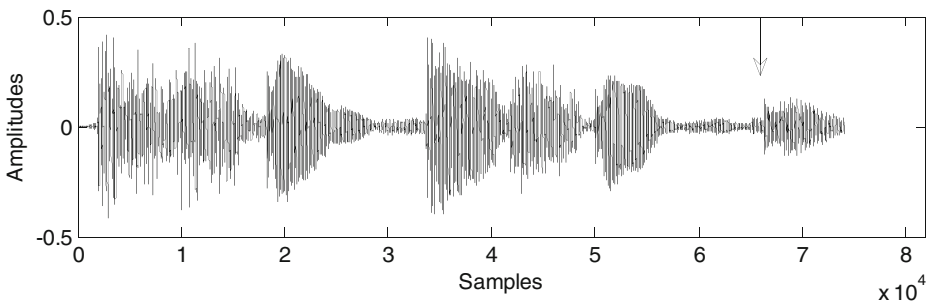


Fig. 21 Watermarked speech subjected to deletion attack

desynchronization attacks are based on synchronization code [8, 21, 22]. Based on the analysis in [12], we can get that the scheme based on synchronization code has some shortcomings. ①It can locate the watermarked signal only, but not locate the signal attacked. ②There are some segments, but not only one, from which the synchronization codes can be extracted, for short-time stationarity of speech signals. In order to confirm the conclusion, some test results are shown in the following.

Select one segment of speech signal randomly, and the length is 3000. Denote $SC = \{1010011101\}$ as the synchronization code. The test signal is divided into 3 frames, and synchronization code is embedded into 1st and 3rd frame, shown in Fig. 11.

- ① If the 2nd frame is subjected to attack, deletion, insertion or substitution, the attacked content cannot be detected by the two adjacent synchronization codes, extracted from the 1st and 3rd frame. As an example, Fig. 12 shows the test result for deletion attack.
- ② Search the synchronization code from the test signal, and the result is shown in Fig. 13. It is found that the synchronization code can be extracted correctly from the signals within rectangle. So, the method can locate the watermarked signal roughly not precisely.

4.4.2 The ability of precisely tamper location for the scheme proposed

The reason of the scheme tamper location roughly is that, one code is embedded by quantifying the feature generated by many samples. The more samples selected, the more tamper

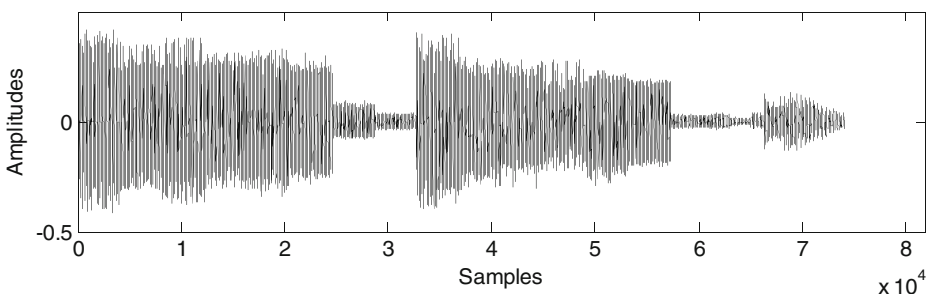


Fig. 22 Authentication result for intact frames

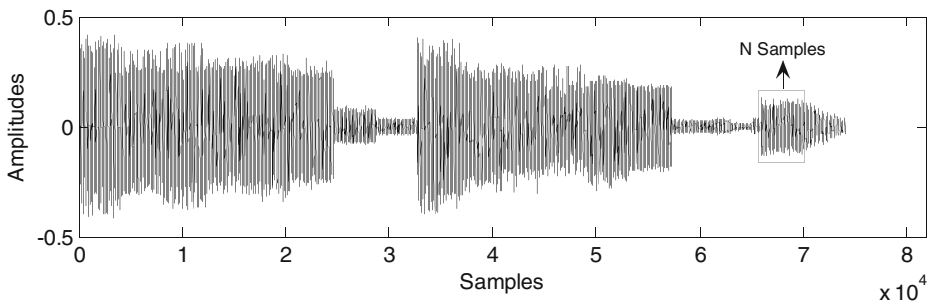


Fig. 23 Search result for next frame through authentication

location roughly is. To address the above problems, and in consideration of inaudibility and robustness, one code is embedded into a few samples (three samples) in this paper.

For the scheme, one frame is located and authenticated by the sequence of n integers. The probability of false tamper location for one frame is $1/2 \cdot 10^n$, approximate to zero with the increasing of n . So the method can tamper location precisely, and the probability of tamper location R_L is computed theoretically by Eq. (10).

Based on the analysis in this section, Table 4 gives the comparison of various abilities for some different schemes [8, 18, 21–23], containing the security and the ability of tamper location precisely, which are denoted by AI and AII, respectively. In Table 4, DYWT represents Dyadic Wavelet Transform. From the results shown in Table 4, it can be concluded that the scheme proposed has many advantages comparing with the schemes, [8, 18, 21–23].

5 Experimental results

We use MATLAB to simulate and analyze the performance of the recovery scheme proposed. And the computer used is powered by a 3.4-GHz Intel Core i7 processor and has 8GBs of RAM and an AMD R9 360 for graphics. 100 test signals are selected, contain 5 types. The number of each type signal is 20. Type 1 to type 4 are recorded by digital voice recorder, SONY PCM-D100, and the recording environments are quiet room, noisy room, open field and noisy station. Type 5 is selected for the library. The signals selected are WAVE format 16-bit quantified mono signals, sampled at 44.1 kHz. The parameters are set as follows, $L = 81920$, $P = 20$, $N = 4096$, $f = 44100$, $f_r = 11025$, $D = 3$, $N_1 = 12$, $N_2 = 4072$, $n = 4$, $k = 0.68$, $\mu = 3.9728$.

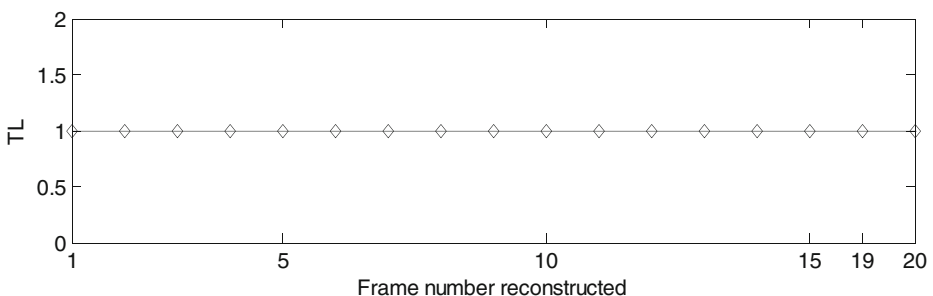


Fig. 24 Tamper location result of deletion attack

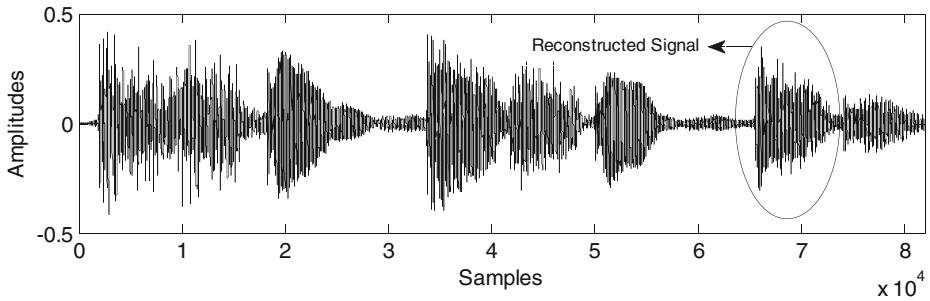


Fig. 25 Tamper recovery result for deletion attack

5.1 Inaudibility

In the paper, subjective difference grades (SDG) and objective difference grades (ODG) are used to test the inaudibility of the watermarked speech. The meaning of the scores in SDG and ODG are listed in Table 5.

The SDG and ODG values of five types watermarked signal are listed in Table 6, in which SDG values are obtained from 10 listeners, and ODG values are acquired by using the PEAQ system. Based on the test results, it can be seen that the watermarked signal is inaudibility.

5.2 Reconstruction quality

One part of the original signal (called approximate signal) can be generated based on the approximate coefficients from DWT, and another part of the original signal (called residual signal) can be obtained by DCT coefficients. And then, original signal are reconstructed approximately by summing the two part signals.

In the following, the difference of original speech and reconstructed one, caused by different ways, are tested and compared. Select one part of the speech signal, from the library randomly, shown in Fig. 14. Compress and reconstruct the signal by using the two different methods, proposed in this paper and in [13]. The reconstructed signals are shown in Figs. 15 and 16, with the same compression ratio. Then the two residual signals are shown in Figs. 17 and 18, respectively.

The energy of the residual signal generated by the method is calculated and compared with that generated by [13]. It's got that the energy the energy in this paper is less than that in [13]. For the signals in Figs. 17 and 18, the energy is 1.5607 and 2.4626, respectively. So, the

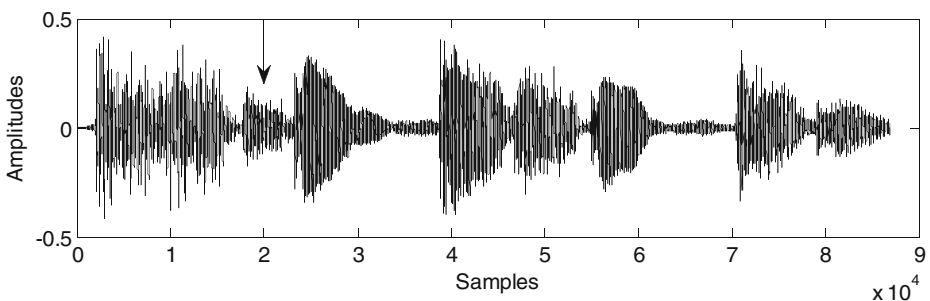


Fig. 26 Watermarked speech subjected to insertion attack

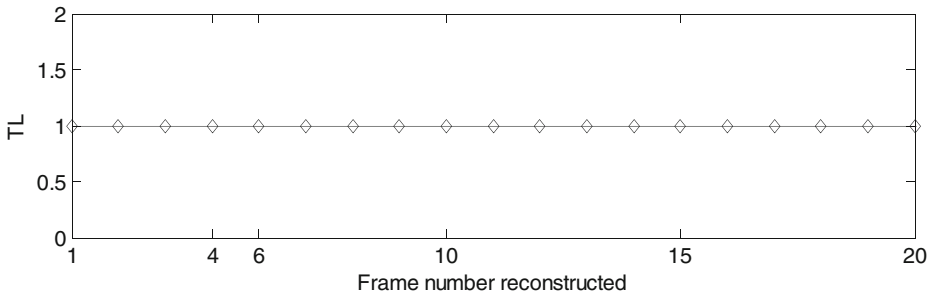


Fig. 27 Tamper location result of insertion attack

reconstruction method in this paper can result in less energy loss than the method in [13]. Table 7 lists the SDG and ODG values of some reconstructed signals, by using different methods. It can be concluded that the reconstructed signals by using the method in this paper have better perceptual quality than that in [13], with the same compression ratio. Therefore the compression and reconstruction method is more effective.

5.3 Tamper Recovery

For all attack channels can be viewed as deletion, insertion and substitution channel for watermarking [23]. In this section, one watermarked speech signal is selected from the library randomly shown in Fig. 19, which will be subjected to the 3 types attack. And Fig. 20 shows the watermarked signal after being scrambled for each frame. Then the corresponding tamper location and tamper recovery results are given.

For deletion attack, the detailed steps are given, aiming to make the method of tamper location and recovery explicitly. While for other attacks, the results of tamper location and tamper recovery are given only.

5.3.1 Tamper recovery for deletion attack

The watermarked speech shown in Fig. 19 is subjected to deletion attack, and the detailed tamper recovery steps are shown as follows.

Step 1 The samples 65001th to 72800th are selected and deleted, shown in Fig. 21.

Step 2 For the attacked signal, each frame is scrambled and authenticated from the first frame using the method proposed, until that the N successive samples cannot

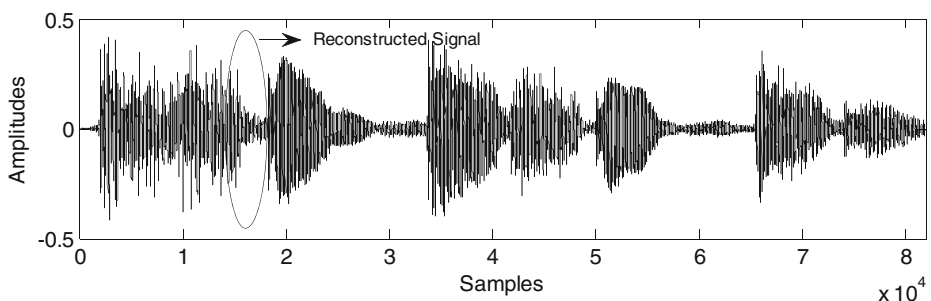


Fig. 28 Tamper recovery result for insertion attack

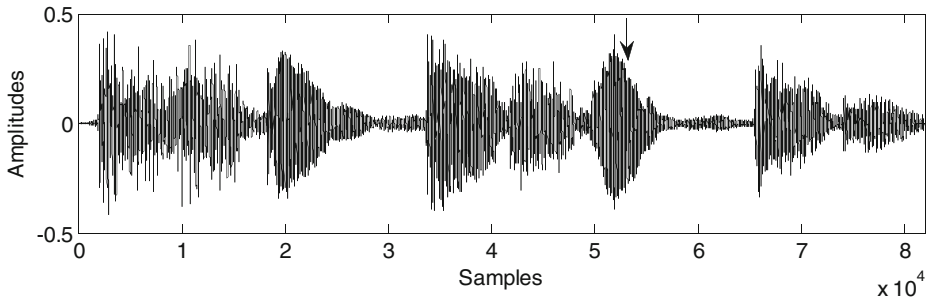


Fig. 29 Watermark speech subjected to substitution attack

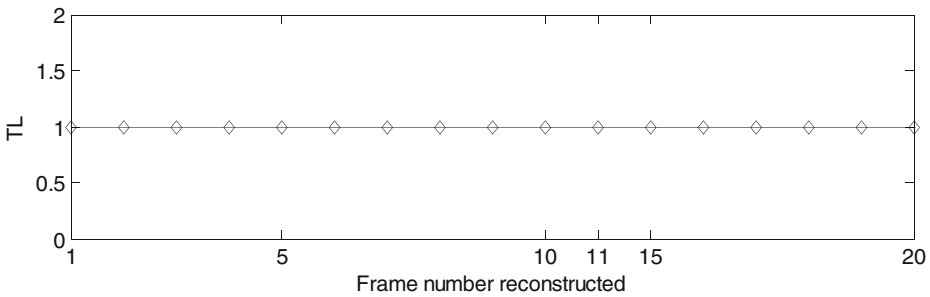


Fig. 30 Tamper location result of substitution attack

pass authentication. The authentication result for intact frames is shown in Fig. 22.

- Step 3 Move and scramble next N successive samples, until the samples can pass authentication successfully. The result is shown in Fig. 23. Then extract and reconstruct the frame number.
- Step 4 The frame number reconstructed, shown in in Fig. 24, is as the tamper location result, in which $TL = 1$ indicates that the corresponding frame is intact. Based on the result, the 16th to 18th frame number can't be reconstructed. So, the 16th to 18th frame of watermarked signal is regarded to be attacked.
- Step 5 Compressed signals of attacked frame are embedded into the 19th, 2nd and 6th frame. Extract the compressed signals from the frames, and reconstruct attacked signals to perform tamper recovery. The tamper recovery result is shown in Fig. 25.

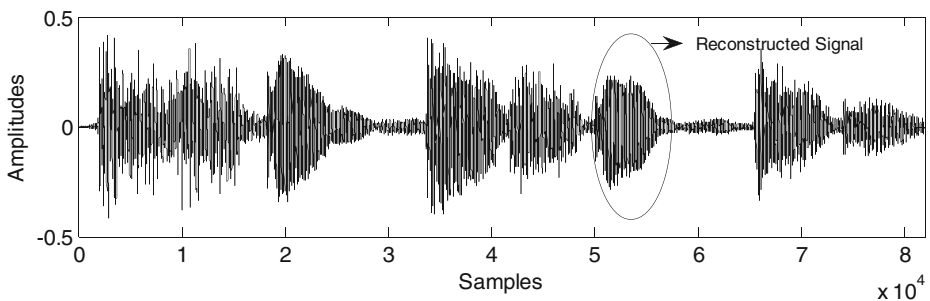


Fig. 31 Tamper recovery result for substitution attack

5.3.2 Tamper recovery for insertion attack

Insert 4000 samples after the 17,000-th sample of watermarked speech. The attacked signal is shown in Fig. 26. And the tamper location result is shown in Fig. 27, from which it can be seen that the 5th frame is attacked. Extract the compressed signal from 1st frame of watermarked speech, and reconstruct the attacked signal to perform tamper recovery. The result is shown in Fig. 28.

5.3.3 Tamper recovery for substitution attack

The samples of 48001th to 55000th of watermarked signal are substituted by using other samples. The attacked signal is shown in Fig. 29, and the tamper location result is shown in Fig. 30, from which it can be seen that the 12th, 13th and 14th frame are attacked. Then extract the compressed signals from 8th, 17th and 4th frame to reconstruct the attacked signals, and the tamper recovery result is shown in Fig. 31.

From the tamper localization and recovery results, it can be concluded that the scheme proposed can locate the attacked frames precisely, and has a good ability of tamper recovery.

6 Conclusion

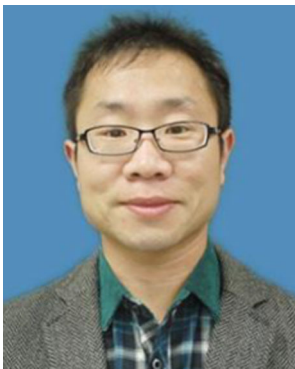
In order to increase the credibility of digital speech signal, an authentication and tamper recovery scheme is proposed. The compression and reconstruction method for speech signal based on DWT and DCT is given. Firstly, speech signal is framed and segmented, and frame number is embedded into the first two segments. Secondly, each frame is compressed, and all the compressed signals are scrambled, which are as watermark and embedded based on block-based method. For the scheme, attacked frames can be located precisely by frame number extracted. And the compressed signal of attacked frame is extracted to perform tamper recovery. Experimental results show that the authentication and recovery scheme is effective.

Acknowledgments This paper is supported by the National Natural Science Foundation of China (Grant No. 61332012, 61272465, 61502409), Shenzhen R&D Program (GJHZ20140418191518323), and Nanhu Scholars Program for Young Scholars of XYNU. We would like to thank the anonymous reviewers for their constructive suggestions.

References

1. Akhaee MA, Kalantari NK, Marvasti F (2010) Robust audio and speech watermarking using Gaussian and Laplacian modeling. *Signal Process* 90(8):2487–2497
2. Chamlawi R, Khan A, Usman I (2010) Authentication and recovery of images using multiple watermarks. *Comput Electr Eng* 36(3):578–584
3. Chen OTC, Liu CH (2007) Content-dependent watermarking scheme in compressed speech with identifying manner and location of attacks. *IEEE Trans Audio, Speech, Language Process* 15(5):1605–1616
4. Fakhr MW (2012) Sparse watermark embedding and recovery using compressed sensing framework for audio signals. *Int Conf Cyber-Enabled Distrib Comput Knowledge Discover* 535–539
5. Herbig T, Gerl F, Minker W (2012) Self-learning speaker identification for enhanced speech recognition. *Comput Speech Language* 26(3):210–227
6. Khan LA, Baig MS, Youssef AM (2010) Speaker recognition from encrypted VoIP communications. *Digit Investig* 7(1–2):65–73
7. Lee TY, Lin SD (2008) Dual watermark for image tamper detection and recovery. *Pattern Recogn* 41(11):3497–3506

8. Lei BY, Soon IY, Li Z (2011) Blind and robust audio watermarking scheme based on SVD-DCT. *Signal Process* 91(8):1973–1984
9. Lei B, Soon IY, Tan EL (2013) Robust SVD-Based audio watermarking scheme with differential evolution optimization. *IEEE Trans Audio, Speech, Language Process* 21(11):2368–2378
10. Li B, Wang M, Li XL, Tan SQ, Huang JW (2015) A strategy of clustering modification directions in spatial image steganography. *IEEE Trans Inform Foren Sec* 10(9):1905–1917
11. Li CL, Wang YH, Ma B, Zhang ZX (2012) Tamper detection and self-recovery of biometric images using salient region-based authentication watermarking scheme. *Comput Standards Interf* 34(4):367–379
12. Liu ZH, Wang HX (2014) A novel speech content authentication algorithm based on Bessel-Fourier moments. *Digit Sign Process* 24(1):197–208
13. Liu ZH, Zhang F, Wang J, Wang HX, Huang JW (2016) Authentication and recovery algorithm for speech signal based on digital watermarking. *Signal Process* 123(1):157–166
14. Luo D, Sun MM, Huang JW (2016) Audio postprocessing detection based on amplitude cooccurrence vector feature. *IEEE Sign Process Lett* 23(5):688–692
15. Navarathna R, Dean D, Sridharan S (2013) Multiple cameras for audio-visual speech recognition in an automotive environment. *Comput Speech Language* 27(4):911–927
16. Park CM, Thapa D, Wang GN (2007) Speech authentication system using digital watermarking and pattern recovery. *Pattern Recogn Lett* 28(8):931–938
17. Peng H, Li B, Luo XH (2013) A learning-based audio watermarking scheme using kernel Fisher discriminant analysis. *Digit Sign Process* 23(1):382–389
18. Pun CM, Yuan XC (2013) Robust segments detector for de-synchronization resilient audio watermarking. *IEEE Trans Audio, Speech, Language Process* 21(11):2412–2424
19. Roldan LR, Hernandez MC, Miyatake MN, Meana HP, Kurkoski B (2013) Watermarking-based image authentication with recovery capability using halftoning technique. *Signal Process Image Commun* 28(1):69–83
20. Sahidullah M, Saha G (2012) Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Comm* 54(4):543–565
21. Vivekananda BK, Indranil S, Abhijit D (2011) A new audio watermarking scheme based on singular value decomposition and quantization. *Circ, Syst, Sign Process* 30(5):915–927
22. Wang XY, Ma TX, Niu PP (2011) A pseudo-Zernike moments based audio watermarking scheme robust against desynchronization attacks. *Comput Electr Eng* 37(4):425–443
23. Wang Y, Wu SQ, Huang JW (2010) Audio watermarking scheme robust against desynchronization based on the dyadic wavelet transform. *J Adv Sign Process* 2010(13):1–17
24. Yuan S, Huss SA (2004) Audio watermarking algorithm for real-time speech integrity and authentication. *The 2004 workshop on Multimedia and Security*, 220–226



Zhenghui Liu was born in 1983. He received the B.S. degree from Luoyang Normal University, Luoyang, in 2005; M.S. degree from Xinyang Normal University, Xinyang, in 2010, and Ph.D. degrees from the School of Information Science and Technology, Southwest Jialong University, Chengdu, in 2014. He currently works as the post doctor with the College of Information Engineering, Shenzhen University, Shenzhen, China. His current research interests include multimedia information security and audio forensics.



Da Luo received the B.S. degree in Computer Science and Technology and Ph.D. degree in Computer Application and Technology from the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China, in 2008 and 2014. He currently works as the post doctor with the College of Information Engineering, Shenzhen University, Shenzhen, China. His current research interests include audio forensics and deep learning.



Jiwu Huang received the B.S. degree from Xidian University, Xi'an, China, the M.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Science, Beijing, in 1982, 1987, and 1998, respectively. He is currently a Professor with the College of Information Engineering, Shenzhen University, Shenzhen, China. Before joining Shenzhen University, he was with the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China. His current research interests include multimedia forensics and security. He serves as a member of the IEEE CASS Multimedia Systems and Applications Technical Committee and the IEEE SPS Information Forensics and Security Technical Committee. He was an Associate Editor of the IEEE Transactions on Information Forensics and Security from 2010 to 2014.



Jing Wang was born in 1992. He received the B.S. degree from Xinyang Normal University Huarui College, in 2014. She is currently pursuing the M.S. degree in Xinyang Normal University. Her current research interests include multimedia information security and digital watermark.



Chuanda Qi was born in 1965. He received the B.S. degree from Xinyang Normal University, Xinyang, in 1985, and the Ph.D. from PLA Information Engineering University, Zhengzhou, in 2007. His current research interests include cipher theory and information security.