SPEECH AUTHENTICATION BASED ON PERCEPTUAL FINGERPRINT HASHING

[1]HONGXIA WANG, [2]JINFENG LI, [3]YONG QIU

[1]Prof., School of Information Science and Technology, Southwest Jiaotong University,
Chengdu, 610031, P.R.China

[2]Ph.D. candidate, School of Information Science and Technology, Southwest Jiaotong University,
Chengdu, 610031, P.R.China

[3]Master, School of Information Science and Technology, Southwest Jiaotong University,
Chengdu, 610031, P.R.China

E-mail: [1]hxwang@home.swjtu.edu.cn, [2]lijinfeng_wu@sina.cn, [3]qyong1985@vip.qq.com

ABSTRACT

A perceptual fingerprint hashing method is proposed for audio authentication in this paper. First, the biologic fingerprint image is divided into overlapping rectangle blocks by randomly using a chaotic sequence, and then the gravity center of each block is calculated. Finally, the perceptual Hashing value is generated based on the gravity center of each block. To authenticate the speech identity source, the perceptual fingerprint hashing that represents speaker's identity information and speech signal must be connected closely. So the perceptual fingerprint hashing value is embedded into the speech signal as a watermark. At the authentication stage, first, the perceptual fingerprint hashing value is extracted from the speech for authentication, and then it matches another perceptual fingerprint hashing value which is restructured by the suspicious fingerprint image. Finally, the authentication result can be decided by the matching result. The experimental results indicate that the proposed algorithm is not only robust against fingerprint rotation processing and noise attack, but also has high security and uniqueness.

Keywords: *Speech Authentication, Speaker Verification, Perceptual Hashing, Biologic Fingerprint, Digital Watermarking*

1. INTRODUCTION

Nowadays, the existing authentication protocol using PKI/CA technology, smart card, digital signature and USB key means to solve the network identity authentication, but the password is easy to be forgotten, vulnerable, leaked, while the smart card and USB certificate token are easily stolen, lost, or spurious. The basic principle of traditional speaker verification technology creates a personality description model for every speaker. The speaker authentication is implemented by comparing the suspicious speech with the saved speaker model. If the similarity is smaller than a specified threshold, the speaker's identity will be confirmed. Otherwise, the authentication doesn't pass. Therefore, the existing speaker verification algorithm mainly consists of three parts: (1) Speech signal processing and feature extraction; (2) Speaker verification model establishing and training; (3) Calculating the matching distance between the suspicious speech and the model. The traditional speaker authentication process is shown in Figure 1.

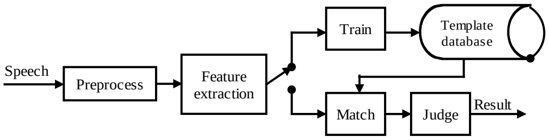


Figure 1: Principle of Traditional Speaker Verification

The main traditional speaker verification methods are presented as follows: (1) Template model method [1]; (2) Probability model method [2]; (3) Artificial neural network method and support vector machine (SVM) [3]; (4) Fusion method [4]. In general, these methods are based on the speech features.

The speech signal identification is relevant to the surrounding environment, emotion, health status and others of the speaker. So, how to find the features that do not rely on the above factors and more accurate to reflect the different speaker is an urgent problem to be solved. Meanwhile, the speech is easy to be imitated. With the development of high capability digital recorder, the attacker can achieve sound features of the speaker with a very low cost. Therefore, the identity reliability is reduced by means of using the voice feature only.

Limited to speech signal itself, simple features extracted from the voice signal to speaker recognition cannot solve the problem commendably. So other aided features for speaker verification are expected to solve these problems such as speaker lip changes [5] or other biologic features. The fingerprint features possess high recognition accuracy, low equipment cost, small data storage and other advantages, so it can be used for speech recognition technology. However, the same individual's fingerprint images data maybe have a certain difference due to two times input angle minor differences, the fingers' pressing strength, and wet / dry degrees. Considering the notable difference between the perceptual hashing and the traditional hashing algorithm that even two perceived similar fingerprint images have numerically different Hashing representations, the perceptual Hashing function enables them to produce the same or similar hashing value [6] [7]. Gravity center, geometric invariant to translation and rotation, can tolerate small geometric distortion. Because the gravity center of fingerprint image is calculated based on image content, the perceptual hashing values generated by the gravity center of different fingerprint images are different. Even if two fingerprint images' perceptual elements are the same, they will also generate different perceptual hashing values due to different keys.

Therefore, in this paper, we propose a speech authentication algorithm by the perceptual hashing model based on the gravity center of biometric fingerprint image. To connect the identity information with the speech signal for the purpose of speaker verification, the perceptual fingerprint Hashing value as a watermark is embedded into the speech signal. The remainder of this paper is organized as follows. In Section 2, the proposed algorithm is described. Experimental results and capability analysis are given in Section 3. Finally, the conclusions are given in Section 4.

2. PROPOSED ALGORITHM

2.1. Basic idea

First of all, the fingerprint image is collected to the computer through the fingerprint scanner, and then gravity center of the fingerprint image is calculated. A perceptual hashing model based on the gravity center of the fingerprint image is created under the control of the key. In order to connect the speaker identity information with the speech signal, the generated perceptual hashing sequence, as a watermark, will be embedded into the speech signal. When we identify the speech identity source, first, we extract the perceptual hashing sequence from the speech signal to be authenticated. Then, we scan the fingerprint image into the computer according to the same processing as before and the same key to generate another perceptual hashing sequence, and then calculate the normalized distance of two perceptual hashing sequences. Set a threshold, if the distance is less than the threshold, the authentication will pass. Otherwise, the candidate person is not the speaker of the speech to be authenticated.

2.2. Perceptual Hashing Generation

The perceptual hashing value is generated and stored as shown in Figure 2.

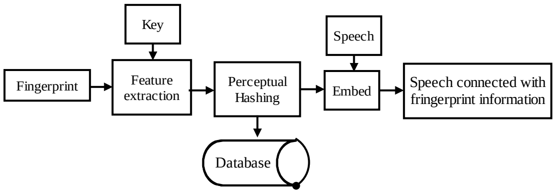


Figure 2: Generation and Storage of Perceptual Hashing Values

The main processes are as follows:

Step 1. Fingerprint image random blocking: In order to enhance the security of the perceptual hashing algorithm, we use a secrete key to generate a pseudo-random sequence Q_p. The pseudo-random sequence will randomly divided the fingerprint image into C overlapping rectangles, and the key controls the number of rectangles and the pseudo-random sequence. The fingerprint image blocking can also makes up the disadvantage that the gravity center can only describe global characteristics of an image.

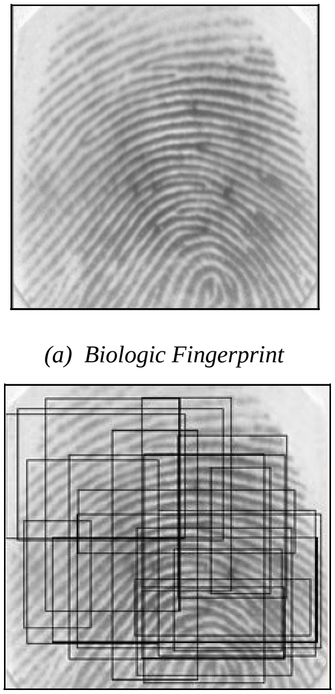
The pseudo-random sequence x_i is generated by Logistic map as follows:

$$x_i = \mu x_{i-1}(1 - x_{i-1}) \qquad 3.5699 < \mu < 4 \qquad (1)$$

Then, we obtain the binary pseudo-random sequence according to the following quantification method:

$$B_i = \begin{cases} 0 & x_i < 0 \\ 1 & x_i \geq 0 \end{cases} \qquad (2)$$

According to the image size, we adaptively select proper bits as the coordinate of x-axis and y-axis, length and width of the random region to prevent boundary, denoted by a quaternion *C* (m_x, m_y, *length*, *width*). So, the fingerprint image is divided into *C* overlapping rectangular regions, which is shown in Figure 3. Because the quaternion is randomly generated, the rectangular areas are random.



(a) Biologic Fingerprint

(b) Dividing into blocks

Figure 3: Fingerprint Divided Randomly into Overlapping Blocks

Step 2. Calculate the gravity center of the rectangular blocks: The original gray fingerprint image is denoted by $f(i,j)$, $1 \le i \le M, 1 \le j \le N$. The coordinate of the gravity center $G(m_x, m_y)$

$$\begin{cases} m_x = \sum_{i=1}^{M}\sum_{j=1}^{N} i \cdot f(i,j) / \sum_{i=1}^{M}\sum_{j=1}^{N} f(i,j) \\ m_y = \sum_{i=1}^{M}\sum_{j=1}^{N} j \cdot f(i,j) / \sum_{i=1}^{M}\sum_{j=1}^{N} f(i,j) \end{cases} \qquad (3)$$

According to step 1, the gravity center coordinates of the pseudo-random rectangular blocks are denoted by $G = \{G_1, G_2, \cdots, G_C\}$, and *C* is the number of rectangular blocks. Through the calculation of each block's gravity center, the local information of the fingerprint image can be obtained, which improves the ability to distinguish different fingerprint images.

Step 3. Gravity center quantification: Note *d* be a quantification step, and the gravity center coordinates *G* are quantified as follows:

$$S_C = \lfloor G_C / d \rfloor \times d \qquad (4)$$

Obviously, if the quantification step *d* is too large, the change of gravity center will not be sensitive, and the false acceptance rate will increase.

Step 4. Perceptual hashing value generation: Let $S = \{S_1, S_2, \cdots, S_C\}$ denote the quantized values by Step 3. Convert *S* to binary bits, denoted by $(b_{1,1}b_{1,2}\cdots b_{1,j}\cdots, b_{2,1}b_{2,2}\cdots b_{1,j}\cdots, b_{C,1}b_{C,2}\cdots b_{C,j}\cdots)_2$, $b_{C,j} \in \{0,1\}$. Finally, the generated perceptual Hashing sequence *H* is

$$H = (b_{1,1}b_{1,2}\cdots b_{1,j}\cdots)_2 \,\|\, (b_{2,1}b_{2,2}\cdots b_{1,j}\cdots)_2 \,\|\cdots\|\, (b_{C,1}b_{C,2}\cdots b_{C,j}\cdots)_2 \,.$$

2.3. Perceptual Hashing Value Storage

The perceptual hashing value is a digest, so the perceptual hashing sequence has a relatively small amount of data, which can save a lot of storage resources. The perceptual hashing sequences can be stored in the database, but it doesn't associate the speaker with the perceptual hashing sequence which can represent the speaker's identity information. Moreover, the cost will increase because of the database management. When the information is transmitted trough the network, it is easy to be attacked, which reduces the security of authentication system. Therefore, in this paper, the perceptual hashing sequences which represent the speaker's identity information will be connected with the speech signal. There are two schemes to combine the perceptual hashing sequences and the speech signal. One is that the perceptual hashing sequence is directly added to the end of the speech file. But the position of authentication data is fixed, so it is easy to be damaged. For example, the attacker cut off the end of the speech files, which contains the authentication data (the perceptual Hashing sequence), and the authentication will be unsuccessful. Another scheme is that the perceptual hashing sequence is embedded into the speech signal as a digital watermark. Because the authentication data are dispersed in the speech signal, the security is improved. In this paper, we use the robust watermarking algorithm to resist the conventional signal processing and malicious attacks. The perceptual hashing sequence is regarded as a digital watermark, and speech signal as the carrier, thus the perceptual hashing sequence will be connected with the speech signal by the watermarking technique for the authentication of speech identity source.

The watermarking algorithm using robust audio watermarking algorithm based on zero-crossing rate against time-scale modification [8], which is robust enough to many attacks, such as the conventional speech signal processing, MP3 lossy compression, low-pass filtering, and the time scale modification. It is proved that the perceptual Hashing sequence can be extracted correctly, so the accuracy of speaker identification can be guaranteed.

2.4. Speech Identity Source Authentication

When we identify the speaker, firstly, the perceptual hashing sequence is extracted from the speech signal, and is matched the perceptual hashing sequence generated by the dubitable fingerprint image. If the distance between two perceptual hashing sequences is less than the specified threshold, the speech will pass the authentication. Otherwise, the authentication is failed.

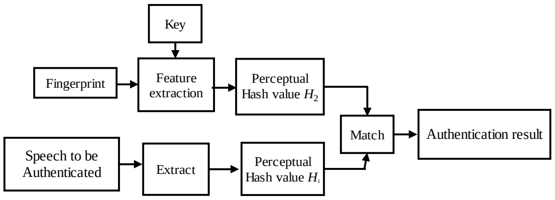

Key
Fingerprint
Feature extraction
Perceptual Hash value H_2
Match
Authentication result
Speech to be Authenticated
Extract
Perceptual Hash value H_1

Figure 4: Principle of Proposed Speaker Verification

The identification framework is as shown in Figure 4. The steps are as follows:

Step 1. The perceptual Hashing sequence H_1 is extracted from the speech signal which contains biometric feature trough digital audio watermarking extracting algorithm.

Step 2. The dubitable fingerprint image is partitioned into C random rectangular blocks by the same secret key as perceptual Hashing generation process, and the gravity center coordinates of the blocks are calculated and quantized. Finally, another fingerprint perceptual hashing sequence H_2 is generated.

Step 3. Set a threshold T, $T > 0$, the normalized Hamming distance is calculated

$$D(H_1,H_2) = \frac{1}{L}\sum_{k=1}^{L} | H_1(k) - H_2(k) | \qquad (5)$$

where L is the length of perceptual hashing sequence. If $D(H_1,H_2) \le T$, the speakers are the same person, otherwise, $D(H_1,H_2) > T$, the speakers are different. Ideally, the normalized Hamming distance of two similar fingerprints is close to 0, while the normalized Hamming distance of two different fingerprints is close to 0.5.

3. EXPERIMENTAL RESULTS AND CAPABILITY ANALYSIS

In the experiment, 100 fingerprint images of size 304×256 are randomly selected, and the format is BMP. The parameters are set as follows: the chaotic initial value is 0.23, the quantization step is 1, and the number of random rectangle blocks is 150.

3.1. Collision Test

Collision represents two different fingerprint images that generate approximate perceptual Hashing values. We calculate the 100 perceptual hashing values of 100 fingerprint images, and the distance between two perceptual hashing values D (H_1, H_2). Finally, we obtain 4950 matching values. The frequency distribution histogram is shown in Figure 5.

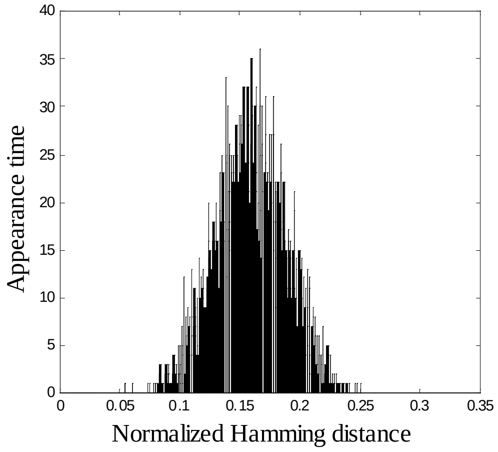


Figure 5: Collision Test

Figure 5 shows the results can be approximate to the Gaussian distribution with the expectation $\mu = 0.1591$ and standard deviation $\sigma = 0.0293$. Set the threshold $T = 0.1$, the conflict probability will be

$$p = \int_0^T \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = 7.6603\times10^{-6}$$

As a result, the conflict probability is very small. Hence, the proposed method can ensure the uniqueness of perceptual Hashing values of the fingerprint image.

3.2. Perceptual Robustness

A. Robustness to Rotation Attack

Rotate the fingerprint image with different degrees, and calculate the perceptual Hashing values H_2' . Compare H_2' with the original fingerprint perceptual Hashing values. The relationship between the rotation angle and the normalized Hamming distance is shown in Figure 6. When the rotation angle of the fingerprint image is limited to ±3°, the normalized Hamming distance $D(H_1', H_2')$ is less than the threshold $T = 0.1$, so the proposed algorithm can tolerate the rotation attack within ±3°.

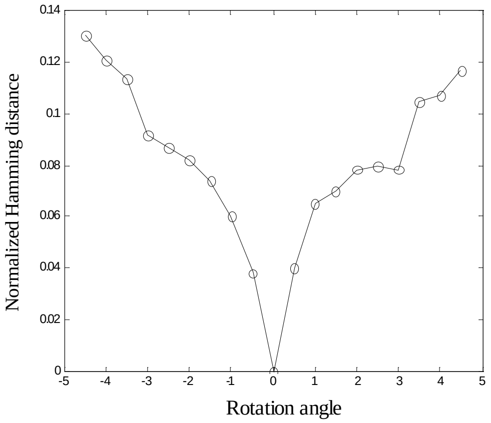


Figure 6: Robustness to Fingerprint Rotation

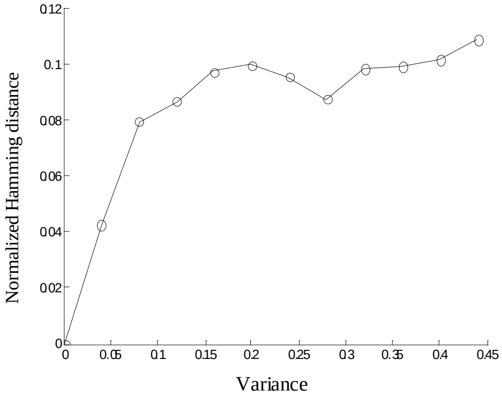


Figure 7: Robustness to Gaussian White Noise

B. Robustness to Noise Attack

The fingerprint image is added to different degrees Gaussian white noise with variance 0~0.44. We calculate the perceptual Hashing value H_2'', and extract the perceptual Hashing value H_1''. Then the normalized Hamming distance $D(H_1'', H_2'')$ is obtained. The robustness to Gaussian white noise is shown in Figure 7. From Figure 7, when the variance is less than 0.36, the normalized Hamming distance $D(H_1'', H_2'')$ is less than the threshold value $T = 0.1$ .Namely, the proposed algorithm is robust enough to resist noise attack that caused by the condition of fingerprint image acquisition, i.e. clean or not, dry or wet, which is equal to the distortion of adding noise with different variances.

3.3. Security

Because the chaotic sequence is non-periodic and sensitive to the initial value, the Logistic map is used to generate a pseudo-random sequence in this paper. The fingerprint image is partitioned into random blocks. The chaotic initial values, as the key, are set 0.23 and 0.38, respectively. The segment of perceptual Hashing value generated by the different keys is listed in Table 1.

Table 1: Segment of Perceptual Hashing Value Generated by Different Keys

…	137	103	177	113	110	206	…
…	193	183	148	92	92	125	…

By calculation, the normalized Hamming distance is 0.4511, which is far greater than the threshold $T = 0.1$, so the security meets the application requirement. Therefore, without knowing the key that the initial value of the chaotic map, even if the watermark extraction algorithm is known, the correct perceptual Hashing value generated by fingerprint image cannot be leaked.

3.4. Authentication capability discussion

The existing traditional speaker authentication method is based on large amount of speech analysis, feature extraction, training speech and modeling [8]. The amount of data to be processed is very large, and features extraction costs a lot of time. This paper introduces the gravity center of fingerprint image as well as perceptual Hashing technology, effectively reducing the consumption of time. Assume the size of fingerprint image is 304×256, and then the image has 77824 pixels with 8 bits. If the image is stored in binary, the storage space needs 622592 bits. But if we store the perceptual hashing sequence of the fingerprint image, the storage space requirement will greatly reduce. Assume the fingerprint image is randomly partitioned into 100 rectangular blocks, and then 100 gravity centers are generated and quantified. If we convert the decimal gravity centers coordinate to the binary, the storage space requires only 1600 bits. This means, the algorithm can greatly reduce the storage cost, at the same time, greatly increase the running efficiency.

The accuracy of traditional speech identification methods are highly susceptible to the environmental factors, such as the speaker's health, age, emotion, environment noise and so on. The proposed method introduces the speaker's fingerprint biological characteristic, which is embedded into the speech to be protected. In fact, our method is that a stable phase is added into the speech, so it can avoid the environmental impact. Moreover, the accuracy of speaker identification depends on the accuracy of speaker's fingerprint authentication. If the fingerprint feature can be nicely extracted, the accuracy of speech identification can reach 100% in some cases according to the performance of fingerprint perceptual Hashing technique.

4. CONCLUSIONS

In this paper, we present a speech identity source authentication scheme based on the perceptual fingerprint Hashing technique. The experimental results show that the proposed algorithm can resist rotation attack within the rotation angle ±3° , Gaussian white noise attack with variance 0.36, and has strong robustness and high security. If the key is not known, the perceptual hashing values will not be matched, even if the fingerprint image is copied. Furthermore, the experimental results prove that the fingerprint perceptual hashing sequence is unique. The proposed algorithm combines the speaker's biologic fingerprint feature with the speech signal, equivalently adding the speech with a stable phase, which is good at avoiding the environmental impact, such as age, health and other factors, which causes the probability of failure in identifying the speaker. Hence, the ability of the resistance of environmental impact is improved. Further work is needed to reduce the amount of the perceptual hashing values by coding, such as huffman coding and arithmetic compression coding.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (NSFC) under the grant No. 61170226, the Fundamental Research Funds for the Central Universities under the grant Nos.SWJTU11CX047, SWJTU12ZT02, and the Young Innovative Research Team of Sichuan Province under the grant No.2011JTD0007.

REFRENCES:

[1] A. Brew and P. Cunningham, "Vector Quantization Mappings for Speaker Verification", *20th International Conference on Pattern Recognition(ICPR)*, Istanbul, Turkey, August 23-26, 2010, pp. 560-564.

[2] M.S. Sinith, K.S. Gowri, K.V.N. Sandeep, and A. Soman, "A Novel Method for Text-Independent Speaker Identification Using MFCC and GMM", *International Conference on Audio Language and Image Processing (ICALIP)*, Shanghai, China, November 23-25, 2010, pp. 292-296.

[3] W.M. Campbell, J.P. Campbell, T.P. Gleason, D.A. Reynolds, and W. Shen, "Speaker Verification Using Support Vector Machines and High-Level Features", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 7, 2007, pp. 2085-2094.

[4] W.M. Campbell, D.E. Sturim, and D.A ReynoldS, "Support Vector Machines Using GMM Supervectors for Speaker Verification", *IEEE Signal Processing Letters*, Vol. 13, No. 5, 2006, pp. 308-311.

[5] V.R. Aravabhumi, R.R. Chenna, and K.U. Reddy, "Robust Method to Identify the Speaker Using Lip Motion Features", *2nd International Conference on Mechanical and Electrical Technology (ICMET)*, Singapore, September 10-12, 2010, pp. 125-129.

[6] X.D. Lv and Z.J. Wang, "Perceptual Image Hashing Based on Shape Contexts and Local Feature Points", *IEEE Transactions on Information Forensics and Security*, Vol. 7, No. 3, 2012, pp. 1081-1093.

[7] J.T. Zhou and Q.C. Au, "Security Evaluation of a Perceptual Image Hashing Scheme Based on Virtual Watermark Detection", *IEEE International Conference on Multimedia and Expo (ICME)*, Barcelona, Spain, July 11-15, 2011, pp. 1-6.

[8] H.X. Wang and M.Q. Fan, "Time-Scale Invariant Audio Watermarking based on Zero-Crossing Rate", *Chinese Invention Patent No. ZL 200910058797.X*, 2011, pp. 1-10.

[9] V. Varshney, D.J. Malakar, and P.K. Das, "Analysis of Patterns in Speaker Authentication Using Discrete Probability HMMs", *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, Shanghai, China, August 26-29, 2004, pp. 3706-3710.