A Robust Speech Content Authentication Algorithm Against Desynchronization Attacks

Bin Han and Enjie Gou
College of Information Security, Chengdu University of Information Technology, Chengdu, 610225, China
Email: binhan_cn@163.com; enjackgo@163.com

Abstract —A robust speech content authentication algorithm against desynchronization attacks is proposed, in order to authenticate the content of digital speech signal. Firstly, the signal is framed, and each frame is divided into two parts. The frame number of each frame, as the watermark bit, is embedded into the first and second part by quantizing Bessel-Fourier moments of the correspond signals. The content of each frame is authenticated by comparing with the watermark bits extracted from the first and second part. Experiment results shown that the authentication scheme against desynchronization attacks proposed is effective. And comparing with some existing schemes, the performance of scheme is improved.

Index Terms—Speech watermarking, tamper detection, content authentication, bessel-fourier moments

I. Introduction

Nowadays, the authenticity of digital speech signal is questioned for the different types of attack. So, it's necessary to authenticate the speech signal obtained. Digital audio watermarking provides a method to authenticate the truth and integrity of digital speech signals.

For digital audio watermark, there are abundant research results, which can be divided into two types according to the purpose of application. The first type is used for protecting the audio copyright [1]-[4]. And the second is used for authentication the content of audio signal [5], [6].

In these two aspects, protecting the copyright and identifying the contents true or not, the great demand is to authenticate the integrity for speech signal. In [7], authors proposed a speech watermarking scheme, which can ensure the integrity of compressed speech data. For the algorithm, watermark is generated according to the line spectrum frequency feature, and embedded in each frame by quantizing the least significant bits (LSB). The embedding methods based on LSB are very fragile to signal processing. For these schemes, the common signal processing motivated by the needs of application will be considered to be malicious attacks. In [8], authors proposed a watermarking scheme using singular value decomposition and quantization index modulation, which is robust against desynchronization attack by using synchronization code [9]. For the synchronization code embedded in each frame is the same, if exchanging the content of two frames, the synchronization codes extracted from the two exchanged frames will be equal. So, the attack cannot be detected. That is, for the schemes [8], [9] using synchronization code to resist desynchronization attack are insecurity.

According to the description above, there are some shortcomings for speech content authentication algorithm currently. 1) The ability against signal processing is weak, such as the embedding method based on LSB [7]. 2) For these schemes based on synchronization code, they are insecurity. 3) The algorithm based on synchronization code can locate the watermarked signal only, but not detect the frame number attacked.

Considering the existing shortcomings of speech content authentication algorithm, a robust speech content authentication algorithm against desynchronization attacks is proposed. In the paper, the watermark bits are generated by the frame number. For different frames, the watermark bits are different. If two frames are exchanged, the different watermark bits are exchanged too. So, the attack can be detected. The watermark bits are embedded by quantizing the Bessel-Fourier moments, and the method is robust [10], [11]. So, our scheme is more robust than that based on LSB. Experimental results show that the algorithm has the ability of tamper detection and localization. Comparison results with previous works show its superior.

The details of the scheme will be addressed in the following sections. Section II and III illustrate the detail procedure of the scheme, containing watermark generation and embedding, and watermark extraction and content authentication. The performance of the scheme and the experimental results are shown in Section IV, which demonstrate that the scheme is effective. Finally, we summarize the conclusion in Section V.

II. Robust Watermark Generation and Embedding

A. Preprocessing

Denote $S=\{S(t),1\le t\le L\}$ as the original speech signal, and $S(t)$ is the *t*-th sample.

1) S is divided into *P* non-overlapping frames, and the length of each frame is *N*, $N=L/P$. The *i*-th frame is denoted by S_i, $1\le i\le P$.

Manuscript received May 5, 2014; revised September 17, 2014.
Corresponding author email: binhan_cn@163.com.
doi:10.12720/jcm.9.9.723-728

2) S_i is divided into two parts, and denoted by S_i^1 and S_i^2. Both of S_i^1 and S_i^2 are divided into 6*M* segments, and *M* is a positive integer. The *j*-th segments are denoted by $S_{i,j}^1$ and $S_{i,j}^2$, respectively, $1 \le j \le 6M$. That is, the *i*-th frame is divided into 12*M* segments. The length of $S_{i,j}^1$ and $S_{i,j}^2$ are equal to M_1.

B. Watermark Generation and Embedding

1) In this paper, we use frame number as the watermark embedded. For the *i*-th frame, the frame number *i* is converted into binary bits, and denoted by $W_i = \{w_{i,j} | w_{i,j} \in (0,1), 1 \le j \le M\}$. When the length of W_i is less than *M*, 0 is added.

2) Calculate the total amplitude of all Bessel-Fourier moments (BFMs) of $S_{i,j}^1$, with a given order *n*, using the following steps.

Step 1: The 1D signal $S_{i,j}^1$ is mapped into 2D form by using the projection as follow

$$\begin{cases} M_1 = K \times K + h, \ 0 \le h < 2K+1 \\ S_{i,j}^1(u,v) = S((u-1)\cdot K + v), \ 1 \le u,v \le K \end{cases} \quad (1)$$

where $S_{i,j}^1(u,v)$ is the 2D signal after projection, *h* is the number of the rest samples, *K* is the width or height in $S_{i,j}^1(u,v)$, the value of which is as large as possible under the constrain of the Eq. (1), and M_1 is the length of the segment $S_{i,j}^1$.

Step 2: Denote $S_{i,j}^1(\rho,\theta)$ as the signal $S_{i,j}^1(u,v)$ in polar coordinates. The total amplitude of all BFMs with the given order *n* can be calculated by using

$$F_{i,j}^1 = \sum_{m=-n}^{n} |B_{nm}| \quad (2)$$

where B_{nm} is the BFMs of order *n* with repetition *m* [10,11]

(3) Watermark embedding

In this paper, one watermark bit is embedded into 6 successive segments, and the method is described in the following.

Step 1: Calculate the average value of $F_{i,6\times i-5}^1$ and $F_{i,6\times i-4}^1$ using the Eq. (3). The average value of $F_{i,6\times i-3}^1$ and $F_{i,6\times i-2}^1$, $F_{i,6\times i-1}^1$ and $F_{i,6\times i}^1$ are calculated in the same way, and the results are denoted by E_2 and E_3, respectively.

$$E_1 = \frac{F_{i,6\times i-5}^1 + F_{i,6\times i-4}^1}{2} \quad (3)$$

Step 2: Denote $E_{\max} = \mathrm{Max}(E_1, E_2, E_3)$, $E_{\mathrm{med}} = \mathrm{Medium}(E_1, E_2, E_3)$, $E_{\min} = \mathrm{Min}(E_1, E_2, E_3)$, and calculate the value *U* and *V* using

$$\begin{cases} U = E_{\max} - E_{\mathrm{med}} \\ V = E_{\mathrm{med}} - E_{\min} \end{cases} \quad (4)$$

Step 3: The watermark bit $w_{i,j} = 1$ is embedded by the expression

$$U - V \ge T_v \quad (5)$$

If $U - V < T_v$, increase the value of $E_{\max}$ and $E_{\min}$, and reduce the value E_{med} when necessary, under the restricted condition $E_{\mathrm{med}} \ge E_{\min}$, to make the condition shown in Eq. (5) is satisfied .

The watermark bit $w_{i,j} = 0$ is embedded by the expression

$$V - U > T_v \quad (6)$$

If $V - U \le T_v$, reduce the value of $E_{\max}$ and $E_{\min}$, and increase the value E_{med} when necessary, under the restricted condition $E_{\max} \ge E_{\mathrm{med}}$, to make the condition shown in Eq. (6) is satisfied .

In Eq. (5) and Eq. (6), $T_v = d \cdot (E_1 + E_2 + E_3)/3$ is the T_v watermark embedding strength, and as the intensity factor.

Step 4: The average value of E_1, E_2 and E_3 after embedding one watermark bit is denoted by E_1', E_2' and E_3', respectively. It is equivalent to scale E_1, E_2 and E_3 by using the corresponding factor α_1, α_2 and α_3, which can be calculated by using the following expressions.

$$\alpha_1 = \frac{E_1'}{E_1}, \alpha_2 = \frac{E_2'}{E_2}, \alpha_3 = \frac{E_3'}{E_3} \quad (7)$$

Step 5: Magnify all the samples of $F_{i,6\times i-5}^1$ and $F_{i,6\times i-4}^1$ using the factor α_1. Similarly, magnify the samples of $F_{i,6\times i-3}^1$ and $F_{i,6\times i-2}^1$, $F_{i,6\times i-1}^1$ and $F_{i,6\times i}^1$ using the factor α_2 and α_3, respectively. Then, the signal obtained is the watermarked signal.

For the segments $S_{i,j}^2$, $1 \le j \le 3M$, watermark bit $w_{i,j}$ is embedded by using the same method described above.

The process of watermark generation and embedding is shown in Fig. 1.

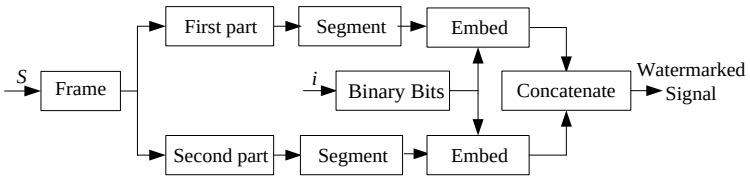


Fig. 1. The process of watermark generation and embedding

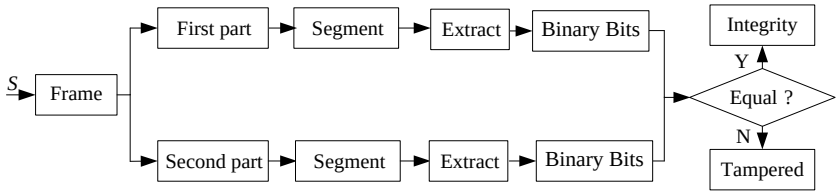


Fig. 2. The process of watermark extraction and content authentication

III. Watermark Extraction and Content Authentication

Suppose that the watermarked signal is denoted by *WS*.

1) Using the preprocessing method shown in section II, we can get P frames, which are denoted by WS_i, $1 \le i \le P$. And WS_i is divided into two parts, denoted by WS_i^1 and WS_i^2. WS_i^1 is divided into $6M$ segments, and the j-th segments are denoted by $WS_{i,j}^1$. Using the same method we can get other $6M$ segments of WS_i^2, and denoted by $WS_{i,j}^2$.

2) $WS_{i,j}^1$ and $WS_{i,j}^2$ is mapped into 2D form using the Eq. (1), and the total amplitude of BFMs are calculated by the Eq. (2), which is denoted by $WF_{i,j}^1$ and $WF_{i,j}^2$, $1 \le j \le 6M$.

3) Calculate the average value of $WF_{i,6\times i-5}^1$ and $WF_{i,6\times i-4}^1$ using the Eq. (3). For $F_{i,6\times i-3}^1$ and $F_{i,6\times i-2}^1$, $F_{i,6\times i-1}^1$ and $F_{i,6\times i}^1$, we can get the results WE_2 and WE_3.

4) Denote $WE_{\max} = \text{Max}\left(WE_1, WE_2, WE_3\right)$, $WE_{\text{med}} = \text{Medium}\left(WE_1, WE_2, WE_3\right)$, $WE_{\min} = \text{Min}(WE_1, WE_2, WE_3)$, and calculate the *WU* and *WV* using

$$\begin{cases} WU = WE_{\max} - WE_{\text{med}} \\ WV = WE_{\text{med}} - WE_{\min} \end{cases} \tag{8}$$

Comparing *WU* and *WV*, we can get the watermark bit embedded by using the following rule

$$w_{i,j}^1 = \begin{cases} 1 & \text{if } WU - WV \ge 0 \\ 0 & \text{if } WU - WV < 0 \end{cases} \tag{9}$$

Using the same method the watermark bit embedded in $WS_{i,j}^2$ can be extracted, which is denoted by $w_{i,j}^2$.

5) If $\sum_{j=1}^{M} w_{i,j}^1 \oplus w_{i,j}^2 = 0$, it indicates that the *i*-th frame is intact. If $\sum_{j=1}^{M} w_{i,j}^1 \oplus w_{i,j}^2 \ne 0$, it indicates that the *i*-th frame has been tampered.

The process of watermark extraction and content authentication is shown in Fig. 2.

IV. Experimental Results

In the following, the comprehensive performance of our scheme will be analyzed and tested. The test speech signals are selected from speech sample library, and they are all 16-bit quantified mono speech signal sampled at 44.1 kHz, wave format. The length $L = 840000$, the frame number $N = 100$, the intensity factor $d = 0.18$, the order of BFMs $n = 13$.

A. Embedding Capacity

In this paper, we define the embedding capacity is the number of watermark bits embedded in per second. Let denote V_w as the embedding capacity, $f_s(\text{Hz})$ as the sampling rate of the original speech, and *N* as the length of each frame. The number of frames in per second is f_s/N. Then V_w can be calculated by the Eq. (10). One watermark bit is embedded in 6 successive segments, while in [12], one watermark bit is embedded in 9 successive segments. That is the embedding capacity of the scheme proposed in [12] is $2V_w/3$. So, Comparing with the scheme [12], the embedding capacity is increased in this paper.

$$V_w = \frac{2M \cdot f_s}{N}(bit/s) \tag{10}$$

B. Method Robust Against Malicious Tampers

1) Substitution attack

If the watermarked signal is subjected to substitution attack, the synchronization of the watermark is not disrupted. So, the location of the watermarked signal is same to that before attacking. For the segments of attacked frame, the watermark bits ($w_{i,j}^1$, $1 \le j \le M$) extracted from the first part WS_i^1 are different to that ($w_{i,j}^2$, $1 \le j \le M$) extracted from the second part WS_i^2. So, the attack can be detected.

2) Desynchronization attack

Desynchronization attack can desynchronize the location of watermark, which cause the location of samples move forward or backward. So, it's necessary for the watermarked signal subjected to desynchronization attack to resynchronize. In this paper, the method of content authentication is processed by comparing the watermark bits extracted from the first and the second part of one frame. The samples are same to that before attack, apart from the location.

In this section, take deletion attack as one example to explain the method against desynchronization attack, and for other attacks, the methods are same. For the scheme proposed, if the watermarked signal is attacked, there will be a frame, denoted by the *i*-th frame, from which the watermark bits extracted are satisfied $\sum_{j=1}^{M} w_{i,j}^1 \oplus w_{i,j}^2 \ne 0$.

Then, move and authenticate the next N successive samples, until to find the N successive samples be authenticated successfully. Then extract the watermark bits from the first part of the N successive samples and convert into an integer, denoted by i'. The difference between i' and $i-1$ is the (frame number) content be deleted. So, the scheme is robust against desynchronization attack. The method robust against desynchronization attack is shown in Fig. 3.

The schemes robust against desynchronization attack proposed in [8], [9] are based on synchronization code. For these schemes, they can locate the watermarked signal only, but not detect the frame number attacked. Based on the scheme proposed, the attacked frame number (e.g. delete one frame) can be obtained (the difference between i' and $i-1$). So, the scheme proposed in this paper is more effective than that proposed in [8], [9].

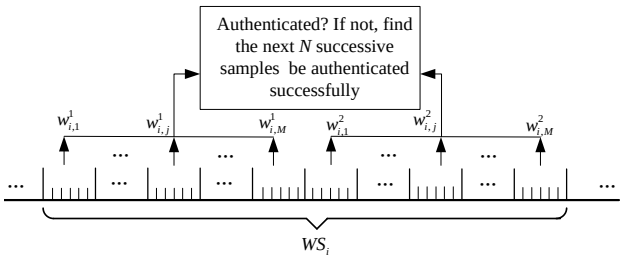


Fig. 3. The method robust against desynchronization attack

C. Inaudibility Test

In the test, different type voice was selected, containing male, female voice. And the performance is measured by subjective difference grades (SDG) and objective difference grades (ODG) [9]. For different type voice, Table I lists the ODG and SDG values, and the SDG was acquired form 4 listeners. Form the test results we get that our scheme is inaudibility.

TABLE I: ODG AND SDG VALUE OF DIFFERENT SPEECH SIGNALS

Voice type	SDG	ODG
Male voice	0	-0.193
Female voice	0	-0.317

TABLE II THE BER VALUES UNDER VARIOUS COMMON SIGNAL PROCESSING OPERATIONS

Signal processing	BER	
	Ref. [3]	Our scheme
Resampling (22.05 kHz)	0.0009	0
Resampling (11.025 kHz)	0.0009	0
Resampling (8 kHz)	0.0029	0
Low-pass filtering (11.025 kHz)	–	0
Low-pass filtering (8 kHz)	0	0
Low-pass filtering (4 kHz)	0	0.0037
MP3 (56 Kbits)	0	0.0448
MP3 (48 Kbits)	0.4578	0.0694
MP3 (40 Kbits)	0.4717	0.0915

D. Robustness Test

In this section, the bit error rate (BER) is used to measure the robustness of our scheme, which is defined

$$BER = \frac{R}{2 \cdot M \cdot P} \times 100\% \tag{11}$$

where $2 \cdot M \cdot P$ is the total number of watermark bits embedded, and R is the number of watermark bits erroneously detected.

In this part, we selected one speech signal randomly, male voice, and calculate the value BER after being subjected to some signal processing operations. The results are shown in Table II, and compared with the schemes [3]. Based on the results, it can be seen that the embedding method is more robust than that [3] for some signal processing.

E. Tamper Location Test

In the following, the tamper detection for the watermarked signal subjected to different attacks is tested, containing substitution attack, deletion attack and insertion attack. Fig. 4 shows the watermarked signal. The results are shown in Fig. 5 to Fig. 7, in which p represents the frame number of speech signal, $L(p)=0$ represents the corresponding frame (p-th frame) is intact, and $L(p)=1$ represents the corresponding frame is tampered.

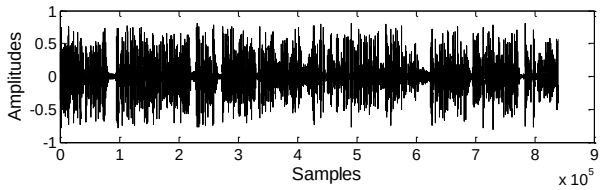
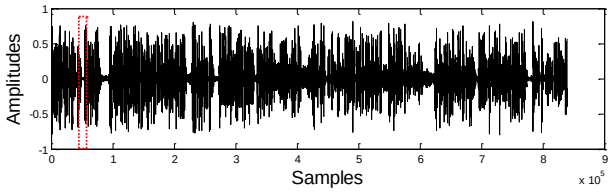
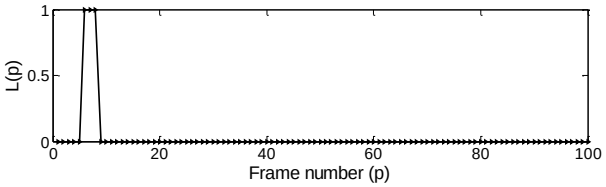


Fig. 4. The watermarked signal



(a) The watermarked signal subjected to substitution attack



(b) The tamper location result for the substitution attacked

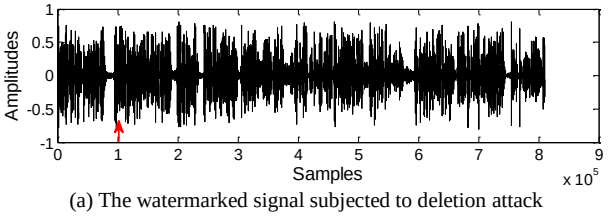
Fig. 5. The substitution attacked signal and the corresponding tamper location result

1) Substitution attack

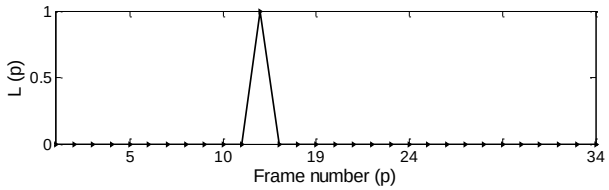
Select one segment signal to substitute the watermarked signal. Assume that the samples of the watermarked signal from 50000-th to 60000-th are subjected to substitution attacked. The attacked signal is shown in Fig. 5 (a), and the tamper location result is shown in Fig. 5 (b). From the tamper location result, it can be seen that the content of frames between 6-th and 8-th are that attacked.

2) Deletion attack

Delete the 100000-th to 130000-th samples of the watermarked signal, which is shown in Fig. 6 (a). For the attacked signal shown in Fig. 6 (a), the tamper location result is shown in Fig. 6 (b). In order to more clearly to show the frames attacked, in Fig. 6(b), the frame number 1 to 34 are shown only, and other frames are all intact. From the tamper location result, we can get that the 12-th to the 16-th frames are that deleted, which is the frames subjected to deletion attack.

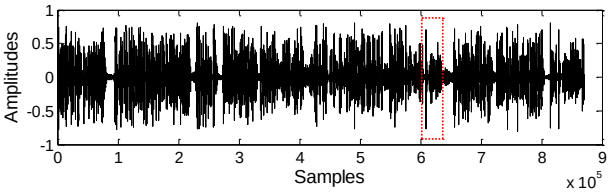


(a) The watermarked signal subjected to deletion attack

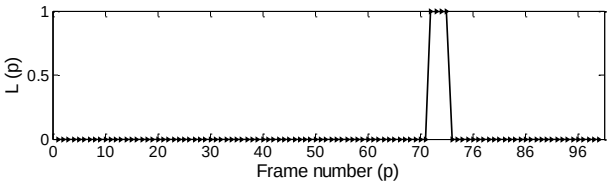


(b) The tamper location result for the deletion attacked

Fig. 6. The deletion attacked signal and the corresponding tamper location result



(a) The watermarked signal subjected to insertion attack



(b) The tamper location result for the insertion attacked

Fig. 7. The insertion attacked signal and the corresponding tamper location result

3) Insertion attack

For insertion attack, select 30000 speech samples and insert to the 600000-th sample of the watermarked signal, which is shown in Fig. 7 (a). For the attacked signal, the tamper location result is shown in Fig. 7 (b). It can be seen that the content of frames between 72-th and 75-th can not through authentication.

Based on the test results shown in Fig. 5 to Fig. 7, it can be concluded that the location attacked is uniform to the tamper location results. So, our scheme proposed in this paper has a good ability of tamper location, and is robust against desynchronization attacks.

V. Conclusions

Considering the shortcomings of speech content authentication algorithm existed, a speech content authentication algorithm robust against desynchronization attacks is proposed. The signal is framed, and each frame is divided into two parts. Then watermark bits generated by the frame number are embedded into the first and the second parts by quantizing the Bessel-Fourier moments of the corresponding signal. For the exchanged frames, they can be detected by frame number converted by the watermark bits extracted. So, comparing with schemes based on synchronization code, the synchronization performance of our scheme is more effective and security. Experiment results shown that the scheme proposed has the ability of tamper location and is robust against signal procession operations, which demonstrate that the scheme is effective.

References

[1] C. M. Pun and X. C. Yuan, "Robust segments detector for de-synchronization resilient audio watermarking," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2412-2424, November 2013.

[2] B. Lei, I. Y. Soon, and E. L. Tan, "Robust SVD-Based audio watermarking scheme with differential evolution optimization," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2368-2378, November 2013.

[3] X. Y. Wang, P. P. Niu, and H. Y. Yang, "A robust, digital-audio watermarking method," *IEEE Multimedia*, vol. 16, no. 3, pp. 60-69, September 2009.

[4] J. Wang, R. Healy, and J. Timoney, "A robust audio watermarking scheme based on reduced singular value decomposition and distortion removal," *Signal Processing*, vol. 91, no. 8, pp. 1693-1708, August 2011.

[5] M. A. Akhaee, N. K. Kalantari, and F. Marvasti, "Robust audio and speech satermarking using gaussian and laplacian modeling," *Signal Processing*, vol. 90, no. 8, pp. 2487-2497, August 2010.

[6] C. M. Park, D. Thap, and G. N. Wang, "Speech authentication system using digital watermarking and pattern recovery," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 931-938, June 2007.

[7] O. T. C. Chen, H. L .Chia, and Y. Chia, "Content-dependent watermarking scheme in compressed speech with identifying manner and location of attacks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1605-1616, July 2007.

[8] B. Y. Lei, I. Y. Soon, and Z. Li, "Blind and robust audio watermarking scheme based on SVD-DCT," *Signal Processing*, vol. 91. No. 8, pp. 1973-1984, August 2011.

[9] B. K. Vivekananda, S. Indranil, and D. Abhijit, "A new audio watermarking scheme based on singular value decomposition and quantization," *Circuits, Systems, and Signal Processing*, vol. 30, no. 5, pp. 915-927, October 2011.

[10] B. Xiao, J. F. Ma, and X. Wang, "Image analysis by bessel-fourier moments," *Pattern Recognition*, vol. 43, no. 8, pp. 2620-2629, August 2010.

[11] F. Li, Q. Q. Pei, and L. J. Pang, "Robust image watermarking based on bessel-fourier moments," *International Journal of Digital Content Technology and its Applications*, vol. 5, no. 11, pp. 394-401, November 2011.

[12] X. Y. Wang, T. X. Ma, and P. P. Niu, "A pseudo-Zernike moments based audio watermarking scheme robust against desynchronization attacks," *Computers and Electrical Engineering*, vol. 37, no. 4, pp. 425-443, July 2011.

Bin Han was born in Sichuan Province, China, in 1974. He received the B.S. degree from the Northeast Normal University of China (NENU), Changchun, in 1997 and the M.S. degree from the University of Electronic Science and technology (UESTC), Chengdu, in 2006, both in computer Science. His research interests include information hiding, and digital watermarking.