Perceptual Hash Secure Speech Authentication System Based on an Improved Chained DES in the Cloud Environment

Yi-Bo Huang[1], Xiang-Rong Pu[1], and Qiu-Yu Zhang[2]
(Corresponding author: Yi-Bo Huang)
College of Physics and Electronic Engineering, Northwest Normal University[1]
Anning District, Lan Zhou, China
Email: huang_yibo@foxmail.com
School of Computer and Communication, Lanzhou University of Technology[2]
Qilihe District, Lan Zhou, China
(Received Mar. 18, 2022; Revised and Accepted June 29, 2022; First Online July 3, 2022)

Abstract

Cloud speech authentication is a technique that can determine the integrity of speech in cloud computing, bringing many conveniences to speech users. However, the authentication process of cloud speech authentication systems faces some security issues that need to be addressed, such as the leakage of speech features and hash sequences in the authentication system and the detection of authentication content after a small-scale tampering attack. In this paper, a novel and effective speech authentication algorithm are proposed. First, NTEO and improved uniform sub-band and frequency band variance features are extracted and fused to construct a feature vector that can be binarized to generate a hash sequence. Then, the hash sequence is encrypted by an improved chained DES algorithm to generate a hash index, and a cloud-based encrypted hash database is established. Next, a hash template is constructed to establish a one-to-one correspondence between the perceptual and encrypted hash. Finally, match the authenticated cloud encrypted hash and the hash to be authenticated after the hash template by the Hamming distance algorithm. For the authenticated speech, a small range of tampering attacks and content preserving operations are first distinguished before the detection algorithm is used for small range detection and positioning. The experimental results show that the proposed method can be authenticated directly in the encrypted hash, effectively balancing discrimination and robustness, accurately locating the tampered regions, and detecting malicious substitutions and muting attacks. In addition, the proposed encryption algorithm, which introduces a random mechanism for key control, encrypts the hash sequence in a "one key at a time" manner, removes the correlation between the hash and the input, and improves the security of the hash in the authentication system. Furthermore, the key space is large enough to resist exhaustive attacks.

Keywords: Cloud Computing; Cloud-based Speech Authentication; Fusion Features; Hash Security; Improved Chained DES Algorithm; Small-Range Tamper Detection Positioning

1 Introduction

With the rapid development of big data and cloud computing, speech recognition authentication systems in cloud environment have been widely used in many fields of life. However, there is a serious leakage problem when transmitting and storing unprotected data, and data loss leads to a huge threat to users' sensitive information. In the face of existing multi-type and deep similarity attacks and forgery attacks, the traditional speech authentication method which constructs hash sequence after extracting speech features and directly uploads cloud authentication is easy to cause leakage of speech features and hash sequence, and it's no longer meet the current authentication needs. In addition, due to the rapid development of complex speech editing software, illegal copying, tampering and forgery of speech content are increasing, in this situation, accurate region tampering detection and positioning techniques are required for speech authentication systems. Therefore, while balancing discrimination and robustness, the study of the security of speech recognition authentication systems in cloud environment has become an important issue.

Facing the diversification of forms and content of Internet data, many features such as fingerprint [13], face shape [12], iris [7], DNA [10], palm print [23], signature [1] and gait [8] are used by more and more authentication systems. Existing speech authentication algo-

rithms are mainly focused on the optimization for robustness and real-time, with less consideration for discrimination and security. Currently, the extraction of speech signal features includes short-time interrelationship coefficients, acoustic features [14], linear predictive cepstral coefficients [2], modulated complex superposition transform (MCLT) [22], resonance peaks [11], multi-feature fusion [3], fundamental frequencies, etc., as well as the fusion between various features. Li *et al.*[13] proposed to improve the cosine transformation and combine the non-negative matrix decomposition to construct the hash sequence. The content preserving operation in different environments of this algorithm has good robustness compared with other algorithms, but it fails to balance the discrimination and robustness, and its efficiency is relatively low. Zhang *et al.* [16] proposed to fuse linear predictive minimum mean square variances and spectral entropy for hash construction, the algorithm has good efficiency but poor robustness in terms of discrimination and partial content preserving operations. Lu *et al.* [19] proposed to using trinary to construct a hash sequence instead of the traditional binary one, demonstrating the flexibility of hash construction, and the algorithm not only has good robustness to content preserving operations, but also has high efficiency.

Existing encryption algorithms include traditional encryption methods and encryption methods based on chaotic systems. Their commonality is direct encryption of speech signals, which leads to big calculation, time-consuming and large storage space. The chaotic system-based encryption method leads to a complete loss of speech features due to multiple rounds of scrambling, spreading and replacement operations, as a result, tampering detection and data recovery cannot be carried out directly. a speech perceptual hashing algorithm with LSP paramenterization as a perceptual feature was proposed in [6], whose hash structure relies on a key. Although the algorithm has good collision resistance and robustness to content preserving operations, its security needs to be improved. Zhang *et al.* [20] used QR decomposition of a given rotation for the wavelet packet coefficient matrix to extract the speech feature parameters and then constructs the perceptual hash sequence, but they didn't consider the security of the hash sequences. Zhang *et al.* [21] proposed to first dislocate the original speech database using Henon mapping to construct the encrypted speech database, and then extract the uniform subspectral variance to construct the hash sequence, but they didn't consider the security of the stored hash values in the cloud, meanwhile, its efficiency is slow and the data volume is large.

In summary, existing cloud-based speech perceptual hash authentication algorithms do not have excessive security measures for the hash value, which in turn makes the algorithm less secure. Most authentication algorithms are only optimized independently for robustness, discrimination, and privacy, without balancing the overall performance of the algorithms. The small-scale tampering attack as a malicious attack is not considered as a content preserving operation and the detection and positioning accuracy of tampering data is not high. To address these problems, this paper proposes a cloud computing perceptual hash speech authentication algorithm based on an improved chained DES to achieve a good balance of the algorithm's performance. The use of long hash sequences can effectively improve the algorithm's discrimination and the use of fusion features can well enhance the robustness of the algorithm and improve the authentication pass rate under different content preserving operations. The use of improved chained DES encryption algorithm, which introduces the random mechanism of key control to encrypt the hash sequence in a way of "one key at a time", can well remove the correlation between hash value and input, and improve the security of hash value in the authentication system.

2 Related Theory

2.1 Construction of the Encrypted Hash Database

The traditional DES encryption algorithm has a single initial key and the use order of extended subkeys is simple, which makes the encrypted data vulnerable to exhaustive attacks. In order to solve this problem, this paper improves the DES encryption algorithm, which uses different keys for data encryption and re-encodes the keys to achieve the effect of " One key at a time ", which greatly improves the security of encrypted data.

Encryption process of improved chained DES:

1) A 3D random number matrix $Q^{3\times N}$ is first generated from the Lorenz chaos measurement matrix $\Phi_{Lorenz}^{M\times N}$.

$$\Phi_{Lorenz}^{M\times N}=\begin{bmatrix}\xi_{\Gamma,\Psi,\Upsilon}^{1,1} & \xi_{\Gamma,\Psi,\Upsilon}^{1,2} & \cdots & \xi_{\Gamma,\Psi,\Upsilon}^{1,N}\\ \xi_{\Gamma,\Psi,\Upsilon}^{2,1} & \xi_{\Gamma,\Psi,\Upsilon}^{2,2} & \cdots & \xi_{\Gamma,\Psi,\Upsilon}^{2,N}\\ \vdots & \vdots & \ddots & \vdots\\ \xi_{\Gamma,\Psi,\Upsilon}^{M,1} & \xi_{\Gamma,\Psi,\Upsilon}^{M,2} & \cdots & \xi_{\Gamma,\Psi,\Upsilon}^{M,N}\end{bmatrix} \tag{1}$$

where $\xi_{\Gamma,\Psi,\Upsilon}$ represents the mapping function of the Lorentzian chaos measurement matrix with the following expression.

$$\begin{cases}\Gamma^{'}=\lambda^{*}(\Psi(i)-\Gamma(i))\\ \Psi^{'}=\Gamma(i)^{*}(\beta-\Upsilon(i)-\Psi(i))\\ \Upsilon^{'}=-\alpha^{*}\Upsilon(i)+\Gamma(i)^{*}\Psi(i)\end{cases} \tag{2}$$

where the initial values of $\alpha=8/3,\beta=28,\lambda=10$, $\Gamma(i),\Psi(i),\Upsilon(i)$ are 0,1,0 respectively.

2) The matrix $Q^{3\times N}$ is arranged in disorder and quantified to generate a sequence of random numbers $\Phi=\{\Phi(i)\mid i=1,2,\cdots N\}$, where $N=12$, set Φ the length of the initial key $Q(i)$ of group 12 of $64bits$ and the value interval length are $768bits$.

3) Firstly, according to the hash construction system, the speech data will be generated based on the authentication algorithm system as a binary hash sequence H. Then the whole hash sequence H is grouped to obtain the grouped hash sequence $\{H_n \mid n = 1, 2, \cdots, \kappa\}$, where the length of H_n is ϑbit. The first set of hash sequences H_1 is triggered by the initial key $Q(1)$, and then the DES cryptographer is executed to generate the first set of ciphertext hash E_1, and the remaining set of hash sequences is triggered by the result of the XOR operation between the ciphertext hash of the previous set and the key of this set, and then the DES cryptographer is executed to generate the ciphertext hash E_n, i.e.

$$E_n = \begin{cases} des(H_1, Q(n)) & n = 1 \\ des(H_n, Q(n) \oplus E_{n-1}) & n = 2, 3, \cdots, \kappa \end{cases} \quad (3)$$

4) The grouped ciphertext hash is stitched together to obtain the total ciphertext hash $E_n = [E_1, E_2, \cdots, E_n]$ and the encryption process is shown in Figure 1.

2.2 Speech Feature Extraction and Fusion Algorithm

In order to better adapt to the special complex speech environment, improve the robustness while ensuring the discrimination, and make the comprehensive performance of the algorithm reach a balance point, fusion features are proposed in this paper. Fuse feature parameters NTEO energy operator and improved uniform sub-band frequency band variance, among them:

2.2.1 NTEO Energy Operator

In this paper, an improved Teager energy algorithm is proposed, which introduces a resolution parameter j, replaces the instantaneous energy of the discrete-time signal $x(n)$ by 3 points adjacent to each other with 3 points separated by j samples before and after, which effectively improves the robustness of the system.

Suppose the time series of the speech signal is $x(n)$, and the i-th frame of the speech signal $x(m)$ is obtained after framing and windowing, and the length of the frame is N.The discrete time signal "energy" is defined by Equation (4).

$$\Psi'[\alpha_i(\beta)] = [\alpha_i(\beta)]^2 - \alpha_i(\beta+1)\alpha_i(\beta-1) \quad \beta = 1, 2, \cdots, N \quad (4)$$

The definition of the discrete-time signal after the introduction of the resolution parameter is given in Equation (5).

$$\Psi'[\alpha_i(\delta)] = \alpha_i^2(\delta) - \alpha_i(\delta-j)\alpha_i(\delta+j) \quad \delta = 1, 2, \cdots, N \quad (5)$$

Let f_c be the fundamental frequency and f_s be the sampling frequency. From the constraint $f_c/f_s < \frac{1}{8j}$, we know that j is upper bounded by $j < \frac{f_s}{8f_c}$.

2.2.2 Improved Uniform Sub-band Frequency Band Variance

Since the traditional uniform sub-band variance only reflects the large difference in characteristics in the frequency domain, and the sensitivity of speech data is not enough for low signal-to-noise ratio, this paper combines the advantages of high resolution and high precision of Constant Q Transform (CQT) to propose a new type of speech parameter—improved uniform sub-band band variance, which improves the robustness of the algorithm. The improved formula for calculating the variance of the uniform sub-band frequency band is as follows:

The time domain waveform of the speech signal is $x(n)$, and the speech signal of frame i obtained after windowing and framing processing is $x_i(m)$,then $x_i(m)$ meets the requirements.

$$x_i(m) = \omega(m) * x(iT + m) \quad 1 \le m \le M \quad (6)$$

Where, $\omega(m)$ is the window function, $i = 1, 2, \cdots, N$, M is the frame length, T is the frame shift length, and the total number of frames is N.

CQT of $x_i(m)$ is defined as follows:

$$X^{CQ}(k, n) = \sum_{j=n-[N_k/2]}^{n+[N_k/2]} x(m) a_k^*(m - n + N_k/2) \quad (7)$$

Where $k = 1, 2, \cdots, K$ is the frequency band serial number, and $a_k^*(n)$ is the conjugate complex number of the primary function $a_k(n)$, $N_k = \frac{f_s}{f_k} Q$ (Q represents the ratio of center frequency to bandwidth, which is a constant independent of k) is a variable window length, and $[\cdot]$ represents floor.

The primary function $a_k(n)$ is defined as follows:

$$a_k(n) = \frac{1}{C}\left(\frac{n}{N_k}\right) \exp[i(2\pi n \frac{f_k}{f_s} + \Phi_k)] \quad (8)$$

Where f_s is the frequency of sample, Φ_k is the phase-steeping, and C is the given scale factor.

For the CQT spectral signal $X^{CQ}(k, n)$, find the uniform sub-band band variance:

Step 1. $(\frac{N}{2} + 1)$ spectral line exists in the positive frequency domain, and this $(\frac{N}{2} + 1)$ amplitude spectral line is divided into q sub-bands, each containing $s = fix[(\frac{N}{2} + 1)/q]$ spectral lines (where $fix[\cdot]$ means taking its integer part), then the sub-bands are formed.

$$XX_i(j) = \sum_{k=1+(m-1)s}^{1+(m-1)s+(s-1)} |X_i(k)| 1 \le j \le q \quad (9)$$

Step 2. Divide the sub-band into r subsets and find the variance of each subset

$$XX_i = \left[XX_i^{(1)}, XX_i^{(2)}, \cdots, XX_i^{(r)}\right] \quad (10)$$

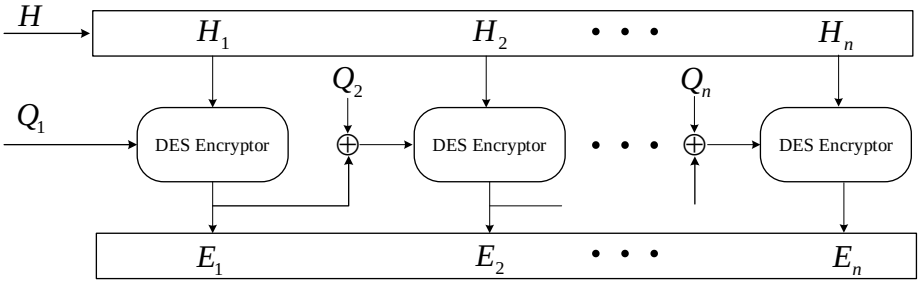


Figure 1: Improved chained DES encryption process

Step 3. The first subset $XX_i^{(1)}$ calculates the mean value as

$$E_i^{(1)} = \frac{r}{q}\sum_{t=1}^{q/r} XX_I^{(1)}(t) \tag{11}$$

The definition of the obtained variance is

$$D_i^{(1)} = \frac{1}{q/r-1}\sum_{t=1}^{q/r}\left[XX_i^{(1)}(t) - E_i^{(1)}\right]^2 \tag{12}$$

Step 4. Within each sub-band there are q-line lines after the original CQT, so it is called a uniform sub-band, that is, each sub-band is equal bandwidth. The subsets used are also the same number of sub-bands, and the variance of all subsets per frame is $D_{r,i} = \left[D_i^{(1)}, D_i^{(2)}, \cdots, D_i^{(r)}\right]$.

3 Speech Perceptual Hashing Authentication Scheme

3.1 Registration Terminal

Speech perceptual hashing authentication scheme is shown in Figure 2.

Step 1. Pre-processing First, the input signal $x(n)$ is pre-emphasis, framing, and processed with the Hamming windowing to obtain a speech signal $x(m) = \{x_i(m)\,|i = 1, 2, \cdots, N, m = 1, 2, \cdots, M\}$ with a total frame count of N, where i represents the i-th frame after the split frame.

Step 2. Feature extraction The Pre-processing signal $x(m)$ constructs the uniform sub-band band variance of the frequency domain signal and calculates the NTEO energy of the time domain signal, and fuses the time domain signal and the frequency domain signal.

1) Uniform sub-band frequency band variance: First, according to Equation (7), the time domain signal $x(m)$ is CQT to obtain the frequency signal $X^{CQ}(k,n)$, and then the frequency domain signal is divided into uniform sub-bands to obtain the frequency domain sub-band matrix $B = \{B_i(m)\,|i = 1, 2, \cdots, N; m = 1, 2, \cdots, q\}$ of the signal. After dividing the sub-band matrix, the sub-band set matrix $C = \{C_i(m)\,|i = 1, 2, \cdots, N; m = 1, 2, \cdots, r\}$, q is an integer multiple of r.

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

The frequency band variance of sub-band separation per frame is calculated by Equations (11) & (12).

$$D = \{D(i)\,|i = 1, 2, \cdots, N\} \tag{13}$$

2) NTEO Energy: According to Equation (5), the Pre-processing speech signal $x(m)$ is subjected to NTEO variation to obtain $\Psi = \{\Psi_m(i)\,|i = 1, 2, \cdots N; m = 1, 2, \cdots, L\}$, and then the average NTEO energy of each speech signal frame is calculated to obtain the feature vector $\hat{\Psi} = \left\{\hat{\Psi}(i)\,|i = 1, 2, \cdots N\right\}$.

3) Constructing time-frequency fusion features: The time-frequency fused feature vector $H = \{H(i)\,|i = 1, 2, \cdots, N\}$ is obtained by fusing the time domain feature parameter $\hat{\Psi}$ with the frequency domain feature parameter D by Equation (14).

$$H(i) = \sqrt{\hat{\Psi}(i) * D(i)} \tag{14}$$

Step 3. Hash sequence construction Binary hash construction is performed on the fusion parameter H.The hash sequence of the speech signal $x(n)$ is

$h = \{h(i)|i = 1, 2, \cdots, N\}$, and the perceptual hash sequence $h(1)$ is set to 0 during the hash construction, then the hash sequence construction formula is as follows.

$$h(i) = \begin{cases} 1 & if(H(i+1) > H(i)) \\ 0 & else \end{cases} \quad (15)$$

Step 4. Improved chained DES encryption Improved chained DES encryption of hash sequence h generates encrypted hash index h_{s1} and uploads it to the cloud to establish a cloud-based confidential database.

3.2 Authentication Side

Step 1. The original speech $x(n)$ at the user side generates a hash index h_{s1} according to Steps 1-4 in Subsection 3.1.

Step 2. The hash index h_{s1} and the cloud server side generates hash index h_{s2} for hash matching authentication. In the matching authentication process, the bit error rate in this paper has been adapted to describe the hash matching, of which the normalized Hamming distance $BER(:,:)$ of the hash sequence is defined as the bit error rate, and the calculation formula is as follows:

$$BER(h_{s1}, h_{s2}) = \sum_{i=1}^{N} (|h_{s1}(i) - h_{s2}(i)|)/N \quad (16)$$

In this paper, the Hamming distance (BER) between the perceptual hash sequences of speech segments h_{s1} and h_{s2} is calculated and compared with the threshold value by setting the size of the matching threshold, i.e.

Ξ_1 : If the perceptual contents of two speech segments h_{s1} and h_{s2} are the same, then

$$BER(h_{s1}, h_{s2}) \leq \tau \quad (17)$$

Ξ_2: If the perceptual contents of the two speech segments h_{s1} and h_{s2} are not the same, then

$$BER(h_{s1}, h_{s2}) > \tau \quad (18)$$

where τ is the threshold value of the perceptual hash and $h(\cdot)$ is the hash function. If the digital distance satisfies Equation (17), the perceived contents of the two speeches h_{s1} and h_{s2} are considered the same and the authentication is passed: otherwise, the authentication is not passed.

4 Experimental Results and Analysis

The experimental hardware platform is AMD Ryzen 5 4600 H with Radeon Graphics, 16 GB, 3.00 GHz.

The original speech database of 1200 segments was created using the speech database for TIMIT (Texas Instruments and Massachusetts Institute of Technology), TTS (Text to Speech) and the self-built speech database together. The frequency of the speech segments is 16000 Hz, the sampling accuracy is 16 bits, the length is 4s, and the format is wav. According to the environment of speech transmission, the original speech database is operated for content preserving. A speech database of 21,600 segments with different content preserving operations is established, which contains eighteen content preserving operations such as echo, noise, low-pass filter, resampling, and MP3 compression.

4.1 Discrimination Test and Analysis

Discrimination is used to evaluate the reliability of the algorithm for distinguishing different speech contents from the same or different people.

The bit error rates of perceptual hash values for different speech contents basically obey normal distribution. In this paper, 1200 speech segments are made pairwise comparison for the perceptual hash value, and 719400 BER data are obtained. As shown in Figure 3, the BER of the hash sequences of different content speech basically obeys normal distribution, of which the horizontal axis is the BER and the vertical axis is the cumulative probability.

From Figure 3, the BER curve is closer to the theoretical curve as the sequence size increases, indicating that the discrimination and collision resistance are better. At the same time, it can also be concluded that the variation of the curve decreases and the enhancement of the discrimination diminishes with the increase of the size. Hash length sequences are used in this paper, and the algorithm has the best comprehensive performance when the sequence length is chosen as 768. Based on the affiliated Morpho-Laplace central limit theorem, the parameters of normal distribution for different length hash sequences are calculated. The specific parameters are shown in Table 1.

Table 1: Parameters of normal distribution

Parameter	Hash length	Theoretical	Actual
μ	532 bits	0.5000	0.4937
μ	639 bits	0.5000	0.4950
μ	798 bits	0.5000	0.4962
μ	1064 bits	0.5000	0.4974
σ	532 bits	0.0216	0.0227
σ	639 bits	0.0197	0.0210
σ	798 bits	0.0180	0.0189
σ	1064 bits	0.0153	0.0164

As shown in Table 1 and Figure 4: When the hash sequence N is 768, the theoretical value has a significant gap with the actual value. As the length increases, the theoretical values of the normal distribution parameters become closer to the actual values, indicating the feasibility of the algorithm in this paper. The hash sequence

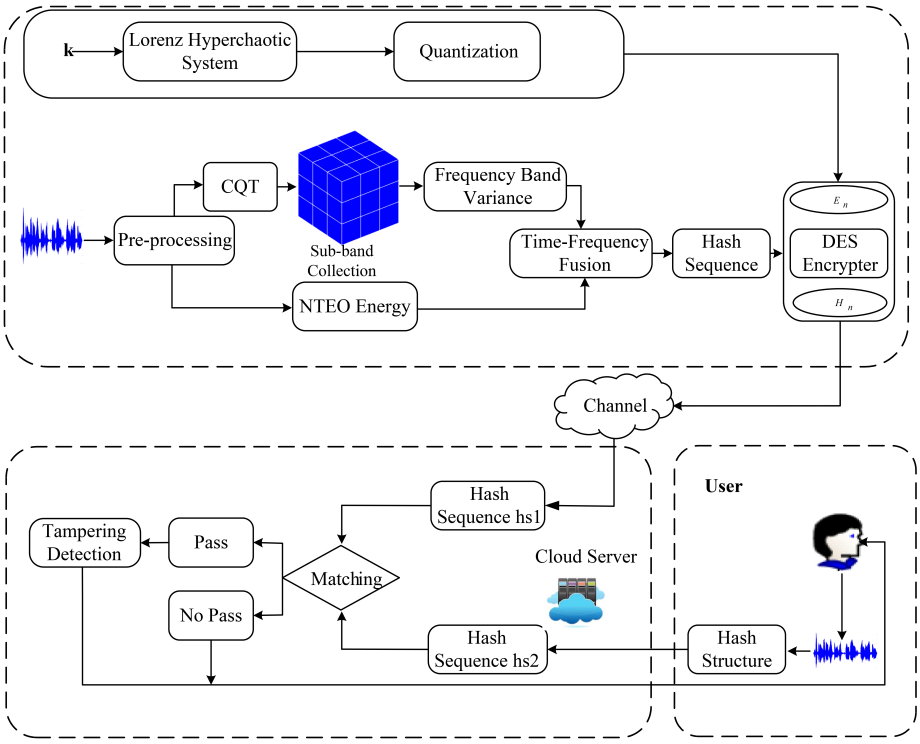
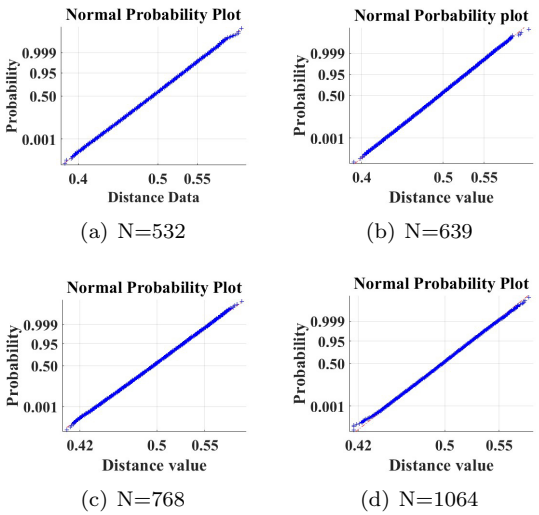


Figure 2: Speech perceptual hashing authentication flowchart



(a) N=532 (b) N=639

(c) N=768 (d) N=1064

Figure 3: BER normal distribution chart

selected by this algorithm has good randomness and collision resistance. To verify the correctness of the experiment, the Fales Accept Rate (FAR) and Fales Rejection Rate (FRR) of the algorithm in this paper can be calculated by Equations (19) & (20).

$$\begin{aligned} FAR(\tau) &= \int_{-\infty}^{\tau} f\left(x|\mu,\sigma\right)dx \\ &= \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \end{aligned} \tag{19}$$

$$\begin{aligned} FRR(\tau) &= 1 - \int_{-\infty}^{\tau} f\left(x|\mu,\sigma\right)dx \\ &= 1 - \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \end{aligned} \tag{20}$$

In Equations (19) & (20): τ is the perceptual authentication threshold; μ is the BER mean; σ is the BER standard deviation. The higher the FRR value, the weaker the perceptual robustness; the larger the FAR value, the worse the discrimination. For different algorithms, the perceptual hash robustness and discrimination can be improved simultaneously.

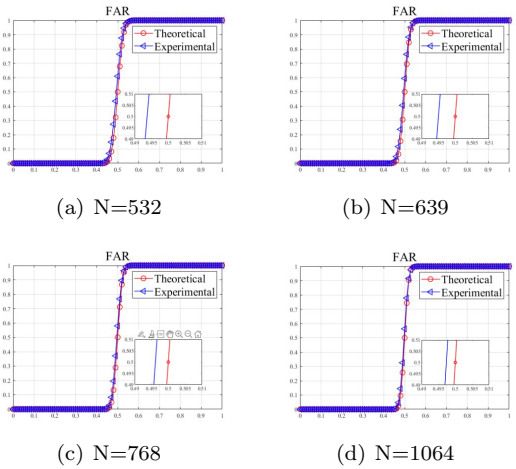
From Table 2, the smaller the setting, the smaller the FAR. When $N = 768$, set τ= 0.1, the algorithm in this paper can completely distinguish between different speech and content preserving operations, when about 4.205 of every 1×10^{98} speech segments are misidentified. As the length of the hash sequence N increases, the FAR becomes smaller, and when N is 1064, only 1.48 of every 1×10^{130} speech fragments are false recognized. As the

Table 2: FAR of different length hash sequences

τ	532 *bits*	636 *bits*	768 *bits*	1064 *bits*
0.10	2.017×10^{-67}	1.526×10^{-79}	4.205×10^{-98}	1.489×10^{-130}
0.20	1.908×10^{-38}	2.656×10^{-45}	8.693×10^{-56}	4.239×10^{-72}
0.25	4.364×10^{-27}	7.123×10^{-32}	3.494×10^{-39}	6.180×10^{-52}
0.30	8.225×10^{-18}	6.720×10^{-21}	1.311×10^{-25}	8.301×10^{-34}
0.35	1.323×10^{-10}	2.271×10^{-12}	4.759×10^{-15}	1.052×10^{-19}

Table 3: Comparison of FAR values for different algorithms

τ	[9]	[18]	[24]	[17]
0.10	3.113×10^{-35}	3.654×10^{-42}	3.191×10^{-30}	6.773×10^{-47}
0.20	1.557×10^{-20}	1.405×10^{-24}	1.113×10^{-17}	2.992×10^{-27}
0.25	9.492×10^{-15}	1.215×10^{-17}	9.588×10^{-13}	1.668×10^{-19}
0.30	5.318×10^{-10}	6.116×10^{-12}	1.075×10^{-08}	3.892×10^{-13}
0.35	2.785×10^{-06}	1.874×10^{-07}	1.601×10^{-05}	3.877×10^{-08}



(a) N=532 (b) N=639

(c) N=768 (d) N=1064

Figure 4: False recognition rate curves for different length hash sequences

length of the hash sequence increases, the more collision-resistant it is, the recognition rate increases, while the FAR decreases. However, the magnitude of the FAR reduction decreases and the robustness also decreases. This is because as the sequence length increases, features containing noise are introduced, so the robustness decreases. The algorithm in this paper fully balances the FAR and FRR, and a hash sequence length of 768 is used to achieves the optimal length for the comprehensive performance of this algorithm. By fusing the features, the FAR and FRR can be effectively reduced, and the robustness and discrimination of the algorithm can be improved.

From Table 3 compared to Table 2. When $\tau = 0.1$, although the algorithms in [9,17,18,24], can also completely distinguish between speech and content preserving operations, the FAR in this paper is much lower than the above algorithms, at which time only 4.025 out of every 1×10^{98} speech segments are false recognized by the algorithm in this paper. The best result among the four algorithms is in [17], where 6.733 are misidentified for every 1×10^{47}. The algorithm in this paper is 7.5×10^{67} times better than in [24], 7.4×10^{62} times better than in [9], 8.6×10^{55} times better than in [18], and 1.6×10^{51} times better than in [17]. It is easy to conclude that the FAR of the algorithm in this paper is much lower than that of other algorithms, which also indicates that the algorithm in this paper has very good collision-resistance and discrimination.

The Entropy Rate (ER) is a comprehensive evaluation index for the discrimination of the hash algorithm, which mainly overcomes the disadvantage that the algorithm is easily affected by the sequence size. The value of ER ranges from 0 to 1, and the higher value indicates the stronger recognition ability, which can be calculated by

Equation (21) and Equation (22).

$$ER = -[q\log_2 q + (1-q)\log_2(1-q)] \quad (21)$$

$$q = \frac{1}{2}\left(\sqrt{\frac{|\sigma^2 - \sigma_1^2|}{\sigma^2 + \sigma_1^2}} + 1\right) \quad (22)$$

In Equation (21) and Equation (22), q is the experimental mean, σ and σ_1 denote the standard deviation of the theoretical and actual BER respectively.

Table 4: Comparison of entropy rates of different length hash sequences

Hash sequence length	ER
532 bits	0.9639
639 bits	0.9535
768 bits	0.9645
1064 bits	0.9494

Table 5: ER of hashing sequences of different lengths

Algorithm	ER
[9]	0.8559
[18]	0.9432
[24]	0.9332
[17]	0.9196

As can be seen from Tables 4 and 5, when the hash sequence length is chosen as $768bits$, the entropy rate of the algorithm in this paper is the highest, and at the same time, the proposed algorithm in this paper has good discrimination compared to the entropy rates in [9,17,18, 24].

4.2 Robustness Test and Analysis

In order to evaluate the robustness of the algorithm in this paper, eighteen content preserving operations are performed for each voice in the speech data, and the average value of BER between the original voice and the hash sequence of the voice after the operations is obtained.

For the 1200 speech segments in the above speech database, the content preserving operation as shown in Table 6 is performed to obtain the various BER of the algorithm $N = 768$ in this paper as shown in Figure 5.

As can be seen from Figure 5, the mean value of bit error rate after 18 content-preserving operations of the algorithm in this paper is distributed in the interval (0.0006, 0.1565), and the maximum value is also relatively low, which indicates that this algorithm has good robustness. Among the above content preserving operation methods, the volume adjustment and resampling have less impact on the speech spectrum, and both have better robustness. The algorithm in this paper improves the anti-noise

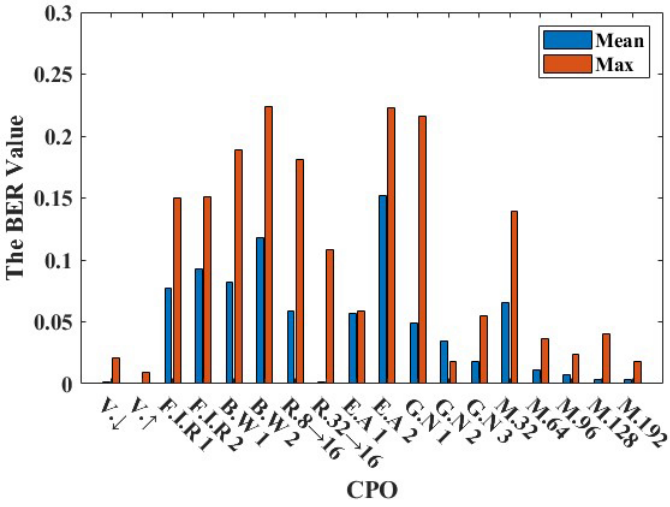


Figure 5: Mean and standard deviation of BER for the algorithm in this paper

performance of the algorithm in feature extraction, so the robustness of the algorithm in this paper is better for narrowband noise operation. To further illustrate the strong robustness of the algorithm in this paper, this experiment compares the means of the algorithm in this paper with those in [9, 17, 18, 24], as shown in Table 7.

As can be seen from Table 7, the BER mean of the algorithm in this paper is not only basically lower than the values of [9, 17, 18, 24], but also more adaptable to many different speech environments, so the robustness of the algorithm in this paper is better than the algorithm in [9,17,18,24].

In order to test the robustness of the algorithm in this paper more comprehensively, the False Rejection Rate (FRR) is introduced in this paper. FRR refers to the percentage of errors in which the same speech segment is judged as a different speech segment in this paper, as shown in Equation (20).

According to the above equation, the FRR values of BER of different speech segments in the speech database after 18 content preserving operations at different hash lengths are derived, and the FRR-FAR curves are plotted by combining the FAR values of BER of different speech segments in the original speech database.

As can be seen in Figure 6, there is no crossover between FRR and FAR curves for different hash lengths, which indicates that the algorithm in this paper has good discrimination and robustness to accurately identify content preserving operations and content malicious operations.

When the authentication pass rate judgment threshold for the above types of attacks is 0.35, the need for authentication can be satisfied. The algorithm proposed in this paper has strong robustness for content preserving operations, especially for echo and low-pass filters, and can best distinguish whether a speech can pass authentication when the judgment threshold $\tau = 0.35$. The

Table 6: Content preserving operations

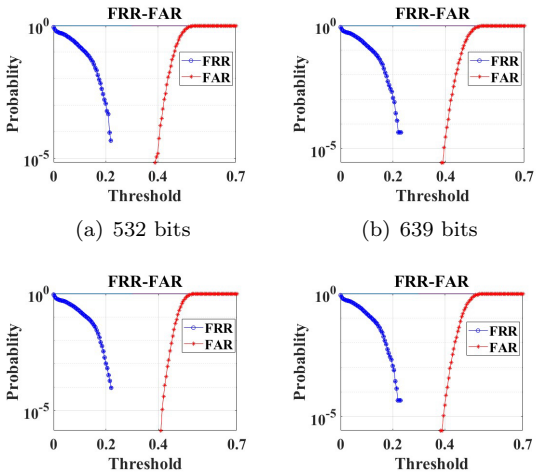
Operating means	Operation method	Abbreviation
Volume Adjustment 1	Volume down 50%	V.1
Volume Adjustment 2	Volume up 50%	V.2
Low-pass Filtering 1	6 order FIR low-pass filtering, Cutoff frequency of 3.4 kHz	F.I.R 1
Low-pass Filtering 2	12 order FIR low-pass filtering, Cutoff frequency of 3.4 kHz	F.I.R 2
Butterworth Filter 1	6 order Butterworth low-pass filtering, Cutoff frequency of 3.4 kHz	B.W 1
Butterworth Filter 2	12 order Butterworth low-pass filtering, Cutoff frequency of 3.4 kHz	B.W 2
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.N 1
Narrowband Noise 2	SNR=40 dB narrowband Gaussian noise, center frequency distribution in $0 \sim 4$ kHz	G.N 2
Narrowband Noise 3	SNR=50 dB narrowband Gaussian noise, center frequency distribution in $0 \sim 4$ kHz	G.N 3
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 64 k	M.64
MP3 Compression 3	Re-encoded as MP3, and then decoding recovery, the rate is 96 k	M.96
MP3 Compression 4	Re-encoded as MP3, and then decoding recovery, the rate is 128 k	M.128
MP3 Compression 5	Re-encoded as MP3, and then decoding recovery, the rate is 192 k	M.192

Table 7: Comparison of average BER of hashing sequences of different lengths

Algorithm	Algorithm of this paper	[9]	[18]	[24]	[17]
Operating means	Average BER				
V.1	0.0019	-	0.00004	0.0004	0.0007
V.2	0.0006	-	0.0264	0.0116	0.0231
F.I.R 1	0.0770	-	-	-	-
F.I.R 2	0.0923	-	-	0.1214	-
B.W 1	0.0822	-	-	-	-
B.W 2	0.1175	-	-	0.1057	-
R.8→16	0.0583	0.0052	-	0.0002	0.0009
R.32→16	0.0015	0.0007	-	0.0008	0.0062
E.A 1	0.0566	0.0041	-	-	-
E.A 2	0.1565	-	0.1427	-	-
G.N 1	0.0941	0.0021	0.0063	0.0581	0.0055
G.N 2	0.0349	-	-	-	-
G.N 3	0.0182	-	-	-	-
M.32	0.0655	-	-	0.0302	0.0214
M.64	0.0115	0.0037	-	-	-
M.96	0.0069	-	-	-	-
M.128	0.0031	0.0045	-	-	0.0019
M.192	0.0035	-	-	0.0011	0.0019

Table 8: Comparison of FAR values for different algorithms

Threshold	F.I.R 2	B.W 2	R.8→16	E.A 2	G.N 1	M.32
0.10	64.83%	35.83%	79.67%	0.00%	64.08%	91.67%
0.15	99.92%	82.15%	98.83%	37.17%	91.50%	100.00%
0.20	100.00%	99.08%	100.00%	99.33%	99.75%	100.00%
0.35	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
0.40	100.00%	100.00%	99.83%	100.00%	100.00%	99.92%
0.45	9.820%	9.540%	9.790%	11.23%	10.75%	10.09%



(a) 532 bits (b) 639 bits

(c) 768 bits (d) 1064 bits

Figure 6: FRR-FAR curves for different length hash sequences

attacked voice authentication pass thresholds are mainly distributed between the interval (0.33, 0.40), therefore, the optimized algorithm in this paper has better robustness to content preserving operations.

4.3 Security Analysis

In order to ensure the security of the hash sequence in transmission and storage and the randomness of the key [4, 5], the hash sequence in this paper adopts an improved chain DES hash encryption algorithm, as follows.

4.3.1 Key Space

In order to ensure the security of the hash sequence in the cloud server for users, this paper proposes an improved chained DES hash encryption algorithm to construct the hash index, which greatly enhances the key space of the hash sequence. The key of the encryption system consists of an initial key $\Phi = \{\Phi(i)|\, i = 1, 2, \cdots, N\}$ and an extended subkey $\Theta = \{\Theta(i)|\, i = 1, 2, \cdots, \varsigma\}$, the algorithm takes the four state variables of hyperchaos, unordered and quantized, and uses the initial value $Q(i) = [Q(1)Q(2)\cdots Q(N)]$ as the initial key. $Q(1)Q(2)\cdots Q(N)$ is a double precision type with a precision of 10^{-15}. The initial key space can be calculated as $2^{64\times 11} = 2^{704}$. The number of values of Θ is 2^{64}, so that the total key space size is 2^{768}, which is much larger than 2^{100}. The spatial comparison of the key length of the algorithm in this paper with other classical algorithms for symmetric encryption is shown in the table.

Table 9: Key lengths for different algorithms

Algorithms	Key length /bits
This article	768
Original DES	56
IDEA	128
AES	128/192/256

As can be seen from Table 9, the key length of this algorithm outperforms several other classical algorithms and can effectively resist exhaustive key attacks and brute force cracking.

4.3.2 Hash Data Security

In order to verify the key sensitivity and security of the encryption algorithm, this experiment analyzes the security of the hash index using the improved chained DES encryption algorithm, as shown in Figure 7.

From Figure 7(a)(b), it can be seen that there is a significant change in the normal distribution parameter A of the hash sequence after the improved chained DES encryption, which indicates that the encryption algorithm in this paper has good security. Similarly, the shift in the normal distribution parameter before and after the key change indicates that the encryption algorithm has good key sensitivity.

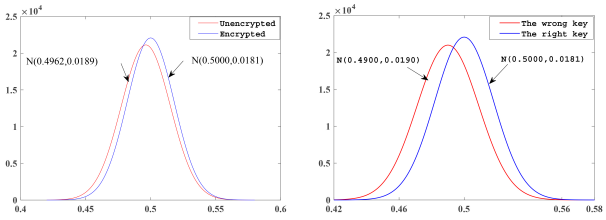
In summary, improved chained DES encryption algorithms for iterative encryption of hash sequences can better resist differential attacks, thus improving the security of the algorithm during transmission and storage.

4.4 Area Tampering Detection and Location

Threshold-controlled speech perceptual content authentication requires that identical and different voices be dis-

Table 10: Comparison of the average running times of different algorithms

—	Algorithms	Main Frequency /GHz	Average running time /s
Hash sequence length		—	
321 bits	Proposed algorithm	3.40	0.0412
532 bits	Proposed algorithm	3.40	0.0657
639 bits	Proposed algorithm	3.40	0.0770
768 bits	Proposed algorithm	3.40	0.0969
360 bits	[2]	3.40	0.0925
360 bits	[9]	3.40	0.1825
360 bits	[18]	3.40	0.0848
360 bits	[15]	3.40	0.0250



(a) Normal distribution before and after encryption (b) Key sensitivity normal distribution chart

Figure 7: Normal distribution of safety

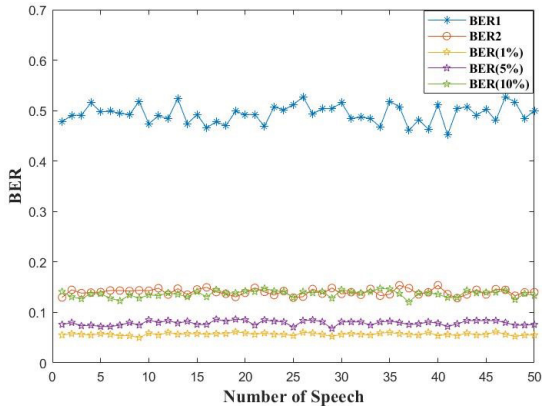


Figure 8: Different states at BER mean

tinguished and that there should be further discrimination between content preserving operational voices and small-scale tampering. Range tampering attacks as a malicious attack can be mistaken for content preserving operations, so we accomplish tampering detection by means of hierarchical detection. The process of determining the threshold is as follows: Calculate the BER of the hash sequence of different content speech, which is denoted as BER1. Calculate the average BER of the hash sequence of the original speech after 18 content preserving operations, which is denoted as BER2. Calculate the BER of the hash sequence of the original speech after different degrees of tampering operations, which is denoted as (BER1%, BER5%, BER10%), as shown in Figure 8.

From Figure 8 it can be seen that there is a clear distinction between the three BER values, so that initial authentication of speech data can be performed within the authentication threshold. When the threshold is in the range of (0.4602, 0.5536) then the voice is a speech with different content. When the threshold is in the range of (0.1235, 0.1690) then it is a speech after the content preserving operation. When the threshold is in the range of (0.0825, 0.0890) then it is a speech with 1% tampering. When the threshold is in the range of (0.0925, 0.1090) then it is a speech with 5% tampering. And when the threshold is in the range (0.1125, 0.1620) then it is a speech with 10% tampering.

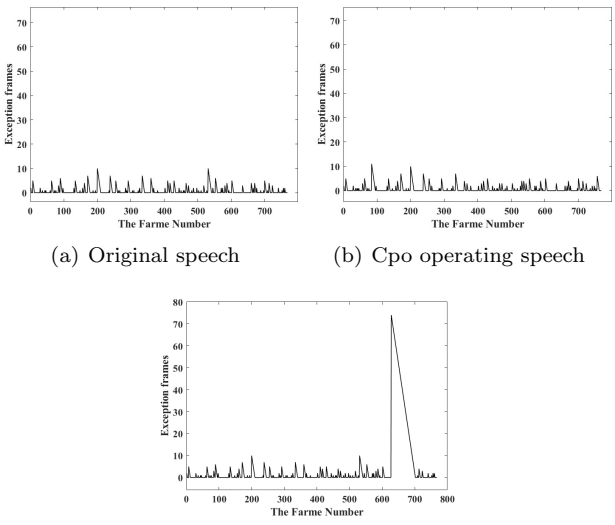
To further determine the location of a small area of tampered speech, a secondary authentication is performed, which is the tamper location process. The location of the change in successive frames is first determined by global traversal, and then the different comparison methods are used to do tampering location for different degrees of tampered speech, with the following equation.

$$\mathrm{T}(i) = \begin{cases} 1, h(i) \neq h'(i) \\ 0, h(i) = h'(i) \end{cases} \tag{23}$$

where $h(i)$ denotes the hash of frame i of the original speech segment and $h'(i)$ denotes the hash of frame i of the maliciously attacked speech segment; $T(i) = 1$ indicates that the content of frame i of the voice clip has been tampered with and $T(i) = 0$ indicates that the content of frame i of the voice clip has not been tampered with.

The original speech changes its hash value after a content preserving operation, but the changes are discontinuous, as shown in Figure 9(b). After the speech has been tampered with, as seen in Figure 9(c), the hash value changes continuously, thus allowing us to further distinguish the content preserving operation from the repeated parts of the different levels of tampering attacks by means of a continuous string detection algorithm.

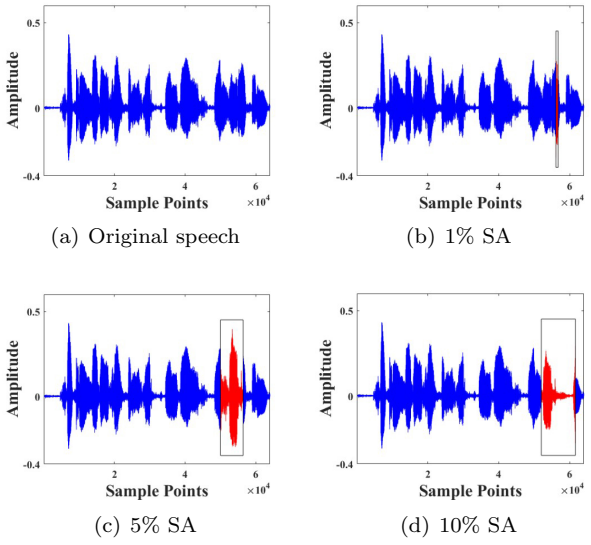
Figure 10 represents the sample points of the signal after a speech segment has been subjected to substitution attack(SA). The blue area in the figure is the speech signal and the black bar represents the area where the speech signal has been tampered with.



(a) Original speech (b) Cpo operating speech

(c) 10% Speech tampering

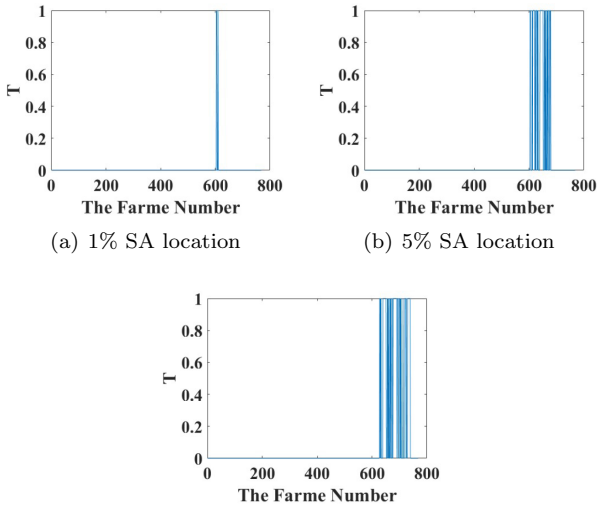
Figure 9: Continuous frame change under different operations



(a) Original speech (b) 1% SA

(c) 5% SA (d) 10% SA

Figure 10: Different types of substitution attacks



(a) 1% SA location (b) 5% SA location

(c) 10% SA location

Figure 11: Frame location for different types of substitution attacks

According to Figure 11, the hierarchical detection algorithm proposed in this paper can detect changes in the speech frame data of substitution attacks, indicating that the algorithm has a good detection and location effect on tampering attacks.

Figure 12 represents the sample points of the signal after a speech segment has been subjected to mute attack (MA). The blue area in the figure is the speech signal and the black bar represents the area where the speech signal has been tampered with.

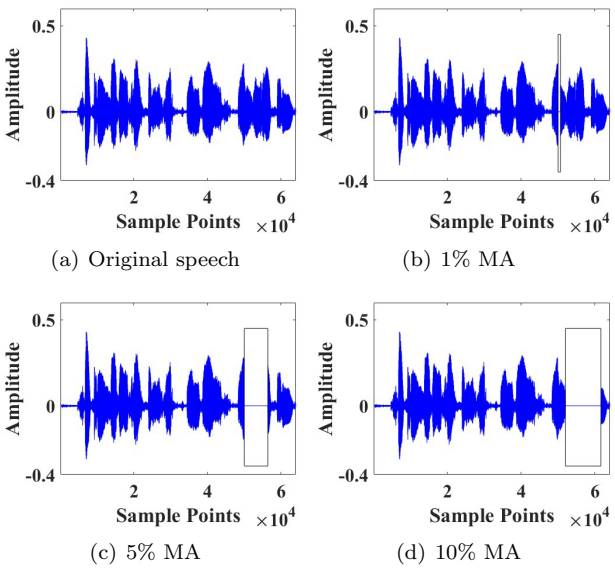
According to Figure 13, the hierarchical detection algorithm proposed in this paper can detect changes in the speech frame data of mute attacks, indicating that the algorithm has a good detection and location effect on tampering attacks.

4.5 Real-time Testing and Analysis

Real-time performance is a very important evaluation criterion in speech authentication. In order to evaluate the real-time performance of the algorithm for this paper, 200 randomly selected speech fragments from the speech database are needed to test the performance of the algorithm in the same experimental environment separately and statistically.

As can be seen from Table 10, when the length of the hash sequence increases, the running time increases, this is because the sample points and frame length are fixed and as the length of each speech segment hash sequence increases, it adds a large number of mathematical operations.

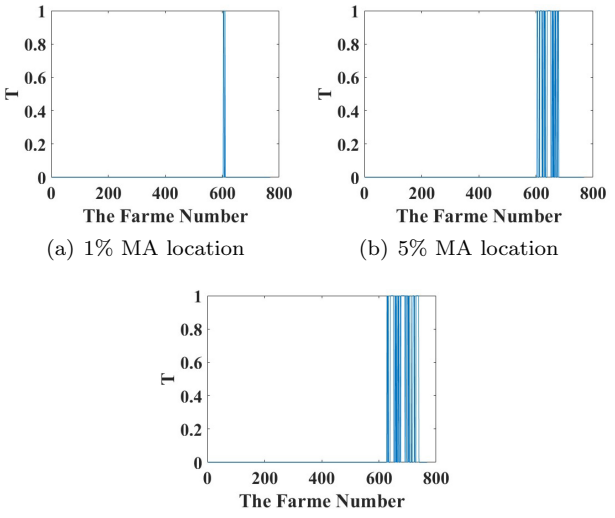
The algorithm in this paper is more efficient than in [2, 9, 18], but less efficient than in [15]. This is because the algorithm in this paper is structurally complex and hash length is long, which leads to a large amount of computation. Overall test shows that this algorithm



(a) Original speech (b) 1% MA

(c) 5% MA (d) 10% MA

Figure 12: Different types of mute attacks



(a) 1% MA location (b) 5% MA location

(c) 10% MA location

Figure 13: Frame location for different types of mute attacks

can well meet the authentication requirements of real-time communication.

5 Summary

This paper presents a perceptual hash-secure speech authentication algorithm based on improved chained DES in a cloud environment. The algorithm balances the overall performance of the algorithm while optimising robustness and real-time performance. Issues such as hash security as well as tamper detection and positioning accuracy are ensured during cloud-based authentication. The experimental results show that: (1) this paper adopts the improved chained DES algorithm to encrypt the hash sequence, introduces the random mechanism of key control, encrypts the hash sequence in the way of " One key at a time ",improves the security of the hash value in the authentication system. At the same time, the introduced tamper detection and positioning can accurately locate the attack location, and meet the authentication requirements of the authentication system. (2) Improved uniform sub-band frequency band variance and NTEO are used for fusion to enhance the robustness of the algorithm and the pass rate problem under different content preserving operations.

Acknowledgments

This work is supported by the National Natural Science Foundation of China(No. 61862041), Natural Science Foundation of Gansu Province of China(No. 21JR7RA120).

References

[1] E. Alajrami, BAM. Ashqar, B. S. Abu-Nasser, "Handwritten signature verification using deep learning," *International Journal of Academic Multidisciplinary Research (IJAMR)*, vol. 3, no. 12,2020.

[2] N. Chen, W. G. Wang, "Robust speech hash function," *ETRI Journal*, vol. 32, no. 2, pp. 345–347,2010.

[3] Y. B. Huang, H. Li, Y. Wang, Q. Y. Zhang, "High Security Speech BioHashing Authentication Algorithm Based on Multi-feature Fusion," *International Journal of Network Security*, vol. 23, no. 6, pp. 962-972,2021.

[4] M. S. Hwang, E. F. Cahyadi, Y. C. Chou, C. Y. Yang, "Cryptanalysis of Kumar's Remote User Authentication Scheme with Smart Card," in *2018 14th International Conference on Computational Intelligence and Security (CIS). IEEE*, pp. 416-420,2018.

[5] B. Irawan, M. S. Hwang, "The weakness of Moon et al.'s password authentication scheme," *Journal of Physics: Conference Series*, vol. 1069, no. 1, pp. 012070,2018.

[6] Y. H. Jiao, L. P. Ji and X. Niu, "Robust speech hashing for content authentication," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 818–821, September 2009.

[7] N. Kumar, M. Rawat, " RP-LPP: A random permutation based locality preserving projection for cancelable biometric recognition," *Multimedia Tools and Applications*, vol. 79, no. 3, pp. 2363-2381,2020.

[8] P. Kumar, S. Mukherjee, R. Saini, P. Kaushik, P. P. Roy, D. P. Dogra, "Multimodal gait recognition with inertial sensor data and video using evolutionary algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 5, pp. 956-965,2018.

[9] F. Li, T. Wu, H. X. Wang, "Perceptual hashing based on NMF and MDCT coefficients," *Chinese Journal of Electronics*,vol. 24, no. 3, pp. 579–583, July 2015.

[10] Q. H. Liao, M. Lüking, DM. Krüger, "Long timescale atomistic simulations of the structure and dynamics of transcription factor-DNA recognition," *The Journal of Physical Chemistry B*, vol. 123, no. 17, pp. 3576-3590,2019.

[11] Z. T. Liu, A. Rehman, M. Wu, W. H. Cao, M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence,"*Information Sciences*, vol. 563, pp. 309-325,2021.

[12] Q. Meng, S. C. Zhao, Z. D. Huang, F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14225-14234,2021.

[13] L. N. Peng, J. Q. Zhang, M. Ling, A. Q. Hu,"Deep learning based RF fingerprint identification using differential constellation trace figure," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 1091-1095, 2019.

[14] S. M. Siniscalchi, T. Svendsen, C. H. Lee, "An artificial neural network approach to automatic speech processing," *Neurocomputing*, vol. 140, pp. 326—338,2014.

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

[16] Q. Y. Zhang, W. J. Hu, S. B. Qiao, T. Zhang, "An efficient voice perception hash authentication algorithm based on LP-MMSE," *Journal of Huazhong University of Science and Technology (Natural Science Edition)*, vol. 44, no. 12, pp. 127–132, 2016.

[17] Q. Y. Zhang, G. L. Li, Y. B. Huang, "An efficient retrieval approach for encrypted speech based on biological hashing and spectral subtraction," *Multimedia Tools and Applications*, vol. 79, no. 39, pp. 29775-29798,2020.

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

[19] Q. Y. Zhang, P. F. Xing, Y. B. Huang, R. H. Dong and Z. P. Yang, "An Efficient Time-Frequency Domain Speech Perceptual Hashing Authentication Algorithm Based on Discrete Wavelet Transform," in *2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pp. 8–10, Nov 2014.

[20] Q. Y. Zhang, P. F. Xing, Y. B. Huang, R. H. Dong, 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, March 2015.

[21] Q. Y. Zhang, D. H. Zhang, L. Zhou, "An encrypted speech authentication method based on uniform subband spectrum variance and perceptual hashing," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 5, pp. 2467-2482,2020.

[22] Q. Y. Zhang, L. Zhou, T. Zhang, D. H. Zhang, "A retrieval algorithm of encrypted speech based on short-term cross-correlation and perceptual hashing," *Multimedia Tools and Applications*, vol. 78, no. 13, pp. 17825-17846, July 2019.

[23] S. P. Zhao, B. Zhang, C. P. Chen, "Joint deep convolutional feature representation for hyperspectral palmprint recognition," *Chinese Journal of Electronics (in Information Sciences)*, vol. 489, pp. 167-181,2019.

[24] Y. B. Zhang, B. Q. Mi, L. Zhou, "Speech Perceptual Hashing Algorithm Based on Short-term Autocorrelation for Speech Authentication," *Radio Engineering*, vol. 49, no. 10, pp. 899-904,2019.

Biography

Huang Yi-bo received ph.D candidate degree form Lanzhou university of technology in 2015, and now working as a Associate Professor in the college of physics and electronic engineering in northwest normal university. He main research interests include Multimedia in-formation processing, infromation security, speech recognition.

Pu Xiang-Rong received the BS degrees in northwest normal university, Gansu, Chain, in 2020. His research interests include speech signal processing and application, multimedia authentication techniques.