Computer Science and Information Systems 14(3):703–718 https://doi.org/10.2298/CSIS170112024H

A Retrieval Algorithm of Encrypted Speech based on Syllable-level Perceptual Hashing

Shaofang He[1,2], Huan Zhao[1,*]

[1] School of Information Science and Engineering, Hunan University
410082 Changsha, China
wxdyzp@sina.com
[2] College of Science, Hunan Agricultural University
410128 Changsha, China
wxdyzp@sina.com
[3] *Corresponding author at: School of Information Science and Engineering, Hunan University
410082 Changsha, China
hzhao@hnu.edu.cn

Abstract. To retrieve voice information in a fast and accurate manner over encrypted speech, this study proposes a retrieval algorithm based on syllable-level perceptual hashing. It implements the function of retrieving speech segment and spoken term over encrypted speech database. Before uploading the speech to the cloud, it needs to embed the digital watermarks (perceptual hashing). In the retrieval process, it does not need search over encrypted speech data directly or decryption, but requires searching the system hash table. Experimental results show that the syllable-level perceptual hashing of the proposed scheme has good discrimination, uniqueness, and perceptual robustness to common speech. In addition, the proposed retrieval algorithm effectively improves the retrieval speed by reducing the matching number of query index. The precision ratio and recall ratio all achieve high under various signal processing.

Keywords: Speech retrieval, Posterior probability, Syllable segmentation, Perceptual hashing.

1. Introduction

Recently, with fast development of multi-media communication, audio and video have been applied more and more widely on the Internet. In particular, the digital audio has virtually become one of the most popular multi-media applications. To satisfy the requirement of large multi-media data management, cloud computing technique presents multi-media services in a new way. Cloud computing is new model of enterprise IT infrastructure which provides on demand high quality application and service from shared pool of computing resources. However, cloud storage servicer is not a trusted third party from a security standpoint. In multi-media applications, many sensitive multi-media data are related to privacy preserving, for example, in the scene of e-health, health related multimedia data is being exponentially generated from healthcare monitoring devices and sensors, coming with it are the challenges on how to efficiently acquire, index, and process such a huge amount of data for effective healthcare and related decision making, while respecting user's data privacy [1]; as well as in the scene of telecommunications, if

the sensitive speech data are stored in the cloud without protecting, it may create issues such as leakage or abuse of personal privacy speech information [2]. Therefore, security protection emerges as an important problem. One of the effective method to protect the security of outsourcing data is data encryption, but it results in the difficulty of encrypted multimedia data retrieval. As we know, encrypted data makes the traditional data utilization service based on plaintext keyword search ineffective.

2. Related works

In last few years, many works have been done for encrypted multimedia database and its retrieval. Qin Liu et al. [3] worked on Secure and privacy preserving keyword searching for cloud storage services which allows the CSP to take part in the decipherment, and returns only files in which user is interested without leaking any information about plaintext. Zhangjie Fu et al.[4] proposed Multi keyword Ranked Search Supporting Synonym Query to overcome the problems of traditional multi keyword scheme and has proposed Two secure schemes to meet up privacy requirements in two threat models as known cipher text model and known background model. The search results achieved when authorized cloud user input the synonym of the predefined keywords, not exact or fuzzy matching keywords. Baojiang Cui et al. [5] worked on Key-Aggregate Searchable Encryption (KASE), in which a data owner only needs to distribute a single key to a user for sharing a large number of documents, and the user only needs to submit a single trapdoor to the cloud for querying the shared documents. Zhangjie Fu et al. [6] proposed flexible and efficient searchable scheme which supports multi keyword and synonym based search. It proposes new text feature weighting function which adds new weighting factor to distinguish keyword on the basis of term frequency keyword and make easy retrieval. Jin Li et al. [7] worked on revocable identity based encryption which offloads all keys generation related operation during key issuing and update, leaving constant no of simple operation so that eligible users can performed locally. All the five schemes mentioned above worked well in encrypted cloud database for retrieval of data files, but as an improvement, Rupali D. Korde et al. [8] suggested new scheme where it was possible for users to upload and download multimedia data.

In multimedia data, privacy-preserving search over encrypted speech data has come into being an important and urgent research field in cloud storage. In the cloud, the rapid increase of speech data size has prompted the need to rapidly and accurately retrieve needed speech data or spoken term from protected speech databases. At present, speech information retrieval over encrypted speech data is in hotspot. Because encrypted speech lose many properties of speech signal, and such loss makes the methods used for plaintext search having highly problematic for encrypted speech retrieval. In traditional retrieval methods, keywords need to be matched exactly, however, the return results will be very less for frequent user access and large number of cloud data. In many existing retrieval methods, a keyword is encrypted as an index and matched with the encrypted data directly. After encryption, keywords lose most of speech features, and the size of encrypted speech data in cloud computing environments is massive, therefore, those algorithms implemented by matching the encrypted keyword and the encrypted data do not possess strong applicability. Ton Kalker first proposed the concept of perceptual hashing in 2001 [9]. Perceptual hashing is described as follows. (1) Bits with little data called perceptual

hash valve can represent multimedia objects with large data; (2) It meets the mapping relationship of multiple objects to one object; (3) For multimedia objects of the same or similar perceptual content, their perceptual hashing sequences are close in mathematical distance [10]. In the field of multimedia information processing and information security, the strong discrimination, uniqueness, and perceptual robustness of perceptual hashing have earned recognition since the concept was presented. The unique characteristics of speech content and mapping speech data to a brief digital digest (called perceptual hashing digest) are the basis of speech perceptual hashing technology. In this technology, a digital representation of multimedia objects is the input data, and the perceptual hashing digest is the output data. For the multimedia information with different contents, its perceptual hashing digest will be significantly differ. In other words, for the multimedia information with the same content regardless of the digital representation, its perceptual hashing digest will remain the same or similarity. The generation of speech perceptual hashing generally involves pretreatment (includes framing and window addition, time-frequency transform), feature extraction, and hash algorithm construction. The method of last two steps make speech perceptual hashing digest different from existing algorithms. Wang et al. put forward a watermark-based perceptual hashing search algorithm over encrypted speech in [10]. In the proposed scheme, the zero-crossing rate is extracted from the digital speech to generate the perceptual hashing as the search digest, which is embedded into the encrypted speech signal; without downloading and decrypting, the search results could be obtained rapidly and accurately by matching and computing the normalized Hamming distance of the perceptual hashing digests between the search target and the extracted one. Based on changes in the characteristics of the time and frequency domain, Hao et al. proposed a speech perceptual hashing algorithm [11]. The scheme also offered good discrimination and robustness and puts forward new ideas for applying perceptual hash technology in large-scale data processing. Recently, after studying existing speech retrieval technologies, Zhao et al. explored a novel perceptual hashing-based retrieval algorithm [12]. In the algorithm, multifractal characteristic of speech data and the technology of piecewise aggregate approximation (PAA) were introduced to generate perceptual hashing sequence. Compared with the methods of [10], [11], the perceptual hashing generated from multifractal characteristics showed better distinctiveness and robustness.

To sum up, the existing methods ([10], [11], [12]) only can search speech segment and need searching the system hashing table completely and matching each index, which makes these methods become inefficient in large-scale data processing. To further improve the retrieval speed, discrimination and perceptual robustness of speech perceptual hashing, the present study proposes a syllable-level perceptual hashing-based retrieval algorithm for encrypted speech. Different from the existing methods, the syllable-level perceptual hashing is introduced in this study for the first time; furthermore, the posterior probability based on acoustic segment models [13] is employed to generate the perceptual hashing digest. Additionally, in the process of retrieving speech segment and spoken term, only the perceptual hashing of equal length and header matching with the target perceptual hashing should be matched in system hash table, and it brings the greatly improvement in retrieval speed.

The remainder of this paper is organized as follows: Section 3 describes the system model of retrieval scenario and a desirable retrieval scheme. Section 4 presents exhaus-

tively the retrieval scheme, which mainly includes the generation of syllable-level perceptual hashing, the retrieval algorithm of speech segment and spoken term. Experimental results and analysis are given in Section 5. Conclusions of the study are drawn in Section 6.

3. System model

As discussed in the introduction, in order to protect speech data privacy, speech need to be encrypted before being transferred to the cloud storage servicer. Speech encryption can be done using state-of-the-art ciphers such as undetermined blind source separation-based dual key speech encryption algorithm [14]. Built upon the established cryptographic speech encryption tools, it is computationally difficult to decrypt speech data. Encryption keeps speech data safe from the server but also makes it difficult for the server to build searchable indexes. A desirable indexing scheme for encrypted speech retrieval, in addition to being efficient and scalable, should retain the similarity between speech pairs. The system model is shown by the left and the retrieval scheme is displayed by the center dash-dotted blocks in Fig.1. The model is mainly composed of generation of encrypted speech with watermarks and retrieval processes. An efficient way of representing speech and potentially enabling fast and scalable search is by the speech perceptual hashing. Before building search index (speech perceptual hashing digest), the posterior probability based on acoustic segment models of speech segment is extracted by employing the method of [13]; meanwhile, syllables are obtained by utilizing the syllable segmentation algorithm [15]. For speech segments, the perceptual hashing sequence of each syllable is generated and embedded into encrypted speech as a digital watermark. The system hash table is formed by the perceptual hashing sequences of all speech segments. In the process of speech retrieval, feature extraction and syllable segmentation of the query speech are conducted, and the perceptual hashing sequences of all the syllables are generated and built the query index (target perceptual hashing). Instead of searching over encrypted speech data directly, the target perceptual hashing digest searches in the system hash table. If the perceptual hash values match successfully, the retrieval result is obtained.

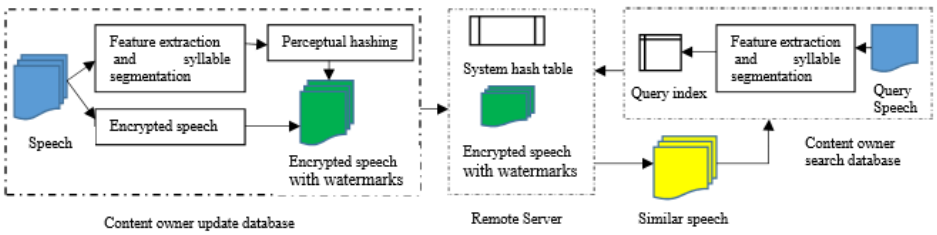

Fig. 1. System model

4. Retrieval scheme

In this section, we consider two retrieval schemes, namely, speech segments retrieval and spoken term retrieval.

4.1. Generation of speech perceptual hashing

For audio retrieval technology based on context, the building of index is one of the key links. It is critical to extract better and shorter digital digest representing audio for enhancing retrieval performance. In the retrieval algorithms of encrypted speech database, perceptual hashing sequence generated from speech features is considered as the index. Generally, the extraction of speech feature for audio signals uses short-time analysis technology. Depending on the extraction method, speech feature involves linear and nonlinear characteristics. There are advantages and disadvantages for linear and nonlinear characteristics of speech signal. Linear characteristics outperform nonlinear features in terms of meaning and computation, but nonlinear features show better robustness for general audio operations, although their extraction is relatively complex [16]. In this work, the posterior probability based on acoustic segment models of speech are chosen for generating speech perceptual hashing. The pending speech data are first divided into ordered syllables by employing a syllable segmentation algorithm. Subsequently, the perceptual hashing value of each syllable is calculated. Finally, the system hash table is constituted by the perceptual hashing sequences of all the syllables.

Supposing a total of t speech segments ($A_1, A_2, \cdots, A_t$) need to generate the encrypted speech with watermarks, and taking the speech segment A for example, the specific generation process of speech perceptual hashing is described as follows:

Step 1. Framing: Pending speech signals A are divided into speech frames with fixed frame lengths. The frame shift is half of the frame length supposing $A = \{a_q, q = 1, 2, ..., n\}$, where q is the frame pointer and n is the total number of frames to be included in A.

Step2. Feature extraction: Through the acoustic segment models, the posterior probability feature vector $P = \{p_1, p_2, ..., p_n\}$ of speech segment A is obtained, where$p_q = \{p_q^1, p_q^2, ..., p_q^D\}, q = 1, 2, ..., n$.

Step 3. Syllable segmentation: Utilizing the syllable segmentation method, the speech data A are divided into ordered syllables$S_i, i = 1, 2, ..., N$, where N is the total number of syllables, supposing each syllable contains up to M frames.

Step 4. Generation of perceptual hashing value: For syllable S_i, whose total frame number is m, $m \leq M$, the posterior probability feature vector is represented by $P_i = \{p_{i1}, p_{i2}, ..., p_{im}\}$, where $p_{iq} = \{p_{iq}^1, p_{iq}^2, ..., p_{iq}^D\}, q = 1, 2, ..., m$. The average value of the D-dimension component is selected to constitute the threshold vector $T = \{T_1, T_2, \cdots, T_m\}$ of perceptual hashing, i.e. $T_q = \frac{1}{D} \sum_{j=1}^{D} p_{iq}^j, q = 1, 2, ..., m$. Comparing the hash threshold vector with D-dimension posterior probability feature vector of each speech frame sequentially, the perceptual hashing sequence H_i of a fixed length being M bits is generated according to formula (1), where $H_i(l, q)$ is the value of the l^{th} row and q^{th} column of perceptual hashing digest. The perceptual hashing sequence of speech signals A is

represented by $H = \{H_1, H_2, ..., H_N\}$.

$$\begin{aligned} &H_i(l,q) = 0, q = 1, 2, ..., M - m \\ &H_i(l,q) = \begin{cases} 1, \ p^l_{i(q-(M-m))} \geq T_{q-(M-m)} \\ 0, \ p^l_{i(q-(M-m))} < T_{q-(M-m)} \end{cases}, \ q = M - m + 1, ..., M \\ &l = 1, 2, ..., D \end{aligned} \tag{1}$$

Step 5. Construction of the system hash table: a system hash table is constructed with the speech perceptual hashing sequences $(H^1, H^2, ..., H^t)$of all the speech segments $(A_1, A_2, , A_t)$.

As discussed in the system model, the perceptual hashing digest needs to be embedded into the encrypted speech as a digital watermark. Because of the modification of encrypted speech resulting in decryption errors (they must be reduced as much as possible), perceptual hashing digests as digital watermarks will be embedded into the least significant bit (LSB) of the encrypted speech data. For syllable S_i, which contains m frames,$m \leq M$, the embedding of perceptual the hashing sequence is described as follows: firstly, in the perceptual hashing sequence H_i with a fixed length M, there are $(M - m)$ zeros before the most significant bit, after removing these zeros, the equivalent perceptual hashing sequence with a length of m bits is obtained; secondly, the sample points of the speech frames are chosen sequentially and converted to binary forms, then, the perceptual hashing value is assigned to the LSB as a digital watermark to produce encrypted speech with watermarks.

4.2. Speech segment retrieval

After the encrypted speech data with digital watermarks is generated and the system hash table is uploaded to the cloud server, the retrieval of speech segment over encrypted speech data can be conducted without decryption as soon as a user sends a retrieval request. Supposing Q is the speech segment to be retrieved, the search process is detailed as follows.

Step 1. The D-dimension posterior probability features of Q are extracted, and syllable segmentation is performed to finally obtain N syllables.

Step 2. For each syllable of Q, a perceptual hashing sequence with a length of M is generated according to the method presented in Section 3.1. The target perceptual hashing sequence $H_Q = \{H_{Q1}, H_{Q2}, ..., H_{QN}\}$that corresponds to the query speech segment is constructed with the perceptual hashing digests of all syllables to be included in Q.

Step 3. The perceptual hashing values with a length of $M \times N$are searched out from the system hash table, supposing one of them is$H_S = \{H_{S1}, H_{S2}, ..., H_{SN}\}$. Before matching H_Q with H_S, the normalized Hamming distance (bit error rate, BER) [12] of perceptual hashing digest between two syllables (such as H_i and H_j) should be first defined, and its formula is displayed as follows:

$$D(H_i, H_j) = \frac{1}{D \times M} \sum_{l=1}^{D} \sum_{q=1}^{M} (H_i(l,q) \oplus H_j(l,q)) \tag{2}$$

Then, the normalized Hamming distance between H_Q and H_S can be calculated according to formula (3).

$$D(H_Q, H_S) = \frac{1}{N}\sum_{i=1}^{N} D(H_{Qi}, H_{Si}) \tag{3}$$

Supposing the similarity threshold is$T', 0 < T' < 0.5$, if$D(H_i, H_j) < T'$, then H_i and H_j are matched successfully; similarly, if$D(H_Q, H_S) < T'$, then H_Q and H_S are matched successfully as well. In the candidate perceptual hashing with a length of$M \times N$, their headers should be matched with H_{Q1}firstly, takeH_S for example, if$D(H_{Q1}, H_{S1}) < T'$, then H_Q and H_S should be matched successively, otherwise, the target perceptual hashing sequence $H_Q = \{H_{Q1}, H_{Q2}, ..., H_{QN}\}$need not to match H_S . It will continue to match the next perceptual hashing with a length of $M \times N$ using the same method. In general, because that the perceptual hashing sequence of each syllable has a fixed length M, the speech segment that corresponds to the perceptual hashing digest with a length of $M \times N$includes N syllables, therefore, the perceptual hashing digests of speech segments without having N syllables do not need matching in the system hash table; furthermore, owning to the speech perceptual hashing sequences matched successfully have the similar header, the candidate perceptual hashing of equal length without similar header do not need matching as well. In this way, it reduces the matching number of retrieval and improves the retrieval efficiency. The illustration is given in Fig.2.

Step 4. The detection results are obtained after the completion of retrieval in the system hash table. The digital watermarks embed in the encrypted speech can be extracted and matched with the perceptual hashing digest of the query speech to verify whether the encrypted speech is damaged or not.

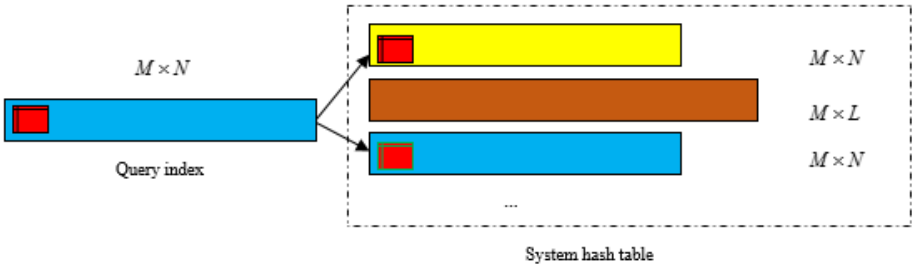

Fig. 2. Illustration of matching selection for speech segment retrieval

4.3. Spoken term retrieval

It is known that the system hash table is constructed with the speech perceptual hashing sequences $(H^1, H^2, ..., H^t)$of all the speech segments $(A_1, A_2, \cdots, A_t)$. The perceptual hashing sequence of speech sentence A_i is$H^i = (H_1^i, H_2^i, ..., H_N^i)$, which is constituted by the perceptual hashing digests of N ordered syllables to be included in A_i. The number of syllables in different speech sentence may be different, and the query spoken term

should be retrieved in the perceptual hashing sequence of each speech sentence. If a user sends a spoken term detection request, supposing K is the query term, the process that K searches in the speech segment A_i is detailed as follows.

Step 1. The D-dimension posterior probability features of K are extracted, and syllable segmentation is performed to obtain L syllables.

Step 2. For each syllable of K, a perceptual hashing sequence with a length of M is generated according to the method presented in Section 3.1, then the target perceptual hashing digest $H_K = \{H_{K1}, H_{K2}, ..., H_{KL}\}$ that corresponds to the query term is constituted by the perceptual hashing sequences of all the syllables in order.

Step 3. The perceptual hashing sequence of speech segment A_i is denoted with$H^i = (H_1^i, H_2^i, ..., H_N^i)$, which generates $N-L+1$ sets $Hv^i = (H_v^i, H_{v+1}^i, ..., H_{v+L-1}^i), v = 1, 2, ..., N-L+1$ to be matched.

Step 4. The target perceptual hashing digest $H_K = \{H_{K1}, H_{K2}, ..., H_{KL}\}$ should be matched to the $N-L+1$ sets in the appropriate order. Before matching, the normalized Hamming distance of perceptual hashing digest between two syllables headers (such as H_{K1} and H_v^i) should be first calculated. Only they are matched successfully, which means their headers are similar, should the target perceptual hashing be matched with candidate perceptual hashing set.

For example, the header of $H_K = \{H_{K1}, H_{K2}, ..., H_{KL}\}$ is first matched to the header of set $H1^i = (H_1^i, H_2^i, ..., H_L^i)$, that is, H_{K1} and H_1^i . Their normalized Hamming distance $D(H_{K1}, H_1^i)$ is calculated according to the formula (2), only it is less than the similarity threshold, should the normalized Hamming distance between the first set and the target perceptual hashing$D(H_K, H1^i)$ be calculated by the formula (3). If $D(H_K, H_1^i) < T'$, then they are matched successfully. The location along with the speech segment Ai should be labeled as one of the retrieval result. Otherwise, the target perceptual hashing H_K is continued to be matched to the next set H_2^i using the same method. The illustration is given in Fig.3.

Step 5. After the completion of retrieval in the system hash table, all the detection results are obtained. In order to verify whether the encrypted speech is damaged, it need to extract the digital watermarks embed in the encrypted speech and match with the perceptual hashing digest of the query speech.

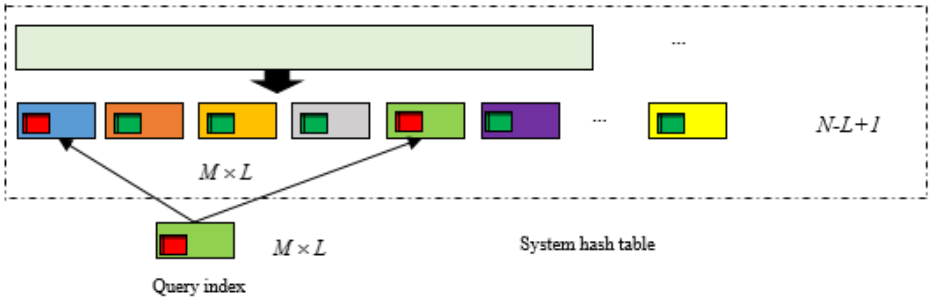

Fig. 3. Illustration of matching selection for spoken term retrieval

5. Experiments and analysis

5.1. Properties of speech perceptual hashing

In the proposed scheme, a binary sequence is used for representing perceptual hashing digest, which have a simple structure and few data. Generally, its mathematical distance is calculated by the normalized Hamming distance (also named BER). Calculating BER can determine whether two perceptual hashing digests represent the same speech. If their BER is less than the preset threshold, then they are deemed to be from the same audio contents. Otherwise, they are deemed to represent different speech contents. The most important properties of perceptual hashing are discrimination and perceptual robustness. In order to clearly describe the properties of perceptual hashing, false accept rate (FAR) is introduced, and it refers to the ratio of speech with different contents that are determined to be the same such that it is accepted by the system [17].

We performed properties of speech segments perceptual hashing experiments on a speech database containing 1,000 different speech segments from 863 Chinese continuous speech database (RASC863). These speech segments with sampling rate of 16 kHz and 16-bit quantization. The sampling signals are added Hamming windows. Each speech is divided into many frames with length of 256 sampling points, and the frame shift is half the length of the frame. Therefore, a frame equals 16 msec. The posterior probability based on acoustic segment models of speech segment is extracted by employing the method of [13]; meanwhile, syllables are obtained by utilizing the syllable segmentation algorithm [15]. The maximum length of syllables is 90 frames, that is $M = 90$, and $D = 64$ in posterior probability feature (there are 64 phonemes in Chinese). Therefore, the dimension of syllable-level perceptual hashing digest is $D \times M$. After that, the perceptual hashing digest of each speech segment is generated using the proposed method (Section 4.1). In order to obtain the statistical characteristics, we conducted a lot of matching calculation. By pairwise matching the generated perceptual hashing value (1,000 999 / 2 = 499,500 matching cases), the statistical results and its histogram of BERs is displayed in Fig.4. Obviously, it can be seen from the figure that the normalized Hamming distance distributes between 0.35 and 0.63, and the result can be approximately 03tted as the Gaussian distribution with the mathematical expectation $\mu = 0.4950$ and standard deviation $\sigma = 0.0352$. Therefore, based on such distribution parameters, the FAR under different thresholds τ (denoted as $R_{FAR}(\tau)$) can be calculated according to formula (4) for the perceptual hashing of speech segments.

$$R_{FAR}(\tau) = P(x < \tau) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{4}$$

Similarly, in the discrimination analysis of syllables perceptual hashing digests, 1,000 syllables are randomly selected from the syllable segmentation results of 1,000 speech segments. By pairwise matching the generated perceptual hashing digests (1,000 999 / 2 = 499,500 matching cases), the statistic results of BERs is obtained. It is found that the BERs distributes between 0.3128 and 0.65, and the results can be approximately fitted as Gaussian distribution with mathematical expectation $\mu = 0.4939$ and standard deviation $\sigma = 0.0463$. The FAR under different thresholds τ for the perceptual hashing of syllable also can be calculated according to formula (4) based on these distribution parameters.

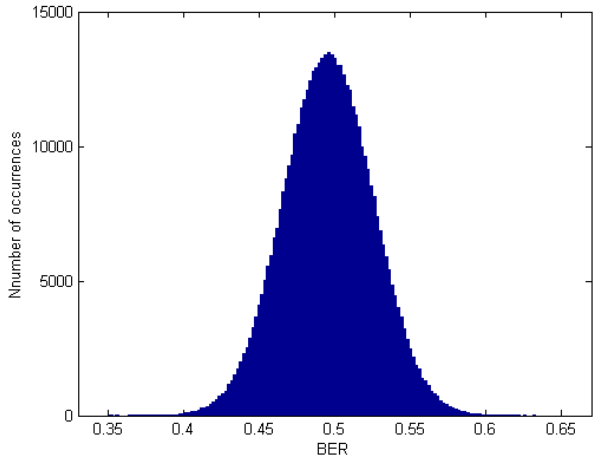

Fig. 4. Statistic histogram of 1,000 speech segments matching results

Given the threshold τ, the lower the value of $R_{FAR}(\tau)$, the better the discrimination of the perceptual hashing scheme. Employing the same speech segments data and comparing the proposed algorithm with those in references [10], [11] and [12], the FARs under different thresholds are calculated and displayed in Table 1. As seen from Table 1, $R_{FAR}(\tau)$ of the adopted scheme is lower than other three algorithms. Therefore, the perceptual hashing of the proposed method have the best properties of uniqueness and discrimination among four schemes. Similarly, according to the distribution parameters of the perceptual hashing of syllable, the FAR under different thresholds for four algorithms can be calculated (Table 2). Obviously, the adopted scheme outperforms other three algorithms in terms of the properties of uniqueness and discrimination. It is remarkable that the properties of uniqueness and discrimination for speech segments are much better than syllables, which result from the length of speech perceptual hashing.

Table 1. Comparison of $R_{FAR}(\tau)$ for the perceptual hashing of 1,000 speech segments

τ	ours	Ref [10]	Ref [11]	Ref [12]
0.02	8.44e-42	5.04e-16	4.18e-19	4.27e-26
0.04	1.60e-38	7.79e-15	1.14e-17	4.25e-24
0.06	2.20e-35	1.07e-13	2.70e-16	3.48e-22
0.08	2.20e-32	1.31e-12	5.59e-15	2.35e-20
0.10	1.59e-29	1.43e-11	1.00e-13	1.30e-18
0.12	8.40e-27	1.39e-10	1.56e-12	5.96e-17
0.14	3.21e-24	1.20e-09	2.12e-11	2.25e-15
0.16	8.90e-22	9.31e-09	2.51e-10	6.97e-14

Table 2. Comparison of $R_{FAR}(\tau)$ for the perceptual hashing of 1,000 syllables

τ	ours	Ref [10]	Ref [11]	Ref [12]
0.02	6.8798e-25	3.3761e-13	4.2121e-18	1.5238e-19
0.06	3.5758e-21	5.0217e-11	4.0171e-15	2.3963e-16
0.10	8.8816e-18	4.7821e-09	2.0809e-12	1.9666e-13
0.14	1.0559e-14	2.9156e-07	5.8549e-10	8.0422e-11

The perceptual robustness of perceptual hashing digest refers that the BER between original speech and the speech under different signal processing (such as noise reduction, compression, resampling, etc.) is less than the preset threshold. By employing four methods above and given the preset threshold 0.005, we used Cool Edit Pro v2.1, Gold Wave v5.68C, and MATLAB R2010b to process the 1,000 syllables, and the average BER between original speech and the speech under different signal processing were listed in Table 3, where the signal processing includes MP3 compression (128kbps), re-quantization (16→8→16bps), decreasing and increasing of amplitude (3dB). From the results listed in Table 3, it can be seen that the average BER of our method is less than the preset threshold under different speech signal processing, which indicates that the proposed perceptual hashing method has good perceptual robustness. Depending on the conclusion that the properties of uniqueness and discrimination for speech segments are much better than syllables, we can infer that the perceptual robustness of speech segments outperforms syllables.

5.2. Performance of speech retrieval

In this paper, we use the recall ratio R and the precision ratio P to evaluate the retrieval performance. In formulas (5) and (6), f_T denotes the number of correct search in the encrypted speech database, f_F denotes error number, and f_L denotes lost number. In the experiments, we generated encrypted speech data with watermarks by employing the undetermined blind source separation-based speech encryption algorithm and the perceptual hashing digests of 1,000 speech segments using the proposed method (they were embedded into the encrypted speech as watermarks). The system hash table was formed by the perceptual hashing digests of 1,000 speech segments. In the process of searching and matching in the system hash table, given the similarity threshold $T', 0 < T' < 0.5$, if the normalized Hamming distance $D(H_Q, H_S) < T'$, the matching succeeds. Obviously, the recall ratio and precision ratio are directly affected by the similarity threshold. In previous experimental results of 1,000 syllables discrimination test, the minimum BER is 0.3128, and in their perceptual robustness test, the maximum BER is 0.0097. Therefore, we chose the similarity threshold T' as 0.25 for avoiding missed detection and achieving a high precision ratio.

$$R = \frac{f_T}{f_T + f_L} \times 100\% \tag{5}$$

$$P = \frac{f_T}{f_T + f_F} \times 100\% \tag{6}$$

Table 3. Comparison of perceptual robustness for 1,000 syllables

Various signal processing	the average BER			
	Ours	Ref [10]	Ref [11]	Ref [12]
MP3	0.0016	0.0093	0.0067	0.0038
Re-quantization	0.0026	0.1895	0.0959	0.0693
Amplitude decrease	0.0042	0.0246	0.0157	0.0139
Amplitude increase	0.0039	0.0498	0.0557	0.0476

100 spoken terms were randomly chosen from 1,000 speech segments as the query speech. Considering that the query fed by user may be corrupted by noise, compression et.al, we processed the query with noise reduction, MP3 compression and re-quantization operation before retrieving. After retrieving in the system hashing table, the recall ratio and precision ratio under different signal processing were shown in Table 4. As can be seen from the table, the number of correct search of the proposed method is high under various signal processing operations, whereas the error number and lost number are small. Obviously, the proposed method have good retrieval performance under various signal processing.

Table 4. Recall and precision ratios under different signal processing

Operation	Noise reduction	MP3	Re-quantization
f_T	98	96	96
f_F	3	4	3
f_L	2	4	2
R	98%	96%	98%
P	97%	96%	97%

Besides, the speech segment retrieval experiments were conducted as well. All the query speech segments were processed by signal processing (shown in Table.2) before retrieving, then their perceptual hashing searched in the system hash table. Take the 600^{th} speech segment for example, it was selected as the retrieval speech and processed by re-quantization (16→8→16bps) operation. The BERs between the perceptual hashing digest of query speech and each perceptual hashing of system hash table were calculated and shown in Fig.5. As can be seen from the figure, apart from the BER between the query speech and the 600^{th} speech segment, all BERs were larger than 0.3. Given the similarity threshold 0.25, only when the BER is less than it the matching succeeds. Additionally, the experimental results of speech segments retrieval were summarized, and it is found that the proposed scheme reached 100% in terms of recall and precision ratios under various signal processing.

Different from references [10], [11] and [12], whose perceptual hashing digest of fixed length was generated for each speech segment, the perceptual hashing sequence of different length will be generated for each speech segment in the proposed algorithm. Supposing the fixed length is 500 frames in references [10], [11] and [12], the retrieval time of query speech segment by employing the adopted algorithm and methods of references

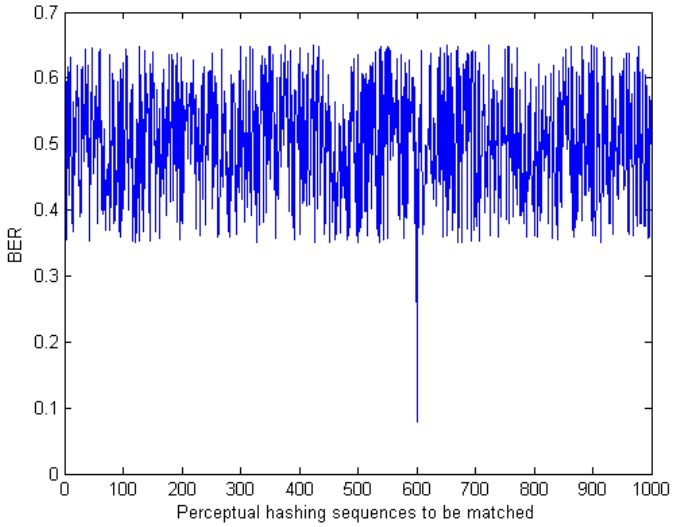

Fig. 5. Matching result of speech segment in system hash table

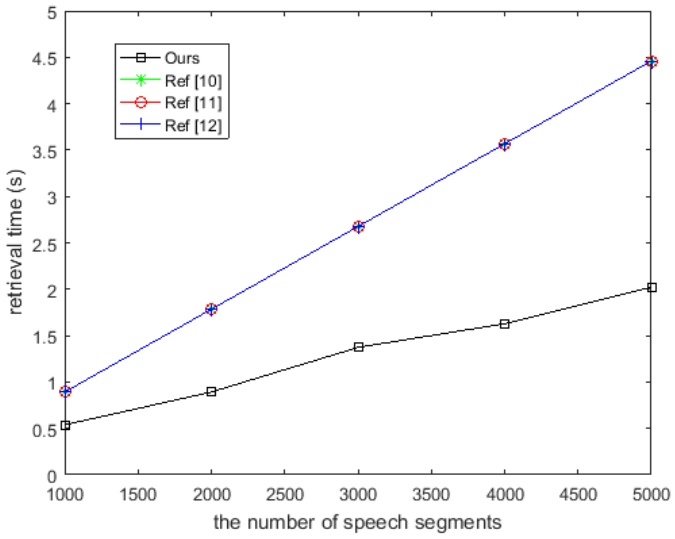

Fig. 6. Comparison of retrieval time for speech segment

[10], [11] and [12] were recorded and displayed in Fig.6. As can be seen from the figure, fixing the number of encrypted speech segments, references [10], [11] and [12] have the same detection time, which due to their same fixed length of perceptual hashing. Compared with them, the proposed method has an advantage in terms of the retrieval speed.

Obviously, as the number of encrypted speech segments increases, the detection time of the proposed algorithm is growing slowly; accordingly, the detection times of references [10], [11] and [12] are growing linearly. This is because that references [10], [11] and [12] need matching all perceptual hashing digests successively in the system hash table; by contrast, the target perceptual hashing only requires matching the perceptual hashing digests of equal length and similar header in proposed method, in other words, if the number of syllables in speech segment is different from that of the query speech, or the number of syllables is the same but without the similar header, then its perceptual hashing sequence does not need to be matched. In this way, the proposed algorithm reduces the matching number of retrieval and improves the retrieval efficiency.

6. Conclusion

Most of the existing retrieval algorithms based on perceptual hashing only can search the speech segments over encrypted speech data, and their retrieval times increase linearly along with the number of encrypted speech segments. If extended them for detecting spoken term, the properties of their perceptual hashing were bad. For the purpose of achieving spoken term retrieval in an encrypted speech database, and further improving the discrimination, uniqueness and perceptual robustness, this study proposes a syllable-level perceptual hashing-based retrieval method. Different from the existing methods, the posterior probability features based on acoustic segment models of syllable are used to generate a perceptual hashing sequence, which is then embedded into encrypted speech as a digital watermark. The perceptual hashing values of syllables obtained from the continuous speech data are constituted the perceptual hashing digest of each speech sentence, and the system hash table is composed of the perceptual hashing sequences of all the speech sentences. Without retrieving the encrypted speech directly or decryption, spoken term retrieval over encrypted speech can implement successfully. In general, the proposed method has three obvious advantages. Firstly, the syllable-level perceptual hashing derived from the posterior probability features based on acoustic segment models show better distinctiveness and robustness than them derived from the time and frequency domain features, which reduces the chance of hash collision & two segments generating the same perceptual hashing values. Moreover, it implements the function of retrieving spoken term over encrypted speech, and effectively improves the retrieval speed by reducing the matching number of query index. Finally, it achieves high recall and precision ratios under various signal processing.

Acknowledgments. This work was supported by the National Natural Science Funds of China (No. 61173106), Key Project Fund of Science and Technology Program of Changsha (No.K1403027-11).

References

1. Yuan X, Wang X, Wang C, et al. Enabling Secure and Fast Indexing for Privacy-Assured Healthcare Monitoring via Compressive Sensing. IEEE Transactions on Multimedia. Vol. 18, 2002 C 2014. (2016)

2. Karan N, Pranav P, Rajesh M. Group Delay Based Methods for Speaker Segregation and its Application in Multimedia Information Retrieval. IEEE Transactions on Multimedia. Vol. 15, 1326 C 1339. (2013)
3. Qin Liu, Guojun Wang , JieWu, Secure and privacy preserving keyword searching for cloud storage services, Journal of Network and Computer Applications, Vol. 35, 927C933. (2013)
4. Zhangjie Fu, Xingming Sun, Zhihua Xia, Lu Zhou, Jiangang Shu, Multi-keyword Ranked Search Supporting Synonym Query over Encrypted Data in Cloud Computing, Proceedings of IEEE. (2013)
5. Baojiang Cui, Zheli Liu and Lingyu Wang, Key-Aggregate Searchable Encryption (KASE) For Group Data Sharing via Cloud Storage, IEEE Transactions On Computers, Vol. 6, No. 1, 1-13. (2014)
6. Zhangjie Fu, Xingming Sun, Nigel Linge, Lu Zhou, Achieving Effective Cloud Search Services: Multikeyword Ranked Search over Encrypted Cloud Data Supporting Synonym Query, IEEE Transactions on Consumer Electronics, Vol. 60, No. 1, 164-172. (2014)
7. Jin Li, Jingwei Li, Xiaofeng Chen, Chunfu Jia, and Wenjing Lou, Identity-Based Encryption with Outsourced Revocation in Cloud Computing, IEEE Transactions On Computers, Vol. 64, No. 2, 425-4371-13. (2015)
8. Rupali D. Korde, Dr. V.M. Thakare. Secure multiple data retrieval over encrypted cloud data. International Journal of Research in Science & Engineering, 330-334. (2016)
9. Kalker T, Haitsma J, Oostveen J C, et al. Issues with digital watermarking and perceptual hashing. Proc SPIE, 189-197. (2001)
10. Wang H, Zhou L, Zhang W, Liu S. Watermarking-based Perceptual Hashing Search over Encrypted Speech. 12th International Workshop on Digital-Forensics and Watermarking (IWDW 2013), Auckland, New Zealand, 1-12. (2013)
11. Hao G Y, Wang H X. Perceptual Speech Hashing Algorithm Based on Time and Frequency Domain Change Characteristics. Symposium on Information, Electronics, and Control Technologies. (2015)
12. Zhao H, He S F. A retrieval algorithm for encrypted speech based on speech perceptual hashing. 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2016), Changsha, China, 1840-1845. (2016)
13. Chan C, Lee L. Model-Based Unsupervised Spoken Term Detection with Spoken Queries. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, No. 7, 1330-1342. (2013)
14. Zhao H, He S F, Chen Z, et al. Dual Key Speech Encryption Algorithm based on Underdetermined BSS. Scientific World Journal, Vol. 1, 57-78. (2014)
15. Andreas S, Neville R, Vikramjit M, et al. Highly accurate phonetic segmentation using boundary correction models and system fusion. IEEE International Conference on Acoustics, Speech & Signal Processing, 5552-5556. (2014)
16. Zhao H, He S F. Analysis of Speech Signals Characteristics based on MF-DFA with Moving Overlapping Windows. Physica A. Vol. 442, 343-349. (2016)
17. Whitman M, Mattord H. Principles of Information Security. Beijing: Tsinghua University Press, 252. (2006)

Shaofang He received her B.Sc. degree in Mathematics and Applied Mathematics and M.S. degree in computational mathematics at Hunan normal University in 2003 and 2006, respectively. Currently, she has received her Ph.D. in Computer Science and Technology of Hunan University. Her current research interests include speech information processing and information security.

Huan Zhao is a professor at the School of Information Science and Engineering, Hunan University. She obtained her B.Sc. degree and M.S. degree in Computer Application Technology at Hunan University in 1989 and 2004, respectively, and completed her Ph.D. in Computer Science and Technology at the same school in 2010. Her current research interests include speech information processing, embedded system design and embedded speech recognition. Prof. Zhao is a Senior Member of China Computer Federation, Governing of Hunan Computer Society, China and China Education Ministry Steering Committee Member of Computer Education on Arts. She has published more than 40 papers and 6 books.

Received: January 12, 2017; Accepted: July 1, 2017.