IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 22, 2020, accepted February 10, 2020, date of publication February 14, 2020, date of current version February 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974029

A Long Sequence Speech Perceptual Hashing Authentication Algorithm Based on Constant Q Transform and Tensor Decomposition

YIBO HUANG, HEXIANG HOU, YONG WANG, YUAN ZHANG, AND MANHONG FAN
College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China

Corresponding author: Yibo Huang (huang_yibo@foxmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61862041, and in part by the Youth Science and Technology Fund of Gansu Province of China under Grant 1606RJYA274.

ABSTRACT Most speech authentication algorithms are over-optimized for robustness and efficiency, resulting in poor discrimination. Hashing shorter sequence is likely to cause the same hashing sequence to come from different speech segments, which will cause serious deviations in authentication. Few people pay attention to the research on the discrimination of hashing sequence length, so this paper proposes a long sequence speech authentication algorithm based on constant Q transform (CQT) and tensor decomposition (TD). In this paper, hashing long sequence is used to solve the problem of poor collision resistance of existing algorithms, fast and accurate authentication can be achieved for important speech fragments with large data volumes. The sub-band in the frequency domain are first divided into different matrix, then the variance set of sub-band in the frequency domain is obtained, and finally the feature values are obtained by CQT and TD transformation. The obtained feature values have strong robustness and can cope with the interference of complex channel environment. In this paper, Texas Instruments and Massachusetts Institute of Technology (TIMIT) speech database and the Text to Speech (TTS) are used to establish a database of 51600 speeches to verify the performance of the algorithm. Experimental results show that compared with the existing speech authentication algorithms, the proposed algorithm has the characteristics of high discrimination, strong robustness and high efficiency.

INDEX TERMS Speech authentication, perceptual hashing, CQT, TD, hashing long sequence, discrimination.

I. INTRODUCTION

With the development of multimedia technology, the speech not only has a huge amount of data, but also has the characteristics of high redundancy and low confidentiality. Therefore, speech authentication, integrity verification and content recognition face great challenges. At present, speech authentication methods mainly include watermarking technology and digital signature. The disadvantage of watermarking technology is that the original data will be modified and the quality of the speech will be degraded after embedding the watermark [1], [2], [3], [36]. Digital signature technology is too sensitive to changes in the binary level of speech data to be suitable for speech content [35]. The perceptual hash function converts the speech data into a short binary string. When the speech datas are the same or similar, they generate the same hash value. For those different speech datas, the hash function could produce different hashing sequence [4]. Therefore, the speech content authentication based on the perceptive hashing just solves the disadvantages of the above method and is also suitable for the speech authentication in the big data environment.

Speech perception hashing authentication mainly consists of two parts: hash construction and matching, among which hashing construction has a very important impact on the performance of the algorithm. At present, the features extracted from speech signals include short-term energy, short-term correlation, Mel-frequency cepstral coefficient (MFCC) [7], [28], cochleagram [9], spectral entropy [11], short-term zero-crossing rate [12], discrete

The associate editor coordinating the review of this manuscript and approving it for publication was Aniello Castiglione.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

wavelet transform (DWT) [10], [13], linear prediction coefficient (LPC) [14], spectrogram [22], [27], formant [24], bark frequency Cepstral coefficients [29] and multiple fusion features. Li *et al.* [8] proposed an audio hash scheme based on non-negative matrix factorization (NMF) of modified discrete cosine transform (MDCT) coefficients. The algorithm has good robustness, especially compression aspects such as MP3 and AAC, but its processing efficiency is relatively low. Zhang *et al.* [11] proposed an efficient perceptual hashing based on improved spectral entropy for speech authentication. The algorithm has higher efficiency, but its collision resistance performance and robustness performance at the MP3 compression is relatively poor. In Ref. [25], the speech authentication algorithm used a ternary hashing sequence instead of a binary hash sequence, and the hash construct proved to be flexible. The algorithm is not only robust to content preserving operations, but also highly efficient. Jiang *et al.* [26] proposed an audio fingerprinting extraction algorithm based on lifting wavelet packet and improved optimal-basis selection. Although the algorithm has strong robustness and efficiency, it reflects fragmentary speech data and has certain limitations. Hammad and Wang [39] proposed a secure multimodal biometric system by fusing electrocardiogram (ECG) and fingerprint based on convolution neural network (CNN). The proposed algorithm is efficient, robust and reliable, and provides a new idea for speech authentication.

Although the length of hashing sequence can affect the collision resistance performance of speech authentication, there is a lack of research on the length of hashing sequence. In Ref. [10], the algorithm given rotating QR decomposition is used to extract speech feature parameters for wavelet packet coefficient matrix, and then perceptual hashing sequence is constructed. Although the algorithm compares the effects of different length hashing sequences on the discrimination of the algorithm, the algorithm only adopts the hashing sequence of 250 bits, without in-depth discussion of the characteristics of the hashing long sequence. Zhang *et al.* [13] proposed a high-performance speech perceptual hashing authentication algorithm based on DWT and measurement matrix. The algorithm adopts the length of 360 bits hashing sequence. Although the discrimination of the algorithm has been improved, its comprehensive performance remains to be improved. Therefore, the increase of hashing sequence length can improve the algorithm discrimination.

To sum up, it can be found that the existing speech perception hashing algorithms adopt shorter hashing sequences, which easily leads to the mapping of multimedia numbers of different perception contents to the same perception hashing value, thus making the algorithm lower discrimination. Most authentication algorithms are optimized independently for robustness and authentication efficiency, without balancing the performance of the whole algorithm. To solve the above problems, this paper studies a novel long sequence speech perception hashing algorithm based on tensor decomposition.

TABLE 1. Notations and symbols.

Symbol	Definition
$x(n)$	Input speech signal
$\omega(m)$	The window function, Hamming window
M	The length of a frame of speech signal
N	The number of frames of the speech signal, and the length of the hashing sequence $(M < N)$
q	Number of sub-band in frequency domain
r	Number of sub-band variance sets
b	Determines the weight of time-frequency resolution
Q	The ratio of center frequency to bandwidth
K	The frequency band number of CQT
f_k	The central frequencies of each frequency band
δ_f	The bandwidth
G	The core tensor
W	The target tensor
P_i	Feature vector, $i = 1, 2, \cdots, N$
h	A one-dimensional binary hashing long sequence (bit vector), $h \in \{0, 1\}^N$
τ	A perceptual authentication threshold

Hashing long sequence can improve the discrimination of the algorithm. Using uniform sub-band variance and CQT can enhance the robustness of the algorithm. In this paper, the algorithm is optimized in structure and the authentication efficiency is also improved.

The rest of this paper is organized as follows: Section II describes the related theory. Section III illustrates the detailed proposed algorithm on a long sequence speech perceptual hashing authentication based on CQT and TD. Section IV gives the experimental results and the performance analysis compared with other related methods. Finally, Section V concludes the paper with future work. The major symbols used in this paper are summarized in Table 1 for easy reference.

II. RELATED THEORY INTRODUCTION

A. UNIFORM SUB-BAND VARIANCE

The features of speech and noise are different in spectrum domain. The energy of speech varies greatly with the frequency band. There is a large peak at the formant, and a small energy at other frequencies. However, the noise energy is much smaller than the speech energy, and it is more evenly distributed in the frequency band. In this paper, the frequency band variance can not only reduce noise interference, but also enhance the robustness of the algorithm.

The time-domain waveform of the speech signal is $x(n)$, and $x_i(m)$ is the i-frame speech signal obtained after preprocessing by adding window division, then it is satisfied

$$x_i(m) = \omega(m) * x(iT + m) \quad 1 \le m \le M \tag{1}$$

where $\omega(m)$ is the window function; M is the frame length; T is move the frame length.

The spectrum is obtained by applying $x_i(m)$ to the discrete Fourier transform (DFT).

$$X_i(l) = \sum_{m=0}^{M-1} x_i(m) exp(-j\frac{2\pi lm}{M}) \quad 0 \le l \le M-1 \tag{2}$$

In the frequency domain, the data length of each frame is M, and there are $(\frac{M}{2}+1)$ spectral lines in the positive frequency domain after DFT. The $(\frac{M}{2}+1)$ spectral lines $X_i = X_i(1), X_i(2), \cdots, X_i(\frac{M}{2}+1)$ are divided into q sub-bands, and each sub-band contains $p = fix[(\frac{M}{2}+1)/q]$ spectral lines ($fix[\cdot]$ represents the integer part).

$$XX_i(j) = \sum_{l=1+(m-1)p}^{1+(m-1)p+(p-1)} |X_i(l)| \quad 1 \le j \le q \tag{3}$$

In this paper, sub-bands are divided into r sub-band sets, and the variance of each sub-band set is obtained.

$$XX_i = [XX_i^{(1)}, XX_i^{(2)}, \cdots, XX_i^{(r)}] \tag{4}$$

where the first sub-band set is $XX_i^{(1)} = [XX_i^{(1)}(1), XX_i^{(1)}(2), \cdots, XX_i^{(1)}(q/r)]$, the mean and variance can be obtained from Equations (5) and (6).

$$E_i^{(1)} = \frac{r}{q}\sum_{t=1}^{q/r} XX_i^{(1)}(t) \tag{5}$$

$$D_i^{(1)} = \frac{1}{q/r-1}\sum_{t=1}^{q/r}[XX_i^{(1)}(t) - E_i^{(1)}] \tag{6}$$

Since each sub-band has p spectral lines after the original DFT, it is called uniform sub-band. In other words, each sub-band is of equal bandwidth. Each sub-band set contains the same number of sub-bands. The variance of all sub-band sets per frame is $D_{r,i} = [D_i^{(1)}, D_i^{(2)}, \cdots, D_i^{(r)}]$.

B. CONSTANT Q TRANSFORM

The essence of CQT is variable resolution processing, that is, the low frequency part has high frequency resolution and the high frequency part has high time resolution. CQT not only inherits the advantages of high resolution and high precision of DFT, but also has good robustness [16], [17]. In CQT, the relation between the central frequencies of each frequency band f_k is defined as Equation (7).

$$f_k = f_{min}2^{\frac{k-1}{b}} \tag{7}$$

where f_{mim} is the lowest frequency of CQT spectrum; b is the parameter, which determines the weight of time-frequency resolution. It can be seen from Equation (7) that the frequency domain of each frequency band is different. This is different from the frequency domain of the DFT, where each band has an equal frequency domain.

In CQT, Q represents the ratio of center frequency to bandwidth, which is a constant independent of k.

$$Q = \frac{f_k}{\delta_f} = \frac{f_k}{f_{k+1}-f_k} = (2^{1/b}-1)^{-1} \tag{8}$$

where δ_f is the bandwidth.

CQT of discrete signal $x(n)$ is shown in Equation (9).

$$X^{CQT}(k,n) = \sum_{j=n-\lfloor N_k/2\rfloor}^{n+\lfloor N_k/2\rfloor} x(j)a_k^*(j-n+N_k/2) \tag{9}$$

where $k = 1, 2, \cdots, K$ is the frequency band number; $a_k^*(n)$ denotes the complex conjugate of $a_k(n)$; N_k are the variable window lengths; $\lfloor\cdot\rfloor$ denotes rounding towards negative infinity.

$$K = \lfloor blog_2(\frac{f_{max}}{f_{min}})\rfloor \tag{10}$$

$$N_k = Q\frac{f_s}{f_k} \tag{11}$$

$$a_k(n) = \frac{1}{\Theta}\omega(\frac{n}{N_k})e^{i(\frac{2\pi nQ}{N_k}+\phi_k)} \tag{12}$$

where f_s is the sampling rate; $\omega(t)$ is a window function (e.g. Hamming window); ϕ_k is the phase shift; Θ is a given scaling factor.

$$\Theta = \sum_{l=-\lfloor N_k/2\rfloor}^{\lfloor N_k/2\rfloor} \omega(\frac{l+N_k/2}{N_k}) \tag{13}$$

C. TENSOR DECOMPOSITION

TD are efficient tools for data analysis and has been successfully applied in many applications, such as data mining, graph analysis, signal processing and computer vision [19], [21], but its use in speech perceptual hashing authentication is rarely discussed. In this paper, TD is used to derive perceptual hash, and Tucker decomposition algorithm is selected to realize TD.

For a third-order eigentensor $V \in \mathbb{R}^{Q_1\times Q_2\times Q_3}$, the Tucker decomposition will decompose it into a core tensor $G \in \mathbb{R}^{I\times J\times K}$ and three orthogonal factor matrices $U_1 \in \mathbb{R}^{Q_1\times I}$, $U_2 \in \mathbb{R}^{Q_2\times J}$, and $U_3 \in \mathbb{R}^{Q_3\times K}$. Mathematically, Tucker's decomposition is expressed in Equation (14).

$$\begin{aligned} V &\approx [\![G; U_1, U_2, U_3]\!] \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} g_{i,j,k}(u_i^{(1)} \circ u_j^{(2)} \circ u_k^{(3)}) \end{aligned} \tag{14}$$

where $u_i^{(1)}$, $u_j^{(2)}$, and $u_k^{(3)}$ are the column vectors of the matrix U_1, U_2, and U_3 respectively; $g_{i,j,k}$ represents the core tensor G; the symbol '$\circ$' represents the cross product of the two vectors; and the symbol '$[\![\cdot]\!]$' is a concise representation of Tucker decomposition. Equation (14) can be rewritten as Equation (15).

$$v_{w,h,r} \approx \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} g_{i,j,k}u_{w,i}^{(1)}u_{h,j}^{(2)}u_{r,k}^{(3)} \tag{15}$$

where $v_{w,h,r}$, $u_{w,i}^{(1)}$, $u_{h,j}^{(2)}$ and $u_{r,k}^{(3)}$ are the elements of V, U_1, U_2 and U_3 respectively. Calculation of Tucker decomposition is equivalent to solving an optimization problem as follows.

$$\begin{aligned} &[\![G; U_1, U_2, U_3]\!] \\ &= arg\min_{G,U_1,U_2,U_3} \|V - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} g_{i,j,k}(u_i^{(1)} \circ u_j^{(2)} \circ u_k^{(3)})\|_2 \end{aligned} \tag{16}$$

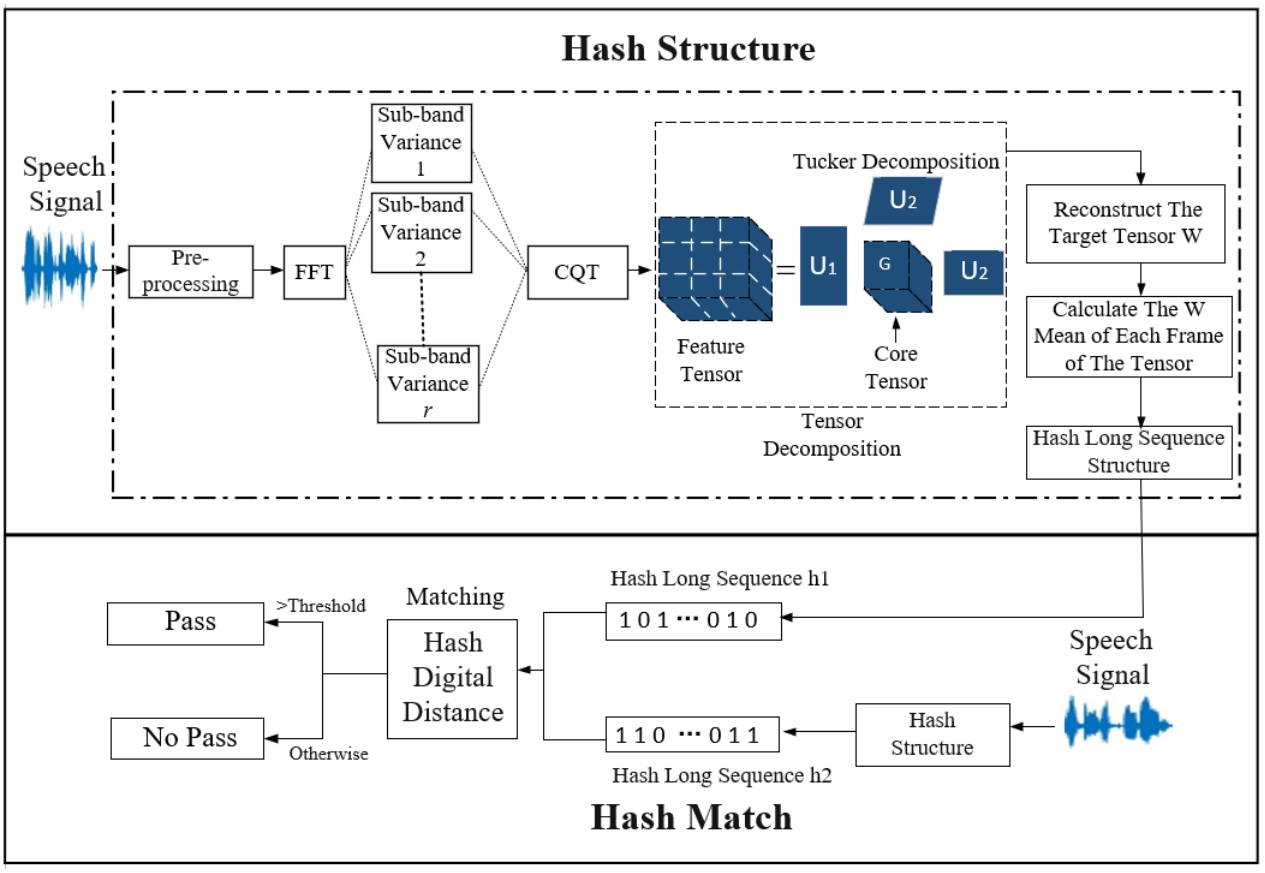


FIGURE 1. Block diagram of the proposed speech perceptual hashing algorithm.

where $\|\cdot\|_2$ is the Frobenius norm. In general, this optimization problem can be solved by alternating least squares (ALS). Tucker decomposition is shown in Fig 1.

In this paper, the sub-band variance set matrix of speech frequency domain is converted into a third-order tensor $V \in \mathbb{R}^{K\times r\times N}$ by CQT. Tensors can embody the whole framework of speech features in three-dimensional space, and the sub-band variance features of each frame of speech signal are carefully positioned in space. In this paper, the target tensor is adopted. The target tensor is very similar to the original feature tensor. The target tensor eliminates the noise and enhances the speech feature. To obtain the target tensor, the core tensor should be combined with the orthogonal matrix for reconstruction, as shown in Equation (17).

$$W = G \times_1 U_1^T \times_2 U_2^T \times_3 U_3^T \tag{17}$$

where W is the target tensor. By reducing the dimension of the target tensor W, a one-dimensional long matrix P is obtained.

$$\begin{aligned} P &= [p_1, p_2, \cdots, p_N] \\ &= \left[\frac{\sum W_{K,r}^{(1)}}{K*r}, \frac{\sum W_{K,r}^{(2)}}{K*r}, \cdots, \frac{\sum W_{K,r}^{(N)}}{K*r}\right] \end{aligned} \tag{18}$$

where $W_{K,r}^{(1)}, W_{K,r}^{(2)}, \cdots, W_{K,r}^{(N)}$ are the matrix for each frame of the target tensor W; $p_1, p_2, \cdots, p_N$ are the mean value of the eigenmatrix of the target tensor W for each frame.

III. THE PROPOSED ALGORITHM

The generic block diagram of the proposed long sequence speech perceptual hashing authentication algorithm based on CQT and TD is shown in Fig 1. The hash structure and matching of the speech signal are carried out, and the processing steps are as follows :

Step 1: Pre-processing Pre-processing includes pre-emphasis, framing and windowing. The speech signal $x(n)$ is obtained by pre-emphasis the input signal $s(n)$. Pre-emphasis can increase the features of the speech signal's high-frequency components, which is advantageous to further spectrum analysis. Then, the processed signal is framed and windowed, where in the window function selects a Hamming window to smooth the edge of the frame. the speech $x(n)$ is divided into N frame, and signal $x(m) = \{x_i(m)|i = 1, 2, \cdots, N, m = 1, 2, \cdots, M\}$ is obtained, where the subscript i represents the i frame after framing.

Step 2: FFT The time domain signal is converted into the frequency domain signal, and the frequency domain signal $A = \{A_i(m)|i = 1, 2, \cdots, N; m = 1, 2, \cdots, M\}$ is obtained.

Step 3: Construct sub-band variance set matrix The frequency domain signal is divided into uniform sub-bands, and the frequency domain sub-band matrix $B = \{B_i(m)|i = 1, 2, \cdots, N; m = 1, 2, \cdots, q\}$ is obtained. Then the

sub-band matrix is divided, and the sub-band set matrix $C = \{C_i(m)|i = 1, 2, \cdots, N; m = 1, 2, \cdots, r\}$ is obtained, q is an integer multiple of r.

$$\begin{aligned} C &= \begin{bmatrix} C_i(1) \\ C_i(2) \\ \vdots \\ C_i(r) \end{bmatrix} \\ &= [B_i(1) \quad B_i(2) \quad \cdots \quad B_i(q/r)]^T \\ &\quad [B_i(q/r+1) \quad B_i(q/r+2) \quad \cdots \quad B_i(2q/r)]^T \\ &\quad \vdots \\ &\quad [B_i((r-1)q/r+1) \quad B_i((r-1)q/r+2) \cdots \quad B_i(q)]^T \end{aligned} \tag{19}$$

Finding the variance of each sub-band set, and geting the matrix of sub-band variance set is $D = \{D_i(m)|i = 1, 2, \cdots, N; m = 1, 2, \cdots, r\}$.

Step 4: CQT The sub-band variance set matrix is transformed by CQT to obtain the two-dimensional feature matrix $E^*_{K,N}$. The feature matrix of the variance of each sub-band set is fused to obtain a feature tensor V.

$$V = \left[E^*_{K,1,N}, E^*_{K,2,N}, \cdots, E^*_{K,r,N}\right]^T \tag{20}$$

Step 5: TD The feature tensor is decomposed by Tucker, and then the low-dimensional core tensor G and three orthogonal matrices $U_{(1,2,3)}$ are recombined to obtain the target tensor W.

Step 6: Hashing long sequence structure The mean value of each frame of the target tensor W is calculated to obtain the target matrix P. The target matrix is constructed with hash length sequence to generate a one-dimensional binary hash long sequence h.

$$h_1(i) = \begin{cases} 1, & if\ P_1(i) > P_1(i-1) \\ 0, & Otherwise \end{cases} \tag{21}$$

where $h(1)$=0; $h(i)$ is the perceived hash value of each frame speech signal.

Step 7: Hashing digital distance and matching For the two speech clips $s1$ and $s2$, their hashing digital distance $BER(:, :)$ can be calculated from the formula as follows:

$$BER\left(h_{s1}, h_{s2}\right) = \sum_{i=1}^{N} \left(|h_{s1}(i) - h_{s2}(i)|\right) / N \tag{22}$$

where h_{s1} and h_{s2} respectively represent hashing long sequences for $s1$ and $s2$; N is the length of the hashing sequence.

In this paper, we use the hypothesis test of hashing digital distance $BER(:, :)$ to describe the hashing matching.

W_0: if the perceptual content of the two speech clips $s1$ and $s2$ are the same:

$$BER(h_{s1}, h_{s2}) \leqslant \tau \tag{23}$$

W_1: if the perceptual content of the two speech clips $s1$ and $s2$ are not the same:

$$BER(h_{s1}, h_{s2}) > \tau \tag{24}$$

where τ represents the perceptual authentication threshold, $h(\cdot)$ is a perceptual hashing function. By setting the size of matching threshold τ, calculating the digital distance between perceptual hashing sequences of the speech clip $s1$ and $s2$. If the digital distance $BER(:, :) \leqslant \tau$, then when their perceptual content are treated as the same, the authentication is passed, and otherwise it is failed.

In order to evaluate the performance of the authentication algorithm, the False Accept Rate (FAR) and False Reject Rate (FRR) of the algorithm can be calculated by Equations (16) and (17).

$$FAR\left(\tau\right) = \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\, dx \tag{25}$$

$$FRR\left(\tau\right) = 1 - \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\, dx \tag{26}$$

where τ is the perceptual authentication threshold, μ is the expected value, σ is the standard deviation. Generally speaking, FAR and FRR are used to evaluate the robustness and discrimination of the authentication algorithm. The lower FAR denotes the better discrimination, and the lower FRR denotes the better robustness.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

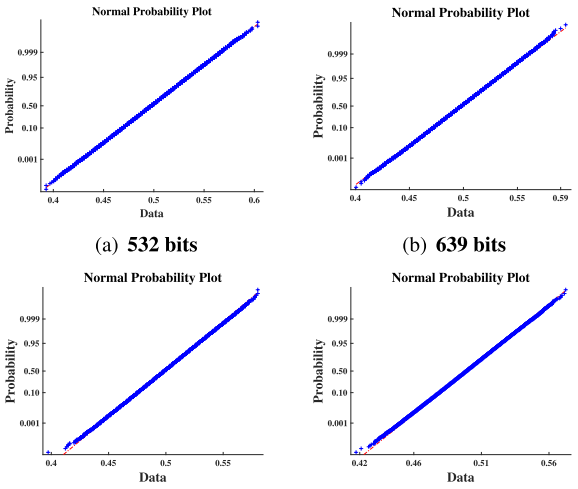
The operating experimental hardware platform is Intel(R) Core(TM) i5-7500 CPU, 3.40 GHz, with computer memories of 4G. The operating software environment is MATLAB R2018b of Windows 7 system.

In this study, after lots of experiments, we found that the following parameters are given the best results after applying it to the proposed algorithm: $M = 178; N = 1064; q = 25; r = 5; b = 12; K = 34$. Where: M is the length of a frame of speech signal, N is the length of the hashing sequence, q is number of sub-band in frequency domain, r is number of sub-band variance sets, b is the parameter, K is the frequency band number.

A. DATASETS

The experimental speech datas comes from TIMIT speech database and TTS speech database. There are different 1200 speech clips in the original speech database. The format of each speech clip is WAV with the length 4 s, which is of the form of 16 bits PCM, mono and sampled at 16 kHz.

According to the environment of speech transmission, the content preserving operations are performed on each speech in the speech database. A speech database of 14400 different content preserving operations was established, including 12 types of content preserving operations, such as echo, noise, low pass filter, resampling and MP3 compression.



(a) **532 bits** (b) **639 bits**

(c) **798 bits** (d) **1064 bits**

FIGURE 2. BER normal distribution of different hashing sequence lengths.

In order to simulate the mixed noises in the real environment, the Noisex-92 noise database was added to the original speech database. A speech database of 36000 different Real background noise was established, including 6 types of noises,such as Gnoisegen noise, Pink noise, Factory floor noise 1, Factory floor noise 2, Babble noise and Volvo noise. The signal-to-noise ratios of noises added are respectively 0db, 5db, 10db, 15db and 20db.

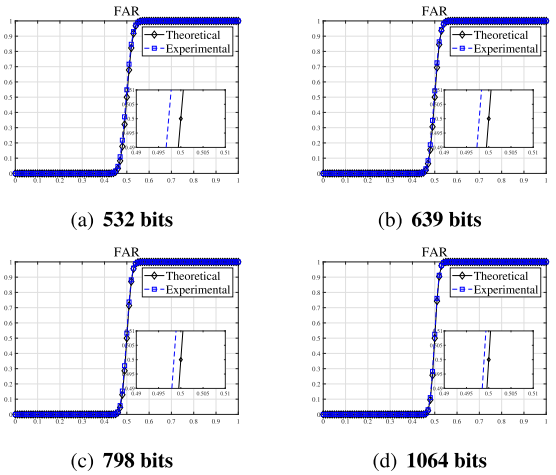
B. DISCRIMINATION TEST AND ANALYSIS

The BER of the perceptual hashing value of different speech contents basically obeys the normal distribution. 719400 BER datas are obtained by comparing the two perceptual hashing values of 1200 speech clips. In this paper, the BER normal distribution of hashing sequence length is shown in Fig 2. The better the BER normal distribution curve is, the better the randomicity and collision resistance performance of the perceptual hashing sequence are. The experimental results show that the probability distribution of BER values of different speechs has a high coincidence degree with the probability curve of the standard normal distribution, and the sequence length of 1064 bits selected in this paper is smaller in BER range than that of 532 bits, 639 bits and 798 bits. The effect is better when 1064 bits are selected for the hashing length.

According to the De Moivre-Laplace central limit theorem, the hamming distance is approximate obeying normal distribution ($\mu = p$, $\sigma = \sqrt{p(1-p)/N}$, N is the number of bits in a hashing sequence, p represents the probability of 0 or 1). In this paper, the length of the perceptual hashing sequence is 1064 bits, and the mean value and standard deviation of the theoretical normal distribution parameters are $\mu = 0.5000$ and $\sigma = 0.0153$. Table 2 describes the mean and standard deviation of the normal distribution of the theoretical and experimental values of different hashing sequences lengths.

TABLE 2. Normally distributed parameters of different hashing sequence lengths.

Parameter	Hashing sequence length	Theoretical value	Actual value
μ	532 bits	0.5000	0.4973
μ	639 bits	0.5000	0.4979
μ	798 bits	0.5000	0.4985
μ	1064 bits	0.5000	0.4989
σ	532 bits	0.0216	0.0222
σ	639 bits	0.0197	0.0202
σ	798 bits	0.0177	0.0181
σ	1064 bits	0.0153	0.0156



(a) **532 bits** (b) **639 bits**

(c) **798 bits** (d) **1064 bits**

FIGURE 3. FAR curves with different hashing sequence lengths.

Fig 3 shows FAR curves with different hashing sequence lengths.

It can be seen from Table 2 and Fig 3 that the values of μ and σ measured in this paper are very close to the parameters theoretically calculated. As can be seen from Fig 3, the actual curve is getting closer to the theoretical curve with the increase of hashing sequence length, indicating that the hashing sequence generated by this algorithm has good randomicity and collision resistance performance.

In order to evaluate the discrimination ability of the algorithm in this paper under different thresholds, the FAR is obtained from Equation (25). Table 3 compares FAR of different long hash sequence algorithms and different algorithms.

As shown in Table 3, the smaller the matching threshold τ is, the smaller the FAR value is. When the hashing sequence length is 1064 bits and the threshold $\tau = 0.35$ is set, about 1.31 of each 10^{21} speech clips are false accepted. As the length of the hashing sequence increases, FAR is decreasing, indicating that discrimination is increasing. Compared with the FAR of other hashing short sequences in this algorithm, the FAR of the hashing long sequence selected by this algorithm is the best, and it is also proved that the long hashing sequence has a high discrimination. When $\tau = 0.35$ occurs, 1.37 of each 10^{07} speech clips in Ref. [8] false accepted, 6.10 of each 10^{05} speech clips in Ref. [11] are false accepted,

TABLE 3. Hashing sequences algorithms of different lengths and the FAR value of different algorithms are compared.

τ	Hashing sequence length (The proposed algorithm)				Algorithm		
	532 bits	639 bits	798 bits	1064 bits	Ref.[8]	Ref.[11]	Ref.[12]
0.10	7.3267×10^{-72}	1.8371×10^{-86}	1.0030×10^{-107}	1.5898×10^{-142}	1.4855×10^{-43}	1.7668×10^{-28}	2.1420×10^{-47}
0.20	3.6414×10^{-41}	2.0984×10^{-49}	2.1302×10^{-61}	5.2575×10^{-81}	2.5886×10^{-25}	1.9604×10^{-16}	1.9220×10^{-27}
0.25	4.2363×10^{-29}	7.7001×10^{-35}	3.4383×10^{-43}	7.9066×10^{-57}	3.9877×10^{-18}	9.8689×10^{-12}	1.3754×10^{-19}
0.30	3.2232×10^{-19}	6.5252×10^{-23}	2.7980×10^{-28}	4.9348×10^{-37}	3.2038×10^{-12}	6.6409×10^{-08}	3.8423×10^{-13}
0.35	1.6435×10^{-11}	1.3053×10^{-13}	1.1739×10^{-16}	1.3075×10^{-21}	1.3692×10^{-07}	6.1048×10^{-05}	4.2761×10^{-08}

TABLE 4. ER of hashing sequences of different lengths.

Hashing sequence length	532 bits	639 bits	798 bits	1064 bits
ER	0.9801	0.9818	0.9838	0.9859

TABLE 5. ER of the different algorithms.

Algorithms	Ref.[5]	Ref.[6]	Ref.[8]	Ref.[11]
ER	0.5491	0.6794	0.9112	0.9732

and 4.28 of each 10^{08} speech clips in Ref. [12] are false accepted. In contrast, although Refs. [8], [11], [12] can also completely discriminate different speech clips, the algorithm in this paper has a much lower FAR than these algorithms. Compared with the hashing short sequence used in Refs. [8], [11], [12], the hashing long sequence in this paper has a great advantage in discrimination.

Entropy rate (ER) is a comprehensive evaluation index of discriminative perception hash algorithm, which mainly overcomes the shortcomings of the algorithm being susceptible to sequence size. The value of ER ranges from 0 to 1. The larger value, the stronger the discrimination ability, which can be calculated by Equations (27) and (28).

$$ER = -[c \log_2 c + (1-c)\log_2(1-c)] \tag{27}$$

$$c = \frac{1}{2}\left(\sqrt{\frac{|\sigma^2 - {\sigma_1}^2|}{\sigma^2 + {\sigma_1}^2}} + 1\right) \tag{28}$$

where σ and σ_1 are theoretical and experimental standard deviation of BERs respectively.

According to Table 4 and Table 5, with the increase of hashing sequence length, the ER of the algorithm in this paper is higher. When the hashing sequence length is 1064 bits, the ER of the algorithm is the highest, which proves that the hashing sequence has good discriminability. Compared with Refs. [5], [6], [8], [11], the ER of the algorithm in this paper is the highest, indicating that the discriminative effect of the algorithm in this paper is the best.

C. ROBUSTNESS TEST AND ANALYSIS

In order to evaluate the robustness of the proposed algorithm, the 14400 speech segments in the content preserving operations are extracted to generate hash sequences. According to the hashing sequence of the original speech and the speech after operation, the mean BER between the two is obtained.

Table 6 shows the content preserving operations that simulate a real environment. In this paper, the various BER of different hashing sequence lengths are shown in Table 7.

It can be obtained from Table 7: the mean BER of the whole algorithm in this paper does not exceed 0.1713, and the max BER does not exceed 0.2444. It is shown that the proposed algorithm in this paper holds better robustness for paper various content preserving operations. As the length of the sequence increases, the robustness of the operation of the other contents, except the echo, decreases. These robustness are only slightly reduced, which will not affect the overall robustness of the algorithm. At the same time, the average running time increases as the length of the hash sequence increases. In this paper, 1064 bits are used to balance the discriminability and robustness, and the overall effect is the best.

719400 BER datas are obtained by comparing of the two perceptual hashing values of 1200 speech clips. When the hashing length is set as 532bits, 639bits, 798bits and 1046bits, the FAR-FRR curve is obtained. The comparison results are shown in Fig 4.

As shown in Fig 4, the FRR and FAR curves of different hashing sequence lengths do not overlap, which can accurately discriminate the content preserving operations and the speech of different contents, indicating that the algorithm in this paper has good discrimination and robustness. The mean BER comparison results of this algorithm with Refs. [6], [8], [11] are shown in Table 8.

As can be seen from Table 8, the proposed algorithm is superior to other algorithms in volume, resampling, gaussian noise and MP3 compression for different content preserving operations. Therefore, the proposed algorithm has better robustness. Especially in MP3 compression, this algorithm has better performance than other algorithms. By comparing the algorithm in this paper with Ref. [11], it can be seen that the algorithm in this paper is better than Ref. [11] in resampling, noise, MP3 compression and other aspects. Therefore, this algorithm is suitable for complex communication

TABLE 6. Content preserving operations.

Operating means	Operation method	Abbreviation
Volume Adjustment 1	Volume down 50%	V.1
Volume Adjustment 2	Volume up 50%	V.2
FIR Filter	12 order FIR low-pass filtering, Cutoff frequency of 3.4 kHz	F.I.R
Butterworth Filter	12 order Butterworth low-pass filtering, Cutoff frequency of 3.4 kHz	B.W
Resampling 1	Sampling frequency decreased to 8 kHz, and then increased to 16 kHz	R.8→16
Resampling 2	Sampling frequency increased to 32 kHz, and then dropped to 16 kHz	R.32→16
Echo Addition 1	Superimposed attenuation 30%, delay 100 ms,initial strength were 10% of the echo	E.A1
Echo Addition 2	Superimposed attenuation 60%, delay 300 ms, initial strength were 25% of the echo	E.A2
Narrowband Noise 1	SNR=30 dB narrowband Gaussian noise, center frequency distribution in $0 \sim 4$ kHz	G.N1
Narrowband Noise 2	SNR=50 dB narrowband Gaussian noise, center frequency distribution in $0 \sim 4$ kHz	G.N2
MP3 Compression 1	Re-encoded as MP3, and then decoding recovery, the rate is 32 k	M.32
MP3 Compression 2	Re-encoded as MP3, and then decoding recovery, the rate is 128 k	M.128

TABLE 7. The BER mean value of different hashing sequences and run times are compared.

Hashing sequence length	**532 bits**				**639 bits**			
Operating means	Mean	Max	Variance	Time(s)	Mean	Max	Variance	Time(s)
V.1	0.0064	0.0583	0.0088	113	0.0066	0.0548	0.0086	130
V.2	0.0064	0.0508	0.0085	115	0.0064	0.0595	0.0083	129
F.I.R	0.0700	0.1485	0.0188	111	0.0737	0.1534	0.0184	129
B.W	0.0793	0.2030	0.0261	110	0.0830	0.2097	0.0260	130
R.8→16	0.0353	0.1823	0.0311	110	0.0364	0.1784	0.0315	132
R.32→16	0.0018	0.0395	0.0030	110	0.0019	0.0360	0.0028	130
E.A1	0.0663	0.1222	0.0130	111	0.0669	0.1174	0.0130	130
E.A2	0.1703	0.2350	0.0220	109	0.1713	0.2441	0.0226	133
G.N1	0.0839	0.2199	0.0405	109	0.0863	0.2269	0.0403	128
G.N2	0.0096	0.0658	0.0075	111	0.0101	0.0626	0.0070	129
M.32	0.0476	0.1316	0.0232	110	0.0493	0.1268	0.0246	132
M.128	0.0029	0.0320	0.0034	109	0.0030	0.0391	0.0034	131
Average time(s)			110.7				130.3	
Hashing sequence length	**798 bits**				**1064 bits**			
V.1	0.0067	0.0551	0.0085	162	0.0069	0.0583	0.0085	194
V.2	0.0064	0.0589	0.0081	162	0.0064	0.0573	0.0077	193
F.I.R	0.0788	0.1617	0.0185	169	0.0892	0.1673	0.0172	195
B.W	0.0876	0.2043	0.0257	172	0.0974	0.2096	0.0250	196
R.8→16	0.0373	0.1767	0.0320	162	0.0379	0.1945	0.0320	195
R.32→16	0.0021	0.0388	0.0030	161	0.0022	0.0367	0.0027	199
E.A1	0.0662	0.1140	0.0128	161	0.0650	0.1024	0.0122	197
E.A2	0.1713	0.2444	0.0226	162	0.1703	0.2406	0.0221	196
G.N1	0.0888	0.2043	0.0404	161	0.0909	0.2096	0.0403	195
G.N2	0.0106	0.0677	0.0072	168	0.0112	0.0677	0.0072	193
M.32	0.0507	0.1278	0.0258	164	0.0512	0.1372	0.0264	195
M.128	0.0032	0.0439	0.0035	164	0.0032	0.0367	0.0033	194
Average time(s)			164.0				195.2	

environment. Since this paper takes a long time to TD and construct hash sequences, the average time is lower than that in Ref. [11], which is more suitable for instant messaging. Compared with Ref. [8], although the robustness of the algorithm in this paper is slightly lower in terms of volume and echo, the robustness of the algorithm in this paper is far better than that in Ref. [8] in other aspects. Since the NMF with relatively complex structure is used in Ref. [8], the running time is also much higher than the algorithm in this paper. Compared with Ref. [6], the algorithm in this paper has better overall performance than Ref. [6].

Through pairwise comparison of the perceptual hashing values of 1200 speech clips, 719401 BER datas and FRR-FAR curves are obtained. The comparison results of different algorithms are shown in Fig 5.

As shown in Fig 5(a), the length of hashing sequence used in this paper is 1064 bits. The FRR-FAR curves without overlap is obtained through experiments,which indicates that the algorithm in this paper not only has good discrimination and robustness,but also can accurately identify the content retention operation and the speech of different contents. As shown in Fig 5(b), although the FRR-FAR curves obtained

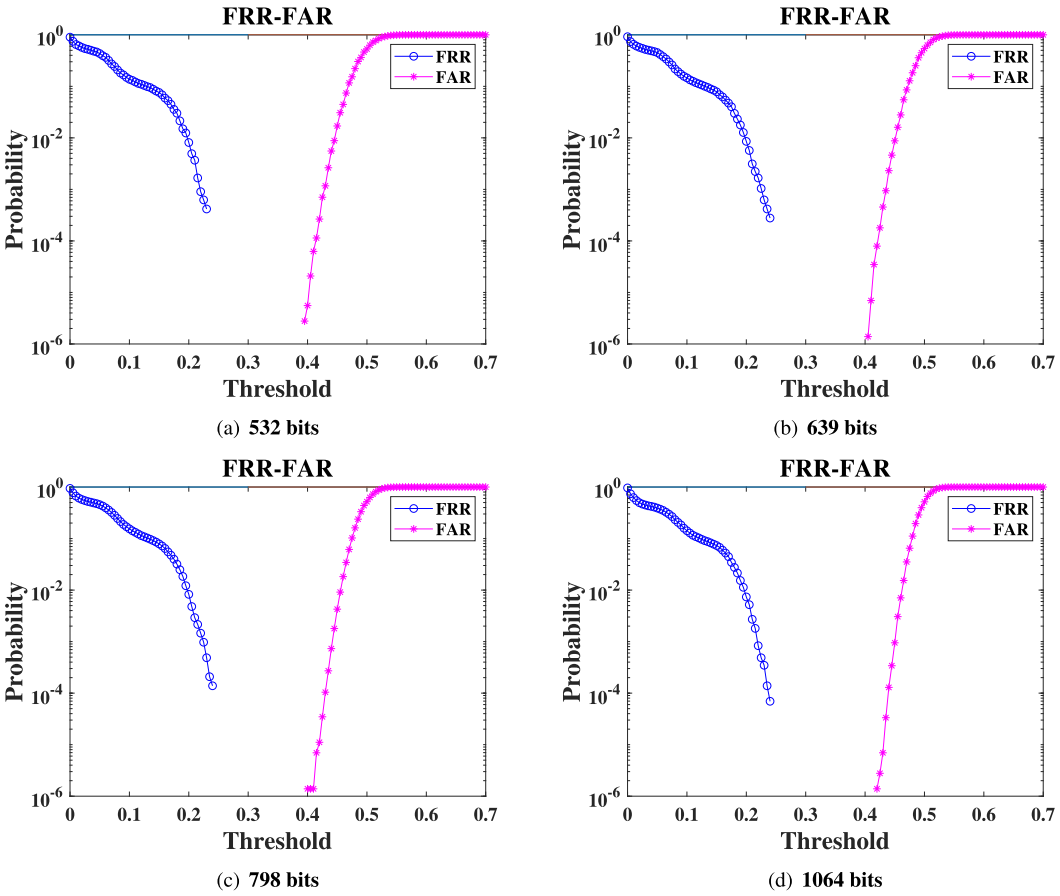


FIGURE 4. The FRR-FAR curves of different length hashing sequences.

in Ref. [11] do not overlap, the two curves are close to each other, which cannot well solve the problems of discrimination and robustness.The comparison results also show that the proposed algorithm is better than that in Ref. [11] in terms of discrimination and collision resistance performance. Comparing Fig 5(c) and Fig 5(d), FRR and FAR curves of the two algorithms intersect, reflecting that the discrimination and robustness cannot be solved well. It can be seen from Table 8 that this algorithm is superior to Ref. [8] and Ref. [6] in terms of discrimination and robustness.

D. PASSING RATE TEST AND ANALYSIS IN REAL NOISE ENVIRONMENT

In order to evaluate the robustness of the proposed algorithm to noise, the passing rate p_r is introduced.

$$p_r = \frac{T_A}{T_A + T_R + F_A} \tag{29}$$

where T_A is the number of speech clips correctly accepted by the system between the speech clips with the same perception content; T_R is the number of speech clips wrongfully rejected by the system; F_A is the number of speech clips wrongly accepted by the system between different speech clips of perceived content. The threshold τ is selected as the minimum BER of FAR curve. Different algorithms select different thresholds: the proposed algorithm is 0.4173, that in Ref. [8] is 0.3593, that in Ref. [11] is 0.3037, and that in Ref. [12] is 0.3677. Fig 6 shows the comparison of the passing rate between the proposed algorithm and that in Refs. [8], [11], [12] under six different noise environments.

As shown in Fig 6, the algorithm in this paper has strong robustness for Gaussian noise, Factory1 noise and Volvo noise. Especially for Volvo noise, the passing rate of different SNR reaches 100%. For all noises, the passing rate of the algorithm in this paper reaches 100% when the SNR is greater than 30db, which is also uncomparable in Refs. [8,11,12]. The stable feature values obtained by TD are robust to different noises. Compared with Ref. [11], the algorithm in this paper has a lower passing rate under the condition of Factory2 noise and Pink noise, indicating that the improved spectrum entropy has a strong robustness against these two kinds of noise. On the whole, the proposed algorithm has the best robustness. Compared with the algorithm in Refs. [8,12], the passing rate of the algorithm in this paper is much higher than that of the two algorithms no matter what kind of noise. Therefore, the proposed algorithm has better robustness for common noises and can meet the needs of speech matching in daily life.

TABLE 8. The BER mean value of different algorithm and run times are compared.

Algorithm	The proposed algorithm				Ref.[6]			
Operating means	Mean	Max	Variance	Time(s)	Mean	Max	Variance	Time(s)
V.1	0.0069	0.0583	0.0085	194	0.0457	0.2382	0.0305	224
V.2	0.0064	0.0573	0.0077	193	0.1042	0.3463	0.0885	234
F.I.R	0.0892	0.1673	0.0172	195	0.4129	0.6177	0.0660	231
B.W	0.0974	0.2096	0.0250	196	0.3939	0.5323	0.0468	234
R.8→16	0.0379	0.1945	0.0320	195	0.2105	0.1784	0.0488	235
R.32→16	0.0022	0.0367	0.0027	199	0.2301	0.5568	0.1815	231
E.A1	0.0650	0.1024	0.0122	197	0.0711	0.1811	0.0200	233
E.A2	0.1703	0.2406	0.0221	196	0.2418	0.4183	0.0617	227
G.N1	0.0909	0.2096	0.0403	195	0.3898	0.5402	0.0776	231
G.N2	0.0112	0.0677	0.0072	193	0.1875	0.5014	0.1634	234
M.32	0.0512	0.1372	0.0264	195	0.1388	0.2089	0.0211	230
M.128	0.0032	0.0367	0.0033	194	0.0312	0.1671	0.0256	236
Average time(s)			195.2				231.7	
Algorithm	**Ref.[8]**				**Ref.[11]**			
V.1	0.0034	0.0390	0.0041	353	0.0090	0.0630	0.0071	61
V.2	0.0011	0.0195	0.0020	356	0.0070	0.0333	0.0056	57
F.I.R	0.3386	0.5292	0.0550	357	0.0413	0.1000	0.0136	66
B.W	0.3240	0.5125	0.0540	371	0.0784	0.1593	0.0262	65
R.8→16	0.0747	0.2145	0.0583	370	0.0529	0.1185	0.0271	66
R.32→16	0.0043	0.0613	0.0051	371	0.0048	0.0444	0.0061	65
E.A1	0.0402	0.1031	0.0113	352	0.0315	0.0778	0.0120	72
E.A2	0.1184	0.1922	0.0205	375	0.1171	0.2000	0.0231	66
G.N1	0.1312	0.2563	0.0407	374	0.1108	0.2333	0.0357	74
G.N2	0.0338	0.1198	0.0218	373	0.0391	0.1259	0.0207	71
M.32	0.1620	0.5905	0.2017	374	0.2535	0.3370	0.0246	75
M.128	0.0070	0.0585	0.0067	383	0.2325	0.3148	0.0243	75
Average time(s)			367.4				67.8	

TABLE 9. Efficiency of the different algorithms.

Algorithms	Hashing sequence length	Working frequency	Average time
The algorithm	532 bits	3.4 GHz	0.0450s
	639 bits	3.4 GHz	0.0533s
	798 bits	3.4 GHz	0.0667s
	1064 bits	3.4 GHz	0.0788s
Ref.[6]	360 bits	3.4 GHz	0.0925s
Ref.[8]	360 bits	3.4 GHz	0.1825s
Ref.[11]	266 bits	3.4 GHz	0.0250s
Ref.[13]	360 bits	3.4 GHz	0.0883s

E. EFFICIENCY TESTING AND ANALYSIS

Efficiency is a very important evaluation criterion in speech content authentication. To evaluate the efficiency of the algorithm in this paper, we need to randomly select 200 speech clips from the speech database, and then calculate the average running time. The same operating environment is adopted, and the speech clips is 4s. Table 9 shows the comparison results between the algorithm in this paper and the algorithm in Refs. [6], [8], [11], [13].

As shown in Table 9, as for the algorithm in this paper, with the increase of hashing sequence length, although the efficiency performance of the algorithm is decreasing, the difference is small, which meets the requirements of efficiency authentication. The length of hashing sequence in this paper is 1064 bits. Compared with the length of other hashing sequences in this paper, the timeliness is relatively low, but the discrimination is greatly improved. Compared with other algorithm, the efficiency of the algorithm in this paper is 1.1 times that in Ref. [13], 2.3 times that in Ref. [8], and 1.2 times that in Ref. [6]. However, compared with Ref. [11], the efficiency of Ref. [11] is 3.2 times of the algorithm in this paper. Since this paper adopts hashing long sequence and tensor decomposition, the complexity is much higher and the average running time is slightly slower than Ref. [11]. Because NMF with large computation and long running time was used in Refs. [6,8], the efficiency performance is lower than the algorithm in this paper. Although the length of hashing sequence in this paper is 4 times of that in Ref. [11] and 3 times of that in Refs. [6,8,13], the algorithm in this paper performs very well in the efficiency performance and can meet the requirements of efficiency authentication.

F. DISCUSSION

We compared the authentication performance of the proposed algorithm with perceptual hashing algorithm based on improved spectral entropy, and perceptual hashing based on NMF and MDCT coefficients. The authentication performance of different algorithms is evaluated in detail. The main highlights of our proposed algorithm are summarized below:

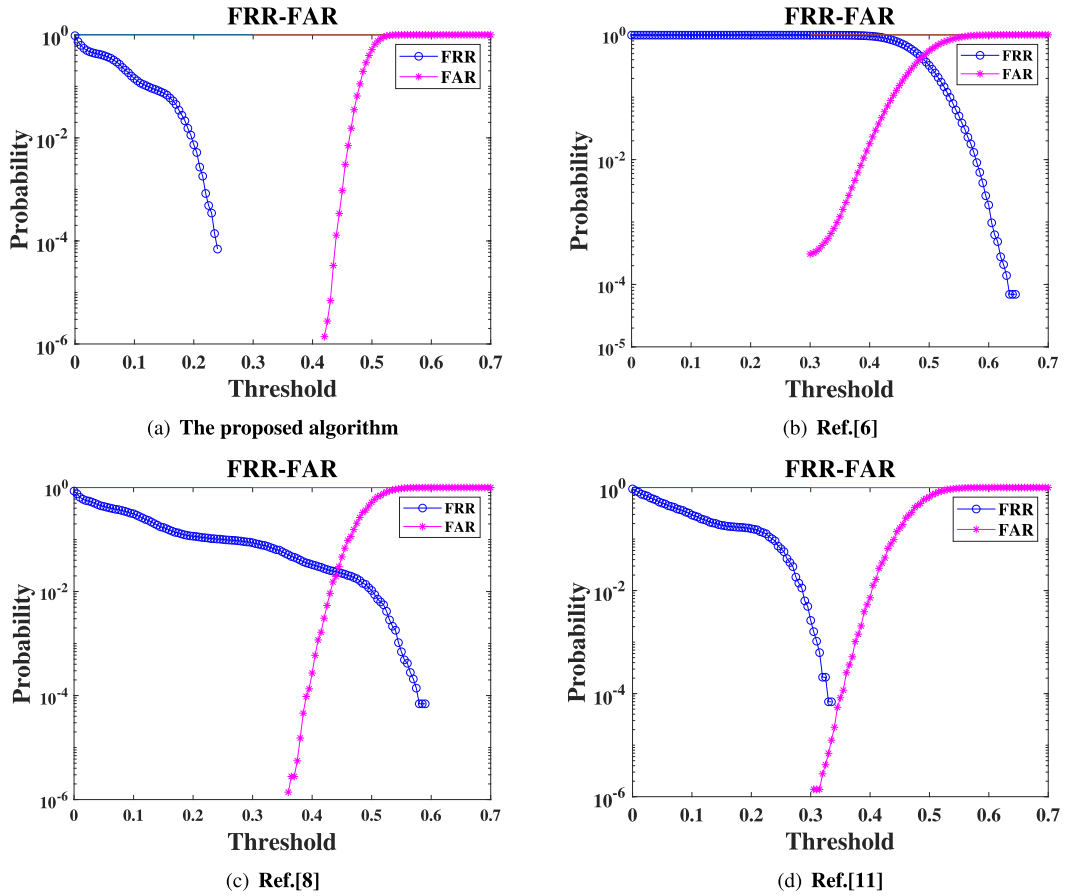


FIGURE 5. The FRR-FAR curves of different algorithm.

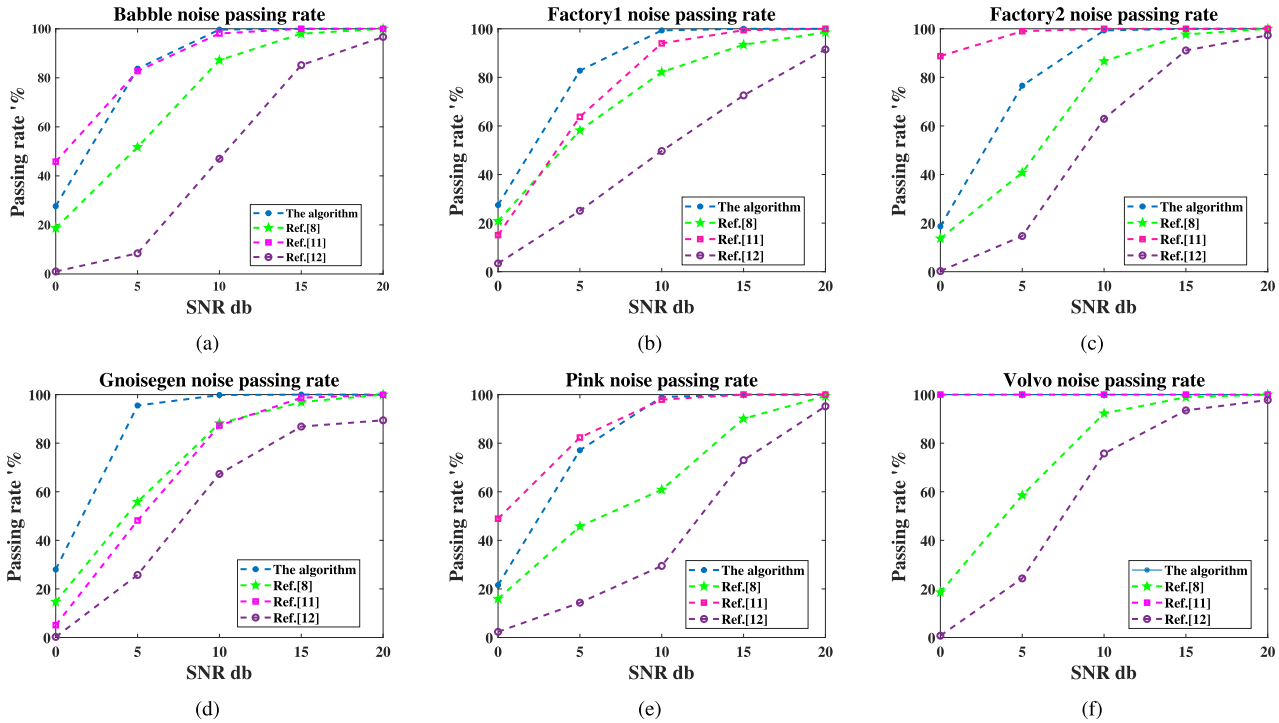


FIGURE 6. Comparison of passing rates of different algorithms under different noises.

1. This algorithm not only improves the length of hash sequence and the recognition rate of algorithm, but also can generate inconsistent hash sequence from a large number of speech data.

2. The features extracted in this paper have strong anti-interference performance, especially various noises with low signal-to-noise ratio.

3. Compared with the existing authentication algorithms, the efficiency of the algorithm in this paper has achieved good results.

According to the advantages of the proposed system, it can be deployed in real speech authentication.

The main disadvantages of the proposed algorithm are:

- The algorithm lacks security and is easy to cause information leakage.
- In the case of speech tampering, this proposed algorithm cannot tamper detection and localization, which is a major flaw in this algorithm.

V. CONCLUSION

This paper presents a long sequence perceptual hashing authentication algorithm based on CQT and TD. This algorithm has good comprehensive performance and solves the existing problems of speech authentication algorithms. The following conclusions can be obtained through the experimental analysis: ***A.*** The algorithm in this paper adopts a long hashing sequence with high discriminability. For different speech clips, different hashing sequences are generated, which effectively reduces the probability that different speech clips are confirmed as the same speech clips and improves the authentication rate of the algorithm. ***B.*** The algorithm in this paper has strong robustness for content preserving operations, especially in the case of resampling, low-pass filtering, noise and MP3 compression, which indicates that the algorithm in this paper is suitable for signal transmission in complex environments. ***C.*** From the perspective of overall performance, when the hashing sequence length of 1064 bits is selected by the algorithm in this paper, it not only gives consideration to the discrimination and robustness, but also has highly efficiency performance, which meets the requirements of speech authentication in the real-time communication environment.

Because the hashing sequence is too long, which will cause the waste of storage space resources and increase the running time, the hashing sequence length of the algorithm in this paper needs to be further optimized, and the security of the algorithm in an open environment needs to be further solved. The proposed algorithm also needs to address location detection in the case of speech tampering.

REFERENCES

[1] M. Hammad, G. Luo, and K. Wang, "Cancelable biometric authentication system based on ECG," *Multimedia Tools Appl.*, vol. 78, no. 2, pp. 1857–1887, Jan. 2019.

[2] Z. Su, G. Zhang, F. Yue, L. Chang, J. Jiang, and X. Yao, "SNR-constrained heuristics for optimizing the scaling parameter of robust audio watermarking," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2631–2644, Oct. 2018.

[3] D. Renza, D. M. Ballesteros L., and C. Lemus, "Authenticity verification of audio signals based on fragile watermarking for audio forensics," *Expert Syst. Appl.*, vol. 91, pp. 211–222, Jan. 2018.

[4] X. M. Niu and Y. H. Jiao, "Overview of perceptual hash," *J. Electionics*, vol. 36, no. 7, pp. 1405–1411, 2008.

[5] N. Chen, W. Wan, and H.-D. Xiao, "Robust audio hashing based on discrete-wavelet-transform and non-negative matrix factorisation," *IET Commun.*, vol. 4, no. 14, pp. 1722–1731, 2010.

[6] N. Chen, "Robust speech hash function," *ETRI J.*, vol. 32, no. 2, pp. 345–347, Apr. 2010.

[7] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1616–1629, 2020.

[8] J. Li, Y. Jing, and H. Wang, "Audio perceptual hashing based on NMF and MDCT coefficients," (in Chinese), *Chin. J. Electron.*, vol. 24, no. 3, pp. 579–588, Jul. 2015.

[9] N. Chen, J. Zhu, H. D. Xiao, W. H. Yuan, Y. Wang, and J. J. Lin, "Robust audio hashing scheme based on cochleagram and cross recurrence analysis," *Electron. Lett.*, vol. 49, no. 1, pp. 7–8, Jan. 2013.

[10] 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," *J. Inf. Hiding Multimedia Signal Process.*, vol. 6, no. 2, pp. 311–322, 2015.

[11] Q.-Y. Zhang, W.-J. Hu, Y.-B. Huang, and S.-B. Qiao, "An efficient perceptual hashing based on improved spectral entropy for speech authentication," *Multimedia Tools Appl.*, vol. 77, no. 2, pp. 1555–1581, Jan. 2018.

[12] Q. Y. Zhang, S. B. Qiao, T. Zhang, and Y. B. Huang, "Multi-format audio perception hashing algorithms based on zero ratio," (in Chinese), *J. Huazhong Univ. Sci. Technol., Natural Sci. Ed.*, vol. 45, no. 9, pp. 33–38, 2017.

[13] Q.-Y. Zhang, S.-B. Qiao, Y.-B. Huang, and T. Zhang, "A high-performance speech perceptual hashing authentication algorithm based on discrete wavelet transform and measurement matrix," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 21653–21669, Aug. 2018.

[14] Y. B. Huang, Q. Y. Zhang, Z. T. Yuan, and Z. P. Yang, "The hash algorithm of speech perception based on the integration of adaptive MFCC and LPCC," (in Chinese), *J. Huazhong Univ. Sci. Technol., Natural Sci. Ed.*, vol. 43, no. 2, pp. 124–128, 2015.

[15] K. Ding, S. Chen, and F. Meng, "A novel perceptual hash algorithm for multispectral image authentication," *Algorithms*, vol. 11, no. 1, p. 6, Jan. 2018.

[16] K. O'Hanlon and M. B. Sandler, "Comparing CQT and reassignment based chroma features for template-based automatic chord recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 860–864.

[17] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, "TimbreTron: A WaveNet(CycleGAN(CQT(Audio))) pipeline for musical timbre transfer," 2018, *arXiv:1811.09620*. [Online]. Available: http://arxiv.org/abs/1811.09620

[18] Z. Shi, H. Lin, L. Liu, R. Liu, and J. Han, "Is CQT more suitable for monaural speech separation than STFT? An empirical study," 2019, *arXiv:1902.00631*. [Online]. Available: http://arxiv.org/abs/1902.00631

[19] G. Zhao, B. Tu, H. Fei, N. Li, and X. Yang, "Spatial-spectral classification of hyperspectral image via group tensor decomposition," *Neurocomputing*, vol. 316, pp. 68–77, Nov. 2018.

[20] Y. Huang and Q. Zhang, "Strong robustness hash algorithm of speech perception based on tensor decomposition model," *J. Softw. Eng.*, vol. 11, no. 1, pp. 22–31, Dec. 2017.

[21] F. Cong, Q.-H. Lin, L.-D. Kuang, X.-F. Gong, P. Astikainen, and T. Ristaniemi, "Tensor decomposition of EEG signals: A brief review," *J. Neurosci. Methods*, vol. 248, pp. 59–69, Jun. 2015.

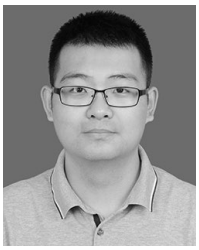
[22] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram and phoneme embedding," in *Proc. Interspeech*, Aug. 2018, pp. 3688–3692.

[23] J. Chatterjee, V. Mukesh, H.-H. Hsu, G. Vyas, and Z. Liu, "Speech emotion recognition using cross-correlation and acoustic features," in *Proc. IEEE 16th Int. Conf. Dependable, Autonomic Secure Comput., 16th Int. Conf. Pervas. Intell. Comput., 4th Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Aug. 2018, pp. 243–249.

[24] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and support vector machine," in *Proc. Int. Conf. Autom. Control Dyn. Optim. Technol. (ICACDOT)*, Sep. 2016, pp. 9–12.

[25] Y. Lu, F. Hu, and X. Li, "Towards the security of big data: Building a scalable hash scheme for big graph," in *Information Technology-New Generations*. Cham, Switzerland: Springer, 2018, pp. 235–243.

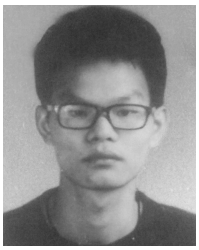
[26] Y. Jiang, C. Wu, K. Deng, and Y. Wu, "An audio fingerprinting extraction algorithm based on lifting wavelet packet and improved optimal-basis selection," *Multimedia Tools Appl.*, vol. 78, no. 21, pp. 30011–30025, Nov. 2019.
[27] Y. Luo and Y. Mao, "Single-channel speech enhancement based on multi-band spectrogram-rearranged RPCA," *Electron. Lett.*, vol. 55, no. 7, pp. 415–417, Apr. 2019.
[28] V. Z. Këpuska and H. A. Elharati, "Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden Markov model classifier in noisy conditions," *J. Comput. Commun.*, vol. 03, no. 6, pp. 1–9, 2015.
[29] C. Kumar, F. ur Rehman, S. Kumar, A. Mehmood, and G. Shabir, "Analysis of MFCC and BFCC in a speaker identification system," in *Proc. Int. Conf. Comput., Math. Eng. Technol. (iCoMET)*, Mar. 2018, pp. 1–5.
[30] T. Afouras, J. S. Chung, and A. Senior, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[31] C. Qin, Y. Hu, H. Yao, X. Duan, and L. Gao, "Perceptual image hashing based on Weber local binary pattern and color angle representation," *IEEE Access*, vol. 7, pp. 45460–45471, 2019.
[32] W. Luo, H. Li, Q. Yan, R. Yang, and J. Huang, "Improved audio steganalytic feature and its applications in audio forensics," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 2, pp. 1–14, Apr. 2018.
[33] F. Tao and W. Qian, "Image hash authentication algorithm for orthogonal moments of fractional order chaotic scrambling coupling hyper-complex number," *Measurement*, vol. 134, pp. 866–873, Feb. 2019.
[34] Y. Cao, H. Qi, J. Gui, S. Li, and K. Li, "General distributed hash learning on image descriptors for k-nearest neighbor search," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 750–754, May 2019.
[35] J. Wang, L. Y. Zhang, J. Chen, G. Hua, Y. Zhang, and Y. Xiang, "Compressed sensing based selective encryption with data hiding capability," *IEEE Trans. Ind. Informat.*, vol. 15, no. 12, pp. 6560–6571, Dec. 2019.
[36] Z. Ali, M. Imran, S. McClean, N. Khan, and M. Shoaib, "Protection of records and data authentication based on secret shares and watermarking," *Future Gener. Comput. Syst.*, vol. 98, pp. 331–341, Sep. 2019.
[37] A. Abozaid, A. Haggag, H. Kasban, and M. Eltokhy, "Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion," *Multimedia Tools Appl.*, vol. 78, no. 12, pp. 16345–16361, Jun. 2019.
[38] M. Hammad, Y. Liu, and K. Wang, "Multimodal biometric authentication systems using convolution neural network based on different level fusion of ECG and fingerprint," *IEEE Access*, vol. 7, pp. 26527–26542, 2019.
[39] M. Hammad and K. Wang, "Parallel score fusion of ECG and fingerprint for human authentication based on convolution neural network," *Comput. Secur.*, vol. 81, pp. 107–122, Mar. 2019.



YIBO HUANG received the Ph.D. degree from the Lanzhou University of Technology, in 2015. He is currently working as an Associate Professor with the College of Physics and Electronic Engineering, Northwest Normal University. His main research interests include multimedia information processing, information security, and speech recognition.

HEXIANG HOU received the B.S. degree in communication engineering from Dezhou University, Shandong, China, in 2018, where he is currently pursuing the M.S. degree in electronic and communications engineering. His research interests include audio signal processing and application, and multimedia authentication.

YONG WANG received the B.S. degree from the Henan Institute of Science and Technology, Henan, China, in 2017. His research interests include audio signal processing and application, and multimedia authentication techniques.



YUAN ZHANG received the B.S. degree in electronic information engineering from the Wuhan Institute of Technology, Hubei, China, in 2017, where he is currently pursuing the M.S. degree in electronic and communications engineering. His research interests include audio signal processing and application, and multimedia authentication.

MANHONG FAN received the M.Sc. degree in circuits and system from Northwest Normal University, Lanzhou, China, in 2012. His research interest includes computer measurement and control.