Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **6**, November **2015**

An Improved Phase Coding-Based Watermarking Algorithm for Speech Perceptual Hashing Authentication

Qiu-Yu Zhang, Shuang Yu, Peng-Fei Xing, Yi-Bo Huang, Zhan-Wei Ren

School of Computer and Communication
Lanzhou University of Technology
Gansu, Lanzhou, 730050, P. R. China
zhangqylz@163.com; 782809502@qq.com

Received January, 2015; revised September, 2015

Abstract. *Speech perceptual hashing authentication provides an efficient way for security authentication of speech content. To solve the problem of perceptual hash value efficient and secure transmission, we present a speech watermarking algorithm, which is based on improved phase coding by bipolar quantization and suitable for speech perceptual hashing authentication. This algorithm embeds the perceptual hash sequence which is regarded as the watermark in the speech signal. At the transmitter, the algorithm embeds watermarks in embedding positions scrambled by Logistic mapping using improved phase coding by bipolar quantization and reconstructs the speech signal. At the receiver, the algorithm extracts watermarks based on quantitative identification after embedding positions determination. Experimental results show that the proposed blind watermarking scheme has good transparency, robustness and security, and can meet the requirements of speech perceptual hashing authentication.*
Keywords: Speech perceptual hashing authentication, Audio watermarking, Phase coding, Bipolar quantization, Logistic mapping

1. **Introduction.** Speech perceptual hashing authentication has become a hot research field in multimedia security these years [1]. Developed from audio fingerprinting, based on perceptual hashing and watermarking, the speech perceptual hashing authentication, which is focusing on authentication of speech content integrity, has been considered as one of the most useful technique for multimedia authentication just like digital signature and digital watermarking [2, 3].

Represented by digital watermarking, information hiding technology transmits perceptual hash sequences in speech perceptual hashing authentication system. This is the material difference between the speech perceptual hashing authentication, especially when the real time performance is required, and other audio security research field in which the database of features is essential, such as music identification [4], speech search [5] and speaker recognition [6]. The most significant requirements of watermarking algorithm in speech perceptual hashing authentication system are blind-detection, transparency, robustness, security. However, in terms of speech perceptual hashing authentication, the vast majority of researches are concentrated on perceptual hash generation.

At present, few watermarking algorithms or information hiding techniques are proposed to provide transmission support for the speech perceptual hashing authentication system. In [7], a watermarking algorithm based on Least Significant Bit (LSB) was proposed, a

classical time-domain watermarking algorithm which is widely used but of poor robustness. An information hiding technique with a good compromise among capacity, transparency and robustness, which is based on Quantization Index Modulation (QIM) was proposed by Xiao et al. [8]. Another algorithm based on QIM focusing on improvement of transparency was proposed by Tian et al. [9]. Other valuable researches watermarking algorithms applied to speech perceptual hashing authentication are as follows: Liu et al. [10] proposed a scheme based on Bessel-Fourier moments, which is not only robust against insertion and deletion attacks, but also secure. Yan et al. [11] proposed an improved semi-fragile speech watermarking with improved transparency scheme by quantization of linear prediction (LP) parameters and modified bit allocation algorithm. Saraswathi [12] realized speech authentication based on audio watermarking by embedding watermark, which is generated from Mel-Frequency Cepstral Coefficients (MFCCs) of speech in the low intensity points detected in the signal.

Especially in the mobile environment, some direct methods such as digital label can be used to transmit perceptual hash sequence when the requirements of transmission efficiency and security are low. In the occasions with high requirement of security, perceptual hash sequence can be coded by Hamming code, Baker code or their like to improve security and robustness.

To solve the problem of perceptual hash value efficient and secure transmission, we present a speech watermarking algorithm, which is based on improved phase coding by bipolar quantization and suitable for speech perceptual hashing authentication. This algorithm takes the perceptual hash sequence which is generated from the perceptual hashing algorithm based on wavelet packet decomposition and QR decomposition proposed in [13] as the watermark to embed in the speech signal. At the transmitter, the algorithm chooses the locations where to embed the perceptual hash sequence based on the capacity and the secret key. Concretely speaking, the watermark embedding positions are scrambled randomly throughout the whole sampling point space of the speech by Logistic mapping. Then the algorithm conducts pre-processing including framing and Discrete Fourier Transform (DFT). After that, first phases of frames in which watermarks will be embedded by bipolar quantization are modified in case of preserving the relative phase of each frame. So the absolute phases are modified respectively but the relative phase difference in each frame is preserved. At last, the speech signal is reconstructed though Inverse Discrete Fourier Transform (IDFT). At the receiver, the algorithm conducts the pre-processing and the determination of embedding positions just like that at the transmitter. The extraction algorithm obtains the watermark by conducting quantitative identification of the first phases of frames in which watermarks are embedded. Experimental results show that the proposed scheme is not only blind but also of good transparency and robustness. What's more, the algorithm is secure and follows Kerckhoffs's principle, which means that for an attacker (without the correct secret key) the watermark (perceptual hash sequence) of a speech will be unpredictable during transmission. The algorithm proposed is satisfied with the requirements of perceptual speech hashing authentication.

The rest of this paper is organized as follows. Section 2 describes related theory including phase coding, bipolar quantization, Logistic mapping. The detailed watermarking algorithm proposed is described in Section 3. Performance evaluation and analysis of experimental results are given in Section 4. Finally, we conclude our paper in Section 5.

2. **Related Theory Introduction.**

2.1. **Phase Coding.**

The phase coding method is the representative of frequency-domain watermarking algorithm, which is based on the characteristic of human auditory system

(HAS) that the HAS is unable to perceive absolute phase, only relative phase, works by substituting phase of an initial audio segment with a reference phase such as $\pi/2$ or $-\pi/2$ that represents the data, adjusting other phases in order to preserving the relative phase and identifying the first phase in embedding frame to get the watermark at the receiver. Comparing with time-domain watermarking methods, such as LSB or echo hiding, phase coding is blind but also of great robustness against content preserving operations. However, considering that the reference phase coding substitutes with the original phase is $\pm\pi/2$, the speech signal changes a lot after having been embedded. That is to say, the transparency of phase coding method needs to be improved. Applying phase coding method to speech perceptual hashing authentication needs to ensure the high transparency of the watermark algorithm.

2.2. **Bipolar Quantization.** The core idea of bipolar quantization is substituting the quantitative objectives with the median of the nearest interval. The bipolar quantization procedures are as follows:

(1) Separate the value space in which $C(i)$ exists into two parts shown in Fig.1 by using Δ.

(2) If $w(i)$=1, substitute $C(i)$ with the median of the nearest A interval. If $w(i)$=0, substitute $C(i)$ with the median of the nearest B interval.

where $C(i)$ is the quantitative objective, $w(i)$ is the watermark bit, Δ is the quantization step.

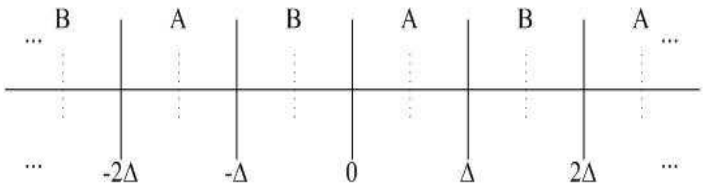


FIGURE 1. Principle of bipolar quantization.

There are many researches on watermarking based on bipolar quantization combined with various signal transform [14, 15]. The quantization step Δ is the key to get some kind of balance between robustness and transparency in watermarking algorithm.

2.3. **Logistic Mapping.** The mathematical expression of Logistic map is described as follows:

$$x_{n+1} = \mu x_n(1 - x_n) \tag{1}$$

where $0 \leq \mu \leq 4$, $x_n \in (0, 1)$, and if $3.56994 < \mu \leq 4$, Logistic mapping is in chaotic state. The characteristics that Logistic mapping in chaotic state possesses, such as aperiodic, non- convergence and sensitive dependence to the initial value, make it widely used in secure communications field.

There is one point to add. Considering the security of watermarking method, the algorithm proposed scrambles embedding positions by Logistic mapping instead of scrambling the watermark itself.

3. Proposed Algorithm.

3.1. **Framework of Speech Perceptual Hashing Authentication.** This paper proposes a method for transmitting speech perceptual hash sequence based on watermarking and the improved framework of speech perceptual hashing authentication adopted in [13]. This algorithm embeds the perceptual hashing value hp generated from the perceptual audio hash function into the original speech signal s to obtain the watermarked speech signal s'. The improved framework of speech perceptual hashing authentication is shown in Fig.2.

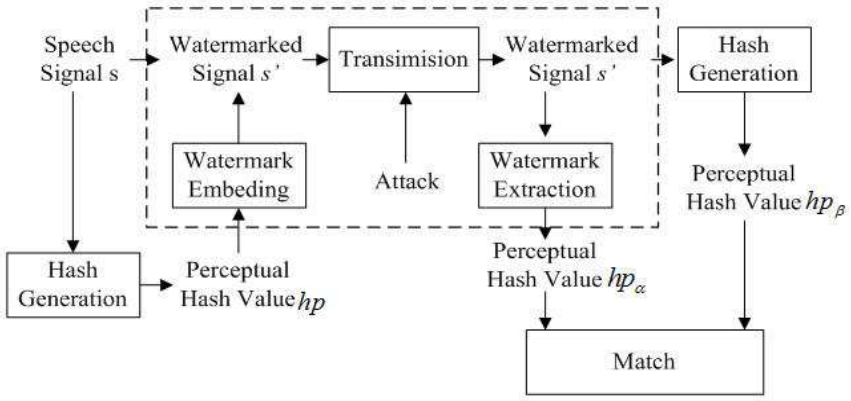


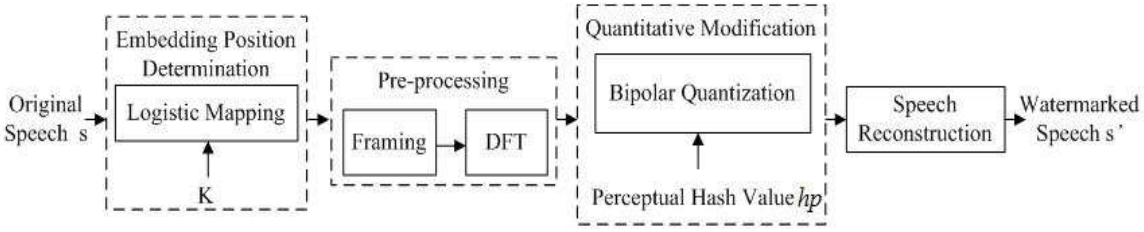
FIGURE 2. Improved framework of speech perceptual hashing authentication based on [13].

There is one point to add. The speech perceptual hashing authentication conducts matching module by comparing the distance between two perceptual hash values, hp_α and hp_β, and a pre-set threshold τ. Considering the watermark embedding, these two values cannot be the same even there are not content preserving operations or any attack during transmission. Match and authentication belong to research of perceptual hashing algorithm and are described in [13] in detail.

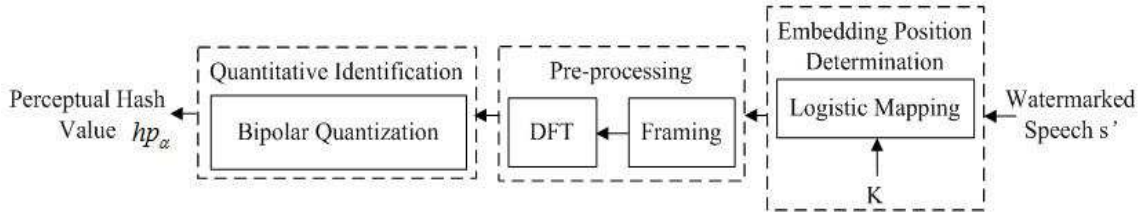
The field of main research work in this paper is show in dashed line in Fig.2. It must be emphasized that the watermarking algorithm applied to speech perceptual hashing authentication must guarantee its robustness to ensure that authentication system can extract the perceptual hash value embedded at the transmitter even after content preserving operations during transmission as much as possible, i.e. ensure that the distance between hp and hp_α as small as possible. At the same time, the watermarking algorithm must guarantee its transparency to ensure that authentication system can structure the perceptual hash value, which is similar with the one structured at the transmitter, from the watermarked signal at the receiver together with robustness of perceptual hashing algorithm, i.e. ensure that the distance between hp and hp_β as small as possible. It is obvious that the overall performance of the speech perceptual hashing authentication is represented by the distance between hp_α and hp_β.

3.2. **Watermarking Algorithm Based on Improved Phase Coding.** As can be seen in Fig.2, the transmission of the perceptual hashing value hp and the original speech signal s is realized by the watermarking algorithm. The detail embedding and extraction procedures are shown in Fig.3.

Concretely speaking, as can be seen in Fig.3(a), at the transmitter, the embedding algorithm chooses the embedding positions of perceptual hash sequence based on the capacity and the secret key. Then the algorithm preprocesses the original signal including framing and DFT. After that, the algorithm modifies the first phases of frames in which watermark bits will be embedded by bipolar quantization. At last, the speech signal is



(a)Watermark Embedding.



(b)Watermark Extraction.

FIGURE 3. The flow chart of watermark embedding and extraction.

reconstructed though IDFT. As can be seen in Fig.3(b), at the receiver, the extraction algorithm conducts the preprocessing and the determination of embedding position just like that at the transmitter. The algorithm obtains the watermark by conducting quantitative identification of the first phase of each frame in which watermark is embedded.

3.2.1. *Watermark Embedding.* Embedding steps are as follows:

Step 1.: Embedding position determination

Scramble the auxiliary array $b(i) = \begin{cases} 1, 1 \le i \le l \\ 0, l < i \le I \end{cases}$ by Logistic mapping and the secret key $K = [\mu, \alpha]$, where l is the length of watermark, I is the watermarking capacity. The element '1' of the array scrambled H_α, the binary array of which length equals the watermarking capacity, represents the embedding position.

Step 2.: Pre-processing

Firstly, segment the original speech signal, denoted as s, to I equal and non-overlapping frames, and create a matrix of frames, $S_{I\times K}$.

Secondly, map the embedding position array H_α which length is I one-for-one to the I row vector of the frames matrix $S_{I\times K}$ and apply a K-points discrete Fourier transform to n-th frame, where $H_\alpha(n)$=1, to create a matrix of the phase, $P_{l\times K} = \{\phi_j(k) | 1 \le j \le l, 1 \le k \le K\}$, and magnitude, $A_{l\times K}(1 \le j \le l, 1 \le k \le K)$.

Moreover, according to formula $\Delta\phi_j(k+1) = \phi_j(k+1) - \phi_j(k)$, store the matrix of phase difference ΔP to be embedded as follow:

$$\begin{aligned} \Delta P &= [\Delta\phi_j(k+1) = \phi_j(k+1) - \phi_j(k)]_{l\times(K-1)} \\ &= \begin{bmatrix} \phi_1(2) - \phi_1(1) & \phi_1(3) - \phi_1(2) & \dots & \phi_1(K) - \phi_1(K-1) \\ \phi_2(2) - \phi_2(1) & \phi_2(3) - \phi_2(2) & \dots & \phi_2(K) - \phi_2(K-1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_l(2) - \phi_l(1) & \phi_l(3) - \phi_l(2) & \dots & \phi_l(K) - \phi_l(K-1) \end{bmatrix}_{l\times(K-1)} \end{aligned} \tag{2}$$

Step 3.: Quantitative modification

Substitute elements in the first row of matrix of the phase $P_{l\times K}$ with watermarks for each frame. Detail procedure is as follow:

(1) Separate the phase space $[-\pi, \pi]$ into two parts shown in Fig.1 by using Δ.

(2) If perceptual hash value $hp(j) = 1$, substitute the first phase $\phi_j(1)$ with the median of the nearest A interval. If perceptual hash value $hp(j) = 0$, substitute the first phase $\phi_j(1)$ with the median of the nearest B interval. A binary set of data is represented as $\phi'_j(1)$ a representing '0' or '1'.

(3) Re-create phase matrixes by using the phase difference to obtain the modified phase matrix $P'_{l\times K}$ as follows:

$$\begin{aligned} P'_{l\times K} &= [\ \phi'_j(1) \quad \phi'_j(k) = \phi'_j(k-1) + \Delta\phi_j(k)]_{l\times(K-1)} \\ &= \begin{bmatrix} \phi'_1(1) & \phi'_1(1) + \Delta\phi_1(2) & \dots & \phi'_1(K-1) + \Delta\phi_1(K) \\ \phi'_2(1) & \phi'_2(1) + \Delta\phi_2(2) & \dots & \phi'_2(K-1) + \Delta\phi_2(K) \\ \vdots & \vdots & \ddots & \vdots \\ \phi'_l(1) & \phi'_l(1) + \Delta\phi_l(2) & \dots & \phi'_l(K-1) + \Delta\phi_l(K) \end{bmatrix}_{l\times(K-1)} \end{aligned} \tag{3}$$

Step 4.: Speech reconstruction

Use the modified phase matrix $P'_{l\times K}$ and the original magnitude matrix $A_{l\times K}$ to reconstruct the watermarked $S'_{l\times K}$ by applying the IDFT and create the $S'_{I\times K}$ with the original frames and watermarked frames in sequence to get the embedded signal s'.

3.2.2. *Watermark Extraction.* Extraction steps are as follows:

Step 1.: Embedding position determination

Step 2.: Pre-processing

Segment the received speech signal, denoted as s', to I equal and non-overlapping frames, and create a matrix of frames, $S'_{I\times K}$. Then extract the frame matrix $S'_{l\times K}$ with watermark based on the embedding position array H_α and apply K-points discrete Fourier transform to get the phase matrix with watermark $P'_{l\times K} = \{\phi_j(k)|1 \le j \le l, 1 \le k \le K\}$.

Step 3.: Quantitative identification

Identify the watermark embedded in the first row $\phi'_j(1)$ of phase matrix $P'_{l\times K}$ in phase space $[-\pi, \pi]$ separated by Δ in sequence. In detail, if $\phi'_j(1)$ lies in interval A, the watermark embedded is '1', $hp(j)_\alpha = 1$. If $\phi'_j(1)$ lies in interval B, the watermark embedded is '0', $hp(j)_\alpha = 0$.

There is one point to add. The watermarking algorithm can adjust sampling rate and number of frames to modify the capacity. As long as the capacity I is bigger than the length of perceptual hash sequence l, the proposed algorithm can be combined with different perceptual hash functions.

For a kind of perceptual hashing algorithm used in speech perceptual hashing authentication, the quantization step Δ, the length of perceptual hash sequence (watermark), denoted by l above, and the capacity of the watermarking algorithm proposed, denoted by I above are constant for speech clips. Moreover, the embedding position can be determined by the secret key at the receiver. So there is no need of any information of the original speech in the extraction process and the watermarking algorithm can realize blind-detection.

4. **Performance Evaluation and Analysis of Experimental Results.** A total number of 150 English speech clips (16-bit signed, 16 kHz sampled and 4 s length) randomly selected from English Language Speech Database for Speaker Recognition (ELSDSR) speech database are used to the evaluation of the proposed algorithm. The length of the perceptual hash sequence generated from perceptual hash function is 256.

4.1. **Transparency.** Signal to Noise Ratio (SNR), which has been widely used in watermarking research fields, and Perceptual Evaluation of the Speech Quality ($PESQ$) [16], which is provided by ITU and whose range is -0.5(worst) to 4.5(best), are used to

evaluate the transparency of watermarking algorithm proposed. SNR can point out the difference between the original speech and the watermarked one and is defined by the following formula.

$$SNR = 10 \times \log \left[\frac{\sum_{i=1}^{L} s^2(i)}{\sum_{i=1}^{L} [s(i) - s_w(i)]^2} \right] \tag{4}$$

where $s(i)$ is the original signal, $s_w(i)$ is the watermarked signal and L is the total number of samples.

It is known from the bipolar quantization theory that the quantization step Δ has great influence on the transparency of watermarking algorithm. The transparency is evaluated with different quantization step and $I = 400$. Average SNR and $PESQ$ of 150 speech clips are summarized as shown in Table 1 and $PESQ$ distribution is shown in Fig.4 with $\Delta = \pi/6$.

TABLE 1. SNR and $PESQ$ of the proposed algorithm

Quantization Step Δ	SNR	$PESQ$
$\pi/6$	30.71	≈4.5
$\pi/9$	37.72	
$\pi/18$	49.75	

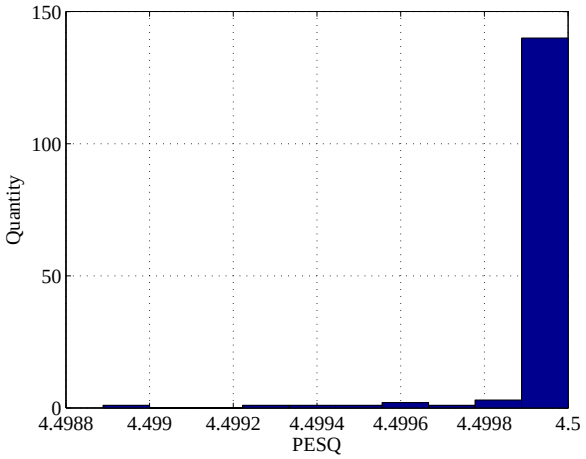


FIGURE 4. $PESQ$ distribution with $\Delta = \pi/6$.

As can be seen in Table 1, according to the range of $PESQ$ and the recommendation from ITU that for a watermarking algorithm the SNR should be more than 20 dB, the algorithm proposed with different parameters all meet the requirement. It can be known from Fig.4 that the majority of $PESQ$ values of 150 speech clips are almost 4.5 even with $\Delta = \pi/6$. Take a speech clip named "FAML-Sa.wav" with the parameters of the worst transparency, $\Delta = \pi/6$ and I=400 for instance, comparison of signal before and after embedding is shown in Fig.5.

As shown in Fig.5, the difference between original waveform and watermarked waveform is not obvious.

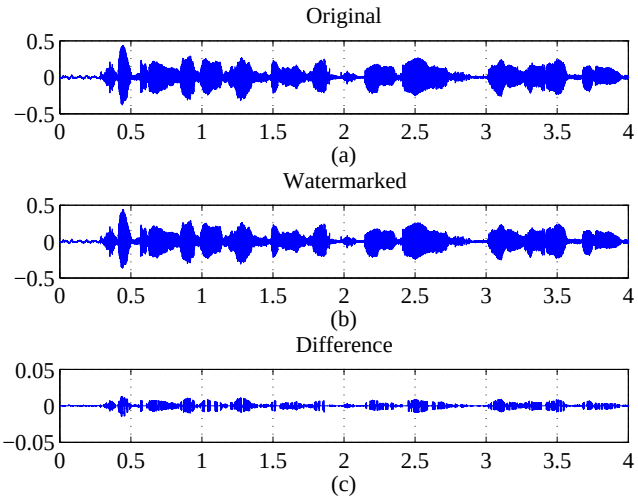


FIGURE 5. Signal before and after embedding.

4.2. **Robustness.** Bit error rate (BER), which has been widely used to evaluate the robustness of algorithms, can points out the error bits percentage in the total number of bits and calculate the distance between the perceptual hash values extracted ph_α and the one regenerated ph_β at the receiver. BER can be used as follow.

$$BER = \frac{\sum_{i=1}^{N}(|ph_\beta(i) \oplus ph_\beta(i)|)}{N} \tag{5}$$

where N is the length of perceptual hash sequence.

The following types of content preserving operations are used to evaluate the robustness of the algorithm proposed:

(1) Decrease volume: volume decreased by 50%;

(2) Increase volume: volume increases by 50%;

(3) Re-sampling 8-16: sampling frequency reduced to 8 kHz, and up to 16 kHz;

(4) Re-sampling 32-16: sampling frequency up to 32 kHz, and reduced 16 kHz;

(5) Narrow-band noise: with the center frequency distribution in $0 \sim 4$ kHz narrow-band Gaussian noise;

(6) FIR filter: using a twelve order FIR low-pass filter with cut-off frequency of 3.4 kHz;

(7) Butterworth filter: using a twelve order Butterworth low-pass filter with cut-off frequency of 3.4 kHz;

(8) Echo addition: stack attenuation was 60%, the time delay for 300 ms.

Considering that speech perceptual hashing authentication is consisted of two parts: the perceptual hash function and the watermarking algorithm, and the robustness of authentication system are influenced by them both, the evaluation of robustness consisted of two parts: the watermarking algorithm only and the whole authentication system, that is calculating BER_α between ph and ph_α, BER_β between ph_α and ph_β with the parameters $\Delta = \pi/6$ and $I = 400$. The results are summarized as shown in Table 2.

As can be seen in Table 2, BER_α, the BER between the ph and ph_α, represents the difference between the perceptual hash value embedded in speech signal at the transmitter ph and the watermark extracted at the receiver ph_α, namely the robustness of the

TABLE 2. BER of watermarking algorithm and authentication system

Operating means		BER_α	BER_β
Volume Adjustment	50%	2.4568e-05	0.0089
	150%	2.4568e-05	0.0084
Re-sampling	16kHz→ 32kHz	2.4568e-05	0.0082
	16kHz→ 8kHz	0.0098	0.0770
Low-pass Filtering	3.4kHz FIR filter	0.0986	0.1638
	3.4kHz Butterworth filter	0.1383	0.2412
Echo Addition		0.1324	0.1894
White Noise Addition ($50dB$)		0.1029	0.1589

watermarking algorithm proposed. Considering the ideological core of the proposed algorithm is the modification of phase, the objective of volume adjustment is the magnitude and the 32 kHz re-sampling does not influence the sampling points obviously also, so the algorithm proposed has great robustness against these content preserving operations. However, other content preserving operations have certain influences on phase.

BER_α, the BER between the ph_β and ph_α, represents the robustness of the whole authentication system. In ideal state, the perceptual hash value extracted at the receiver equals the one embedded as watermark at the transmitter. However, influenced by the watermarking algorithm and content preserving operations during transmission, the two hash values are different. To decrease the difference between the two hash values, the key lies in the improvement of transparency of watermarking algorithm and robustness of perceptual hash function. As can be seen in Table 2, compared with the watermarking algorithm, the robustness of the whole authentication system represented by BER_β decrease somewhat. The robustness of perceptual hash function against content preserving operations and the transparency of watermarking algorithm cause this situation.

Based on the experimental results and analysis above, researches on the robustness of speech perceptual hashing authentication should focus on improving the robustness of both watermarking algorithm and perceptual hash function, and the transparency of watermarking algorithm. However, the robustness and transparency of watermarking algorithm are contradictory. How to balance them two seems essential.

4.3. **Security.** Logistic mapping controlled by $K = [\mu, \alpha]$ guarantees the security of the algorithm proposed. Illegal users with wrong secret key hardly can obtain watermark even the algorithm is open.

The following will be the simulation of behavior from eavesdroppers based on Kerckhoffs's principle that extracting watermarks from signals with different secret.

According to scrambling encryption adopted in this paper, perceptual hash sequences (watermark) from 150 speech clips are embedded into original signals respectively by Logistic mapping with the secret key $K = [\mu, \alpha]$. Eavesdroppers extract watermarks from signals with different secret key $K_1 = [\mu_1, \alpha_1]$ during transmission. Calculate BER between perceptual hash sequences of these two procedures. In addition, the parameters of watermarking algorithm are as follows:$\Delta = \pi/6$, $I = 400$, $K = [3.8, 0.6]$. The results are summarized as shown in Table 3.

As can be seen in Table 3, security of algorithm proposed seems poor according to the BER_s. However, it is resulted from the particularity of scrambling binary arrays in fact. Considering scrambling a unary numeral array, BER between results with different secret keys is 0. In a numeral system, the less the number is, the small the scrambling performance reflecting on BER is. The proof is as follow:

TABLE 3. BER of different key

$Key([\mu_1, \alpha_1])$	BER_s
$[3.7, 0.5]$	0.4396
$[3.9, 0.7]$	0.4277
$[3.95, 0.45]$	0.4505
$[3.6, 0.1]$	0.4418

Proof. Let $x(i)$ be an N carry array, and $x'(i)$ be scramble from $x(i)$, $1 \le i \le I$. Then Event A:$x(i) \neq x'(i)$, $p(A) = \dfrac{N-1}{N}$. Then

$$BER = \frac{\sum_{i=1}^{I}(|x(i) \oplus x'(i)|)}{I} = \frac{p(A) \times I}{I} = \frac{N-1}{N}$$ □

In addition, conduct scrambling encryption with $K = [3.8, 0.6]$ and scrambling decryption with $K_1 = [3.6, 0.9]$ on 500 binary, octal, decimal and hexadecimal arrays of which length is 256. BER between results before encryption and after decryption are as shown in Table 4.

TABLE 4. BER in different numeral systems

N	BER
2	0.4996
8	0.8673
10	0.8939
16	0.9304

As shown and analyzed above, the BER of binary arrays tends to 0.5 and is smaller than other arrays in different numeral system indeed. But it does not mean that the scrambling performance is poor. For the same reason, the small BER in Table 3 does not mean the poor performance of scrambling encryption in the algorithm proposed

5. **Conclusions.** A watermarking algorithm based on improved phase coding applied to speech perceptual hashing authentication, which is used to transmit the perceptual hash sequence efficiently and securely, is proposed in this paper. Doing researches on transmission of the perceptual hash sequence can not only perfects speech perceptual hashing authentication but also expand application scope of the system. Experimental results show that the algorithm proposed is not only blind but also of good transparency, robustness against content preserving operations and security during transmission. The proposed watermarking algorithm meets the requirements of speech perceptual hashing authentication.

Further research on speech perceptual hashing authentication focuses on reduction of computational complexity and combination with low bit rate coding standards.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No. 61363078), the Natural Science Foundation of Gansu Province of China (No. 1310RJYA004). The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

[1] G. Grutzek, J. Strobl, B. Mainka, F. Kurth, C. Pörschmann, and H. Knospe, Perceptual hashing for the identification of telephone speech, *Proc. of the Speech Communication; 10. ITG Symposium*, IEEE, 2012, pp. 1-4.

[2] S. Adibi. A low overhead scaled equalized harmonic-based voice authentication system, *Telematics and Informatics*, vol. 31, no. 1, pp. 137-152, 2014.

[3] H. X. Wang, M. Q. Fan. Centroid-based semi-fragile audio watermarking in hybrid domain, *Science China Information Sciences*, vol. 53, no. 3, pp. 619-633, 2010.

[4] M. A. Nematollahi and S. A. R. Al-Haddad. An overview of digital speech watermarking, *International Journal of Speech Technology*, vol. 16, no. 4, pp. 471-488, 2013.

[5] H. X. Wang, L. N. Zhou, W. Zhang, and S. Liu. Watermarking-Based Perceptual Hashing Search over Encrypted Speech, *Digital-Forensics and Watermarking*, Springer Berlin Heidelberg, pp. 423-434, 2014.

[6] E. Karpove. Efficient Speaker Recognition for Mobile Devices. Dissertation, *University of Eastern Finland*, Finland, 2011.

[7] Y. Q. Cao, S. Bai, K. Cai, and W. D. Li. Study on covert communication system based on G.729 compressed voice stream, *Modern Electronics Technique*, vol. 36, no. 17, pp. 68-70, 2013.

[8] B. Xiao, Y. Huang, and S. Tang. An approach to information hiding in low bit-rate speech stream, *Proc. of the Global Telecommunications Conference (GLOBECOM'2008)*, IEEE, pp. 1-5, 2008.

[9] H. Tian, J. Liu, and S. Li. Improving security of quantization-index-modulation steganography in low bit-rate speech streams, *Multimedia Systems*, vol. 20, no. 2, pp. 143-154, 2014.

[10] Z. Liu and H. X. Wang. A novel speech content authentication algorithm based on BesselCFourier moments, *Digital Signal Processing*, vol. 24, pp. 197-208, 2014.

[11] B. Yan and Y. J. Guo. Speech authentication by semi-fragile speech watermarking utilizing analysis by synthesis and spectral distortion optimization, *Multimedia tools and applications*, vol. 67, no. 2, pp. 383-405, 2013.

[12] S. Saraswathi. Speech authentication based on audio watermarking, *International Journal of Information Technology*, vol. 16, no. 1, pp. 34-43, 2010.

[13] Q. Y. Zhang, P. F. Xing, Y. B. Huang, R. H. Dong, and Z. P. Yang. An Efficient Speech Perceptual Hashing Authentication Algorithm Based on Wavelet Packet Decomposition, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 6, no. 2, pp. 311-322, 2015.

[14] J. Q. Zhang and H. X. Wang. Analysis on Law of Distortion of Audio Signal for Embedding Watermark in DCT and DWT, *Acta Electronica Sinica*, vol. 41, no. 6, pp. 1193-1197, 2013.

[15] C. D. Wang and D. F. Ma. Information hiding based on real-time voice in DCT domain, *Computer Engineering and Design*, vol. 33, no. 2, pp. 474-478, 2012.

[16] ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs, *ITU-T*, Jan. 2002.