IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 15, 2020, accepted October 31, 2020, date of publication November 5, 2020, date of current version November 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036048

A Classification Retrieval Method for Encrypted Speech Based on Deep Neural Network and Deep Hashing

QIUYU ZHANG, (Member, IEEE), XUEJIAO ZHAO, AND YINGJIE HU
School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China
Corresponding author: Qiuyu Zhang (zhangqylz@163.com)

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61862041 and Grant 61363078.

ABSTRACT In order to improve the retrieval efficiency and accuracy of the existing encrypted speech retrieval methods, and improve the semantic representation of speech features and classification performance, a classification retrieval method for encrypted speech based on deep neural network (DNN) and deep hashing is proposed. Firstly, the speech files are classified according to the category tags, and the speech files are encrypted by Rossler chaotic map method and uploaded to the cloud encrypted speech library. Secondly, the Log-Mel spectrogram features of the original speech are extracted, and extract deep semantic features and generate classification results through the trained convolutional neural network (CNN) and convolutional recurrent neural network (CRNN). Finally, the semantic feature hash code is obtained through the constructed hash function, combined with the category hash code encoded by One Hot coding to obtain the final deep hashing binary code, and uploaded to the deep hashing index table. When retrieval, the deep hashing binary code of the query speech is obtained, and the ''two-stage'' classification retrieval strategy and the normalized Hamming distance algorithm are used to match the semantic feature hash. Experimental results show that the proposed two DNN coding models have excellent feature learning performance, and has better recall rate, precision rate and retrieval efficiency.

INDEX TERMS Encrypted speech retrieval, Log-Mel spectrogram, deep neural network, deep hashing, speech classification.

I. INTRODUCTION

With the continuous advancement of Internet and cloud computing technology, more and more companies and individuals choose to store multimedia data (text, image and speech, etc.) in the cloud. In various multimedia data, speech has the special semantic function, for example, it plays an important role in conference recording, court evidence, communication recording and other applications. Since the cloud is not a trusted third party, in order to protect the privacy of speech data, front-end encryption of speech data is one of the methods to protect the security of speech data in the cloud storage environment. But the encrypted speech data often loses most of the features, which increases the difficulty of speech data retrieval. Therefore, how to achieve efficient and secure encrypted speech retrieval is also an urgent problem to be solved [1].

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

At present, in the existing content-based encrypted speech retrieval system, traditional methods use manual perceptual features to construct perceptual hash digest [2]–[5] to achieve retrieval and matching of cloud encrypted speech data, but these manual perceptual features are largely subjective, computationally intensive and unable to reflect the semantic structure information inside the speech [6], which reduces the retrieval accuracy and precision of the retrieval system to a certain extent. In recent years, ''deep learning'' and ''deep hashing'' technologies show excellent performance in image classification / retrieval [7]–[10], speaker recognition [11], speech / language recognition [12]–[14] and audio classification / retrieval [15]–[18], various DNN models have been proposed one after another. Compared with the traditional feature extraction methods, these DNN models have powerful feature self-learning ability, they can mine the deep semantic information of data by using the internal complex topological structure, and have the advantages of high precision and fast speed. Therefore, the introduction of the deep

hashing method can well solve the defects of manual perceptual features in the existing encrypted speech retrieval system, improve the semantics of features, and further improve the retrieval efficiency and accuracy of the encrypted speech retrieval system.

In order to solve the problems of low classification accuracy and complex classification model construction of traditional classification methods, and to improve the semantics of speech features, retrieval efficiency and retrieval accuracy, etc. In this paper, two end-to-end deep hashing coding models, CNN model and CRNN model, are designed based on the advantages of deep hashing method in feature learning and retrieval. The model uses CNN structure and CRNN structure as feature extractors of input speech data, to mine the deep semantic information of speech data and generate category information. Firstly, the semantic feature hash code is obtained through the learned hash function, and convert the category information into One Hot coding as the category hash code; then the category hash code and semantic feature hash code are combined to obtain the final deep hashing binary code. In the retrieval process, the ''two-stage'' classification retrieval strategy is adopted, the category hash is first retrieved to get the candidate set of the same category, and then the semantic feature hash is retrieved from the candidate set, thereby improving the retrieval efficiency and accuracy of the encrypted speech retrieval system. The main innovations of this paper are as follows: (1) Two end-to-end deep hashing coding models, such as CNN model and CRNN model, are designed to improve the semantics of speech features and generate high-quality deep hashing binary code to improve the retrieval accuracy and retrieval efficiency of the retrieval method; (2) Integrating semantic feature learning and hash coding into an overall learning framework, it can directly generate the deep feature hash code of input speech data, and One Hot coding is used to obtain the category hash code, and add it to the construction of deep hashing binary code, which improves the efficiency of speech deep feature extraction and can generate high-quality deep hashing binary code; (3) A ''two-stage'' classification retrieval strategy is proposed, the category hash is first retrieved to find the candidate set of the same category, and then the semantic feature hash is retrieved from the candidate set, which further improves the retrieval efficiency and retrieval accuracy of the retrieval method.

The rest of this paper is organized as follows: Section II analyzes related research work. Section III describes the relevant theoretical basis in detail. Section IV presents the encrypted speech retrieval scheme and its processing process. In Section V, the encrypted speech retrieval scheme is verified by experiments, and its performance is compared with the existing methods. Finally, we conclude our paper in Section VI.

II. RELATED WORKS

At present, most of the existing encrypted speech retrieval methods use manual perceptual hashing for retrieval and matching [2]–[5], [19]. For example, He and Zhao [2] proposed a retrieval algorithm based on syllable-level perceptual hashing, which uses the posterior probability feature of speech segment model to generate perceptual hash sequence, and realizes the spoken retrieval based on encrypted speech. Zhang *et al*. [3] proposed an encrypted speech retrieval algorithm based on Chirp-Z transform and perceptual hashing second feature extraction, through Chirp-Z transform and sparse matrix to extract the perceptual hash digest, and encrypt the speech file according to the m sequence, which has good retrieval performance for noisy speech. Zhao and He [4] proposed an encrypted speech retrieval method based on perceptual hashing, which uses multifractal features and piecewise aggregation approximation to generate perceptual hash sequences, which has good discriminability and robustness. In these traditional perceptual hashing methods, the generation of hash codes is divided into two steps, firstly, the perceptual features of the speech data are manually extracted to obtain the real-valued feature vector, and then the real-valued vector is mapped to binary hash code by the constructed hash function. This hash construction method is slightly complex, and the manual perceptual feature needs a lot of prior knowledge in the construction process, and the semantic representation is poor, which unable to reflect the speech semantic information, so the retrieval accuracy and efficiency of the system are not high.

In recent years, deep learning technology has been widely used in the fields of content-based image retrieval / classification [7]–[10], [20], [21], speaker / language recognition [11]–[14], [22], [23] due to its deep feature self-learning ability. Zeng *et al*. [7] proposed a new convolutional neural network structure, which added maximum pooling and average pooling to each layer of the network, so as to preserve the effective information of the image as much as possible, and realized the rapid image retrieval. Qin *et al*. [8] proposed a novel end-to-end deep hashing model, using an optimized AlexNet network to extract discriminative image features and generate high-quality hash codes, and designed a new loss function to ensure that the similar images are ranked at the top of the search list. Fan *et al*. [11] proposed a new deep hashing method-Deep Additive Margin Hash (DAMH), which integrates the feature learning and hash function mapping into the end-to-end architecture to improve the performance of speaker recognition and retrieval tasks. Bartz *et al*. [14] proposed a language identification (LID) system, which uses the hybrid CNN and CRNN to extract features from the spectrogram of speech segments, the model has well scalability and recognition accuracy.

In addition to the excellent performance in the image field, deep learning methods also have better performance in audio (speech and music) feature extraction [15], [16], [24], classification / retrieval [15]–[18], [22], [23], [25] and other applications. Many studies have built DNN to extract audio feature vector, which has well feature representation capability, and has good classification accuracy in applications such as classification tags. Xu *et al*. [15] proposed a weakly

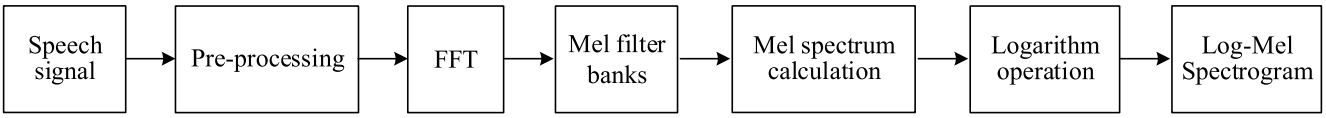


FIGURE 1. The processing flowchart of the Log-Mel spectrogram feature.

supervised audio classification method based on gated convolutional neural network, applying a CRNN with learnable gated linear units (GLUs) on the Log-Mel spectrogram, it helps to select the most relevant features corresponding to labels, so as to classify and predict audio events. Patil and Nemade [16] proposed a method based on machine learning and neural network, which combines fuzzy logic and probabilistic neural network (PNN) features to form a fuzzy probabilistic neural network (FPNN), to perform feature extraction, classification and retrieval tasks on audio. Wang *et al*. [17] proposed two novel deep CNNs: sparse coding CNN (SC-CNN) and multi-convolutional- channel CNN (MSC-CNN), which use the spectrogram as input, perform hierarchical feature learning, and have better accuracy in the recognition and retrieval of sound events. Zhang *et al*. [25] proposed an encrypted speech retrieval method based on CNN-BiLSTM and deep hashing, using the CNN-BiLSTM fusion model to generate deep perceptual hash feature of speech data, and adopting 4D hyperchaotic encryption algorithm to encrypt the original speech, the whole system has high recall rate, precision rate and security.

In summary, there are still some shortcomings in the extraction of speech features, retrieval efficiency and accuracy of existing encrypted speech retrieval methods, while deep learning and deep hashing methods show excellent performance in image / audio and other fields. Taking advantage of the powerful feature self-learning ability of the DNN model, and the advantages of high retrieval accuracy and speed of the deep hashing method, a novel encrypted speech classification retrieval method based on DNN and deep hashing is proposed in this paper, so as to improve the semantics of speech features, and further improve the retrieval efficiency and retrieval accuracy of the system, and realize the content-based encrypted speech retrieval in the cloud environment.

III. RELATED THEORIES ANALYSIS

A. LOG-MEL SPECTROGRAM FEATURE

Perception experiments show that the human ear's perception of sound signals focuses on a specific frequency region, rather than in the whole spectrum envelope. There is no linear relationship between the sound level and the actual frequency (Hz), using Mel frequency is more suitable for the hearing characteristics of human ears [18], [25], that is, it is linear distribution below 1000 Hz and logarithmic growth above 1000 Hz, the relationship between Mel frequency and actual frequency is shown in (1):

$$f_{mel} = 2595 \times log(1 + f/700) \tag{1}$$

where f_{mel} is the Mel frequency, f is the actual frequency.

The Mel frequency scale is closer to the human nonlinear auditory system, Fig. 1 shows the processing flowchart of the Log-Mel spectrogram feature. The specific processing steps are as follows:

Step 1: Speech signal preprocessing. Perform pre- emphasis, framing and adding window operations on the original speech signal $S(n)$ to obtain the processed speech signal $S'(n)$, the process of adding window as shown in (2).

$$S'(n) = S(n) \times W(n), \quad 0 < n < N - 1 \tag{2}$$

where N is the frame length, and $W(n)$ is the Hamming window function. The expression of Hamming window function as shown in (3).

$$W(n, \alpha) = (1 - \alpha) - \alpha \times \cos\left[\frac{2\pi n}{N - 1}\right], \quad 0 < n < N - 1 \tag{3}$$

where α is the Hamming window parameter, and different values of α will produce different Hamming windows, in general, α is 0.46.

Step 2: Fast Fourier Transform (FFT). Perform fast Fourier transform on the processed speech signal $S'(n)$ to get the spectrum $X_n(k)$ of each frame, and perform modular square operation on the spectrum to get the energy spectrum $|X_n(k)|^2$ of the speech signal, the spectrum calculation formula is shown in (4).

$$X_n(k) = \sum_{n=0}^{N-1} S'(n) e^{(-j\frac{2\pi k}{N})}, \quad 0 \le k \le N \tag{4}$$

where k is the point number, and N is the number of Fourier transform points.

Step 3: Mel filter banks. The energy spectrum is filtered by a set of Mel scale triangular filter banks $H_m(k)$, the calculation formula of filter output is shown in (5).

$$H_m(k) = \begin{cases} \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k \le f(m+1) \\ 0, & others \end{cases} \tag{5}$$

where $f\ (m)$ is the center frequency, and $m = 1, 2, \ldots, M$, M is the number of Mel filters, k still represents the point number.

Step 4: Mel spectrum calculation. Apply the output of the Mel filter to the energy spectrum to get the Mel spectrum $MelSpec(m)$, the calculation formula is shown in (6).

$$MelSpec(m) = \sum_{k=0}^{N-1} H_m(k) \times |X_n(k)|^2 \tag{6}$$

where $|X_n(k)|^2$ represents the energy of the k-th point in the energy spectrum.

Step 5: Logarithm operation. Perform logarithm operation on the Mel spectrum $MelSpec(m)$ obtained by (6), to obtain the Log-Mel spectrogram feature, the calculation formula is shown in (7).

$$M(n) = \log(MelSpec(m)) \tag{7}$$

B. CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN [20], [26] is a feed-forward neural network, which is composed of input layer, convolutional layer, pooling layer, fully connected layer and output layer. Among them, the convolutional layer and the pooling layer cooperate to form multiple convolutional groups, to extract features layer by layer, and finally complete classification through several fully connected layers. Because the feature detection layer of CNN learns through training data, it avoids explicit feature extraction when using CNN, and can learn implicitly from training data. In addition, because the weights of neurons on the same feature map are the same, the network can learn in parallel, which is also a major advantage of convolution network compared with the network of interconnected neurons. CNN has unique advantages in speech recognition and image processing because of its special structure of local weight sharing, weight sharing reduces the number of parameters and greatly reduces the complexity of the network. Fig. 2 shows the basic structure diagram of CNN model.

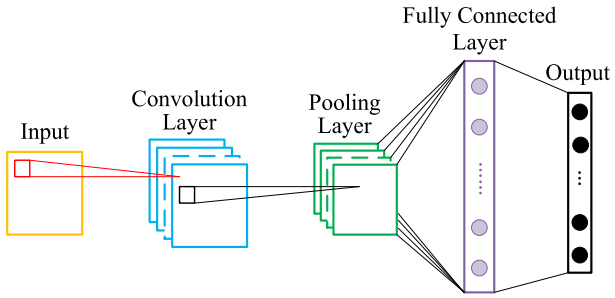


FIGURE 2. The basic structure diagram of CNN model.

C. RECURRENT NEURAL NETWORK (RNN)

RNN [14], [27] is a special neural network structure, which is mainly composed of input layer, hidden layer and output layer. It is different from other DNNs in that RNN can realize some kind of "memory function", the network state information at the previous moment will act on the network state at the next moment, the specific manifestation is that the network will memorize the previous information and apply it to the calculation of the current output. Because of its "memory function", RNN is the best choice for time series analysis, so it is widely used in natural language processing, speech recognition and other fields.

Fig. 3 shows the basic structure diagram of RNN model, where A represents a neural network module, X_t represents the input at time t and h_t represents the output at time t, and tanh represents the activation function.

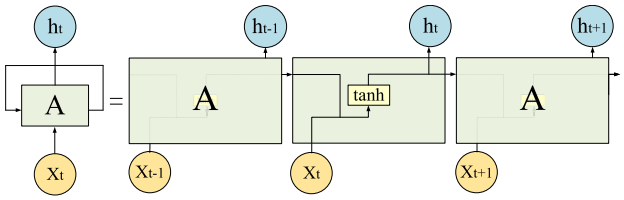


FIGURE 3. The basic structure diagram of RNN model.

Although RNN is effective in dealing with time series problems, there are still some problems, among which the more serious problems are gradient disappearance and long-term dependence, in order to solve these problems, long short-term memory network (LSTM) was designed and proposed. LSTM is a special RNN structure, which can learn long-term dependence, and it is mainly composed of input gate, forget gate, output gate and memory cell [28], [29]. In traditional RNN, the hidden layer recurrent module has a simple structure, only a simple tanh operation, while the internal structure of LSTM is more complex, there are four layers in each recurrent module: three sigmoid layers and one tanh layer, which can select and adjust the transmitted information through the gated state, remember the information that needs long-term memory, and forget the unimportant information. Fig. 4 shows the basic structure diagram of LSTM model, where A still represents a neural network module, X_t represents the input at time t and h_t represents the output at time t, the pink circle represents the point-by-point operation, and the yellow line frame represents the learned neural network layer, which cooperate with each other to control the output of information.

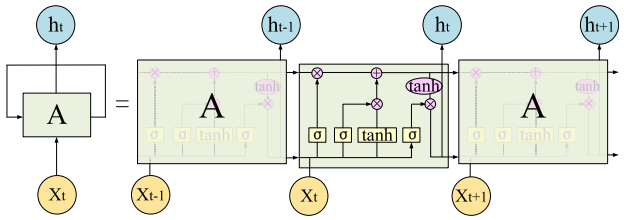


FIGURE 4. The basic structure diagram of LSTM model.

D. CHAOTIC ENCRYPTION BASED ON ROSSLER MAP

With the continuous development of encryption and decryption technology research, some low-dimensional chaotic encryption methods have appeared some cracking methods. High-dimensional chaotic systems have higher complexity, randomness and periodicity, so they can provide better encryption effects to ensure the privacy and security of data.

Rossler mapping [30] is one of the famous nonlinear dynamic systems, which has good randomness and encryption effect, and is often used for multimedia data encryption. The dynamic system model of three-dimensional Rossler

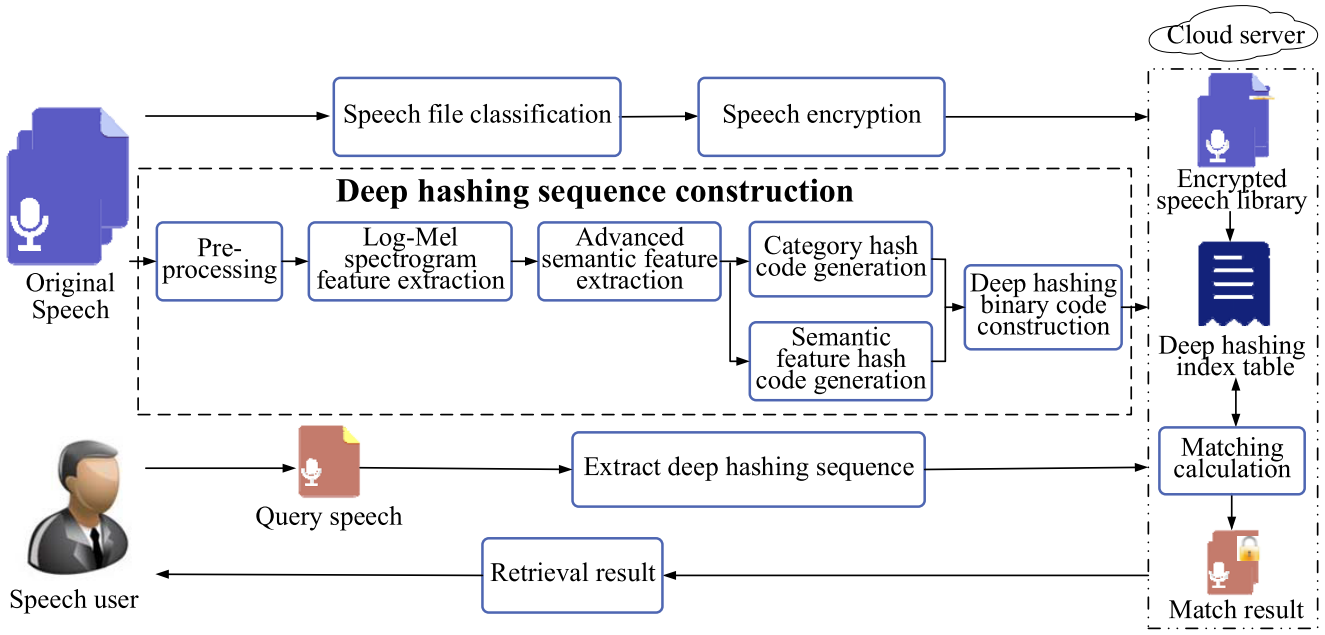


FIGURE 5. The flowchart of the system model.

chaotic map is shown in (8):

$$\begin{cases} \dfrac{dx}{dt} = -\text{y-z} \\ \dfrac{dy}{dt} = x + ay \\ \dfrac{dz}{dt} = z(x - c) + b \end{cases} \tag{8}$$

where t represents the independent variable time, $\frac{dx}{dt}$, $\frac{dy}{dt}$, and $\frac{dz}{dt}$ represent the derivative of the independent variable time t, x, y, and z represent the state variables of the system. a, b, and c are system parameters, when $a = 0.2$, $b = 0.2$, $c = 5.7$, the system is in a chaotic state.

Because the chaotic system has good sensitivity to the initial value, and the three-dimensional chaotic system has more complex system structure, the key space of the three-dimensional Rossler chaotic map is much larger than that of the low-dimensional system, which can better ensure the security encryption of multimedia data.

IV. THE PROPOSED METHOD

A. SYSTEM MODEL

Fig. 5 shows the system model of the encrypted speech retrieval system. The system model is mainly divided into three processes: the establishment of encrypted speech library, the construction of deep hashing and deep hashing index table, and speech retrieval and decryption. In the process of constructing the deep hashing scheme, the Log-Mel spectrogram features of the original speech data are used as the input of the designed CNN coding model and CRNN coding model, to learn deep semantic features and hash function at the same time, and generate classification results; In the process of speech retrieval, the "two-stage" classification retrieval strategy is adopted in this paper, the category hash is first retrieved to find the candidate set of the same category as the query speech, and then the normalized Hamming distance is used to match the semantic feature hash in the candidate set. The introduction of "deep hashing" method improves the semantic feature representation ability of speech data, it is helpful to produce high-quality and strong semantic deep hashing binary codes. By introducing classification retrieval strategy, the retrieval efficiency and retrieval accuracy of the encrypted speech retrieval system are further improved.

B. ESTABLISHMENT OF ENCRYPTED SPEECH LIBRARY

Since the cloud is not a trusted third party, it can not guarantee the data privacy and security in the cloud storage environment, so the front-end encryption operation of speech data is essential. In this paper, the chaotic encryption method based on Rossler map is used to encrypt the original speech data, after encryption, the encrypted speech file is uploaded to the cloud to form the encrypted speech library.

The specific encryption steps are as follows:

Step 1: Classify the speech files according to the category label information, and obtain the classified original speech files $S(n)$.

Step 2: Convert the original speech files $S(n)$ into matrix form $T(n)$.

Step 3: Select the initial key $[x_0, y_0, z_0]$, and obtain the encrypted real number sequence $\{K_x\}$, $\{K_y\}$, $\{K_z\}$ according to the Rossler chaotic mapping equation of (8).

Step 4: Multiply the fractional part of the sequence $\{K_x\}$, $\{K_y\}$, $\{K_z\}$ sequence by 256, take the integral part of the

product result, and convert each element into an 8-bit binary number, and then it is composed of the matrix C_x, C_y, C_z of the same size as the speech files matrix $T(n)$.

Step 5: Perform bitwise XOR processing on the elements in the matrix C_x, C_y, C_z and the corresponding elements in the original speech files matrix $T(n)$, to obtain the encrypted speech files $E(n)$ by Rossler chaotic map.

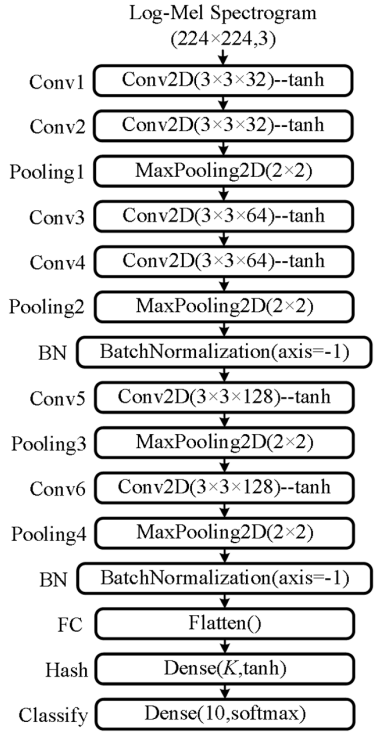
Step 6: Upload the encrypted speech files $E(n)$ to the cloud to complete the establishment of the encrypted speech library.

C. DEEP HASHING CODING MODEL CONSTRUCTION

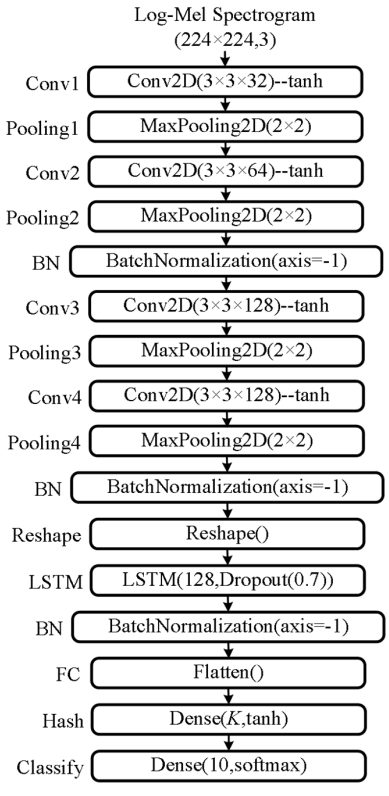
Due to the powerful feature self-learning ability of the CNN model and the excellent performance of the RNN network in processing time series data, in this paper, based on the basic structure of CNN/RNN, two end-to-end deep hashing coding models, CNN coding model and CRNN coding model are designed to learn the deep semantic features and hash functions, and construct high-quality deep hashing binary codes with strong semantic features. Fig. 6 shows the detailed parameter settings of the CNN/CRNN deep hashing coding model designed.

As shown in Fig. 6(a), the CNN coding model consists of 6 convolutional layers, 4 max pooling layers, 2 batch normalization layers, 1 flatten layer and 2 fully connected layers. The input of the model is 224 × 224 Log-Mel spectrogram of 3 channels, the filter size of each convolutional layer is 3 × 3, and the number of filters is (32, 32, 64, 64, 128, 128). The filter size used by each max pooling layer is 2 × 2, and the default step size is 1. The first fully connected layer is the hash layer, and the number of nodes K represents the length of the semantic feature hash. The final fully connected layer is the classification layer, the activation function is softmax, and the number of nodes is set to 10, which means that there are 10 types of speech in the speech library. The activation functions of convolutional layer and hash layer are tanh. In addition, in order to improve the fitting speed of the model, the batch normalization layer is added. At the same time, in order to flatten the extracted features, a flatten layer is added.

As shown in Fig. 6(b), the CRNN coding model consists of 4 convolutional layers, 4 max pooling layers, 3 batch normalization layers, 1 LSTM layer, 1 flatten layer and 2 fully connected layers. The input of the model is still a 224×224 Log-Mel spectrogram of 3 channels, and the parameter settings of convolution layer, max pooling layer, hash layer and classification layer are also the same as the CNN model. Different from the CNN model in Fig. 6(a), the CRNN model adds an LSTM layer after convolution and pooling operations, to capture the temporal features of speech data, and sets Dropout to 0.7 to accelerate the model fitting speed and reduce the risk of over fitting. In addition, the Reshape layer is added to reshape the feature dimension, so as to adapt to the input of the LSTM layer.



(a) CNN coding model



(b) CRNN coding model

FIGURE 6. Parameter settings of the CNN/CRNN coding model.

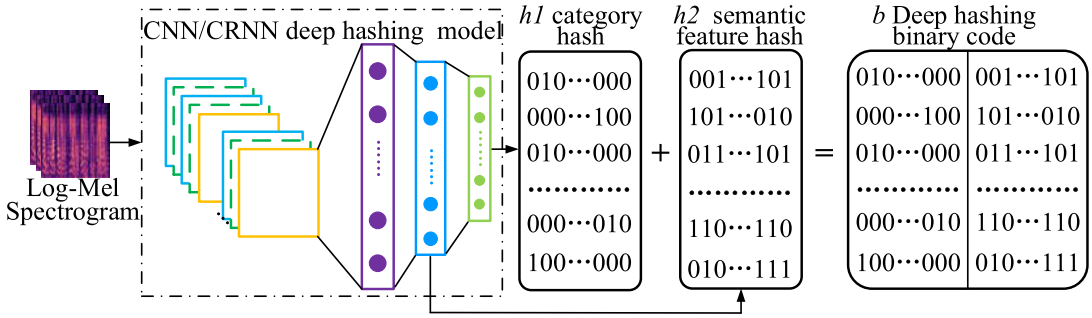


FIGURE 7. The generation process of the deep hashing binary code.

D. HASH FUNCTION LEARNING

Given a training set $X = \{x_1, x_2, \ldots, x_n\}$ with the total number of n and the category number of C, the corresponding label information is $L = \{l_1, l_2, \ldots, l_n\}$, $l_i \in \{0, 1, 2, \ldots, C-1\}$, the purpose is to extract high-level semantic features of the input speech data through two deep hashing coding models (CNN/CRNN) designed in this paper, and generate classification information based on the extracted high-level semantic features. At the same time, through the learned hash function $H(\cdot)$, the abstract semantic features are mapped into the binary feature hash code, $h1_i = H(x_i)$, $h1_i \in \{0, 1\}^K$, where K represents the length of the semantic feature hash code. The learned hash function $H(\cdot)$ must satisfy: when the two speeches x_i and x_j have the same perceptual content, the distance between the mapped semantic feature hash codes $h1_i$ and $h1_j$ is small, which means that the two speeches are the same or similar; when the two speeches x_i and x_j have different perceptual content, the distance between the hash codes $h1_i$ and $h1_j$ is larger.

Fig. 7 shows the generation process of the deep hashing binary code in the end-to-end deep hashing coding model designed in this paper. The generation process of deep hashing binary code mainly includes two parts: category hash code $h1$ and semantic feature hash code $h2$, where $h1 \in \{0, 1\}^C$, C is the number of categories, which is also the length of the category hash code, $h2 \in \{0, 1\}^K$, K is the length of the semantic feature hash code, $C + K$ is the total deep hashing binary code length.

In the learning process of hash coding, the semantic features of the input speech data need to be extracted layer by layer through the network model, the extraction process of the semantic features can be expressed as the definition of (9):

$$\omega(x_i) = W^T f(x_i, \theta_{DNN}) + b \tag{9}$$

where W is the weight matrix of the hash layer, b is the offset vector of the hash layer, θ_{DNN} is the parameter vector of convolutional layer, pooling layer and LSTM layer in the neural network model, $f(x_i, \theta_{DNN})$ is the feature vector obtained by the input data x_i after convolutional layer, pooling layer and LSTM layer calculation, $\omega(x_i)$ is the obtained deep semantic feature vector.

In order to construct the binary hash code, the extra relaxation needs to be added to the hash layer, this paper uses the tanh function to activate the output of the hash layer, and convert the real-valued semantic feature vector to $[-1, 1]^K$, the definition of tanh function is as follows:

$$tanh(\omega(x_i)) = \frac{e^{\omega(x_i)} - e^{-\omega(x_i)}}{e^{\omega(x_i)} + e^{-\omega(x_i)}} \tag{10}$$

Then, the model will map the semantic feature vector into the binary hash code with length K through the designed hash function $H(\cdot)$, in this paper, the $sign(\cdot)$ function is used as the hash mapping function to obtain the binary representation of the semantic feature, its calculation process is shown in (11).

$$h2_i = sign(tanh(\omega(x_i)) - I_{mean}) \tag{11}$$

where I_{mean} represents the mean value of the semantic feature vector.

According to the definition of (10) and (11), the generation process of the semantic feature hash code can be expressed as follows:

$$\begin{aligned} h2_i &= sign(tanh(\omega(x_i)) - I_{mean}) \\ &= \begin{cases} 1, & tanh(\omega(x_i)) \geq I_{mean} \\ 0, & otherwise \end{cases} \end{aligned} \tag{12}$$

The semantic feature hash code representation of the input speech data can be obtained from the semantic feature hash code generation process of (12).

In the process of generating category hash codes, after obtaining the semantic features through (9), the semantic features are input to the later classification layer for learning, and the softmax function is used to classify the input data to obtain the classification results, the softmax classification function is shown in (13):

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{c=1}^{C} e^{x_C}}, \quad i = 1, 2, \ldots C \tag{13}$$

where x_i represents the input data, and C represents the number of categories.

The output result of softmax classification is generally category label information, in order to combine with the semantic feature hash code, in this paper, the $argmax(\cdot)$ function is used to extract the maximum probability of the softmax

classification results, that is, the category of input data, its definition is shown in (14):

$$argmax(softmax(x_i)) = \{z|\forall y : y < z\} \tag{14}$$

where y and z represent the probability values in softmax classification results.

According to the definition of (13) and (14), the category label information can be converted into One Hot coding, and the processing process is as follows:

$$h1_i = argmax(softmax(x_i)) = \begin{cases} 1, & z \\ 0, & otherwise \end{cases} \tag{15}$$

Finally, the category hash code and the semantic feature hash code are combined to generate the complete deep hashing binary code of the input speech data.

E. CONSTRUCTION OF DEEP HASHING BINARY CODES

The deep hashing binary codes of all original speeches obtained by hash function learning are uploaded to the cloud, so as to complete the construction of the cloud deep hashing index table. The detailed construction steps of the deep hashing index table are as follows:

Step 1: Log-Mel spectrogram feature extraction. Perform pre-emphasis, framing, adding window pre-processing operations on the original speech files $S(n) = \{S_1, S_2, \ldots, S_n\}$ to obtain the processed speech files $S'(n)$, then extract the Log-Mel spectrogram features of each speech file according to the method in Section III-A, where set the number of FFT points (*n_fft*) to 1024, frame shift length (*hop_length*) to 512 and Mel frequency band number (*n_mels*) to 128.

Step 2: Network model training. Input the extracted Log-Mel spectrogram feature $M(n)$ in batches into the CNN/CRNN coding model designed in Section IV-C, after parameter adjustment and optimization, a network model with well coding performance is obtained.

Step 3: Advanced semantic feature extraction and semantic feature hash construction. Input the extracted Log-Mel spectrogram feature $M(n)$ into the trained network model, after the semantic feature extraction and hash function learning steps described in Section IV-D, the semantic feature hash code *h2* of the input speech is finally obtained through (12).

Step 4: Category hash construction. The extracted semantic features are sent to the final softmax classification layer to obtain the category information, and the category information is converted into One Hot coding by (15), which is used as the category hash code *h1*.

Step 5: Deep hashing index table construction. Splice the corresponding category hash code *h1* with the semantic feature hash code *h2*, to obtain the complete deep hashing binary code $b(n) = \{b_1, b_2, \ldots, b_n\}$ of each speech file, after establishing a one-to-one mapping relationship between the deep hashing binary codes $b(n) = \{b_1, b_2, \ldots, b_n\}$ and the corresponding encrypted speech files $E(n) = \{E_1, E_2, \ldots, E_n\}$, it is uploaded to the cloud to form the deep hashing index table.

F. SPEECH RETRIEVAL AND DECRYPTION

When the user retrieves the query speech x_q, after the same deep hashing binary code construction process in Section IV-E, the complete deep hashing binary code b_q of the query speech x_q is obtained, then adopt the proposed "two-stage" classification strategy, first, the category hash $h1_q$ is retrieved to find the candidate set of the same category of the query speech, and then the normalized Hamming distance $D(h2_q, h_x)$ is used to match the semantic feature hash $h2_q$ from the candidate set, the encrypted speech file E_q corresponding to the successfully matched hash sequence is decrypted and returned to the query user, and the retrieval successful.

When calculating the normalized Hamming distance (also called bit error rate BER) between the semantic feature hash $h2_q$ of the query speech and the hash sequence h_x in the deep hashing index table, a matching threshold τ ($0 < \tau < 0.5$) needs to be set in advance, if the distance $D(h2_q, h_x)$ is less than the set threshold τ, it means that the speech is the same as the query speech, and the match is successful. The normalized Hamming distance calculation formula is shown in (16):

$$\begin{aligned} D(h2_q, h_x) &= \frac{1}{K}\sum_{i=1}^{K}\left(\left|h_x(i) - h2_q(i)\right|\right) \\ &= \frac{1}{K}\sum_{i=1}^{K} h_x(i) \oplus h2_q(i), \quad i = 1, 2, \ldots, K \end{aligned} \tag{16}$$

where K is the length of the semantic feature hash code, $\oplus$ is the XOR operation.

After the encrypted speech file E_q corresponding to the query speech x_q is retrieved, the encrypted speech file E_q needs to be decrypted.

The specific decryption steps are as follows:

Step 1: Convert the obtained encrypted speech file E_q into matrix form E_q'.

Step 2: Use the initial key $[x_0, y_0, z_0]$ during the encryption operation, to obtain the decrypted real number sequence $\{K_x\}$, $\{K_y\}$, $\{K_z\}$ according to Rossler mapping equation of (8).

Step 3: Multiply the fractional part of the sequence $\{K_x\}$, $\{K_y\}$, $\{K_z\}$ by 256, take the integer part of the result, and convert each value into an 8-bit binary number, and then form the matrix C_x, C_y, C_z of the same size as the encrypted speech file matrix E_q'.

Step 4: Perform bitwise XOR processing on the elements in C_x, C_y, C_z and the corresponding elements in the encrypted speech file matrix E_q', to obtain the decrypted speech file S_q, and return it to the query user.

V. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

A. EXPERIMENTAL ENVIRONMENT

In order to evaluate the performance of the proposed method, this paper uses the speech data from THCHS-30 [31],

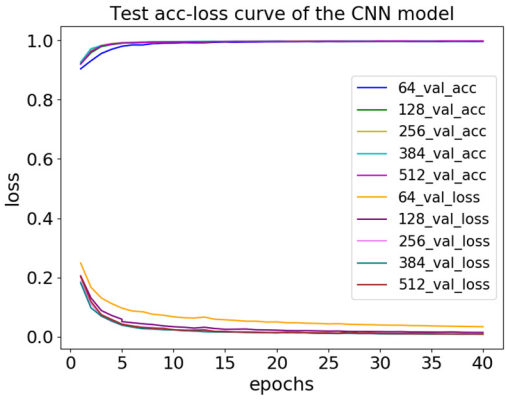
a Chinese speech database released by the center for speech and language technology (CSLT) of Tsinghua University to conduct experiments, the speech sampling frequency is 16kHz and the sampling size is 16bits, the content is 1,000 news fragments with different contents, the length of each speech is about 10 seconds, and the total length of all speech in the database is about 30 hours. The experiment in this paper randomly selects 10 types of speech with different contents, and performs various content preserving operations such as amplitude adjustment, noise addition, re-sampling, re-quantization and MP3, etc., a total of 3,060 speeches are obtained for training, making the features extracted by the system model more robust. In the test experiments of recall rate, precision rate and average precision, randomly select 1,000 speeches in the database to evaluate, and in the retrieval efficiency experiment, randomly select 10,000 speeches in the database to test.

The experimental hardware platform is: Intel(R) Core(TM) i7-10710U CPU @1.10GHz 1.61GHz, the memory is 16GB. The software environment is: Windows 10, MATLAB, JetBrains PyCharm Community Edition 2019.2.4 x64.

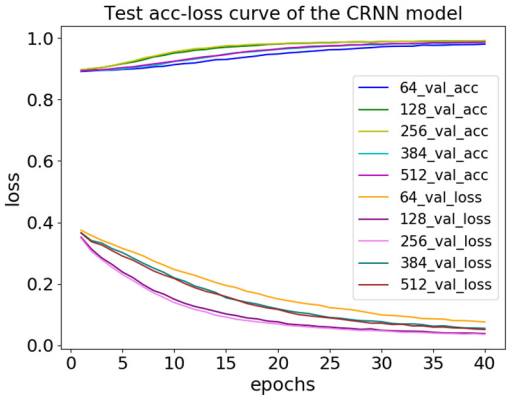
B. PERFORMANCE ANALYSIS OF DEEP HASHING CODING MODEL

In order to extract representative feature vectors and generate high-quality deep hashing binary codes, the performance of coding model is very important. The dimension of the hash layer corresponds to the length of the generated deep semantic hash code, with different hash code lengths, the accuracy and loss of the model will show certain changes. Therefore, under different hash coding lengths, this paper evaluates the test accuracy and test loss of the proposed two deep hashing coding models: CNN and CRNN coding model. Fig. 8 shows the test accuracy-loss curve of the CNN/CRNN coding model under different hash coding lengths. Fig. 9 shows the test loss curve of the CNN/CRNN coding model under different hash coding lengths.

It can be seen from Fig. 8 and Fig. 9 that the test accuracy of the proposed CNN and CRNN models are both close to 1 under different hash coding length, and the performance is excellent. In comparison, the CNN model converges faster than the CRNN model, when the training batch is 10, the CNN model has basically converged, the test accuracy of the model basically does not change, and the test loss is only slightly reduced. The CRNN model basically converges when the batch is 30, the test accuracy is close to 1 and no significant change, and the test loss is also reduced by a small amount. This is because the RNN network structure is added to the CRNN model, the RNN training process is continuous cycle, and the previous data constantly affect the subsequent output, so its convergence speed is slower than the CNN model. In addition, although the accuracy of the two coding models are both tends to 1 and the loss tends to 0, in terms of test loss, the loss value of the CNN model is smaller than the CRNN model. In the CNN model, when the length of hash code is 64, the loss value of the model is relatively



(a) CNN coding model



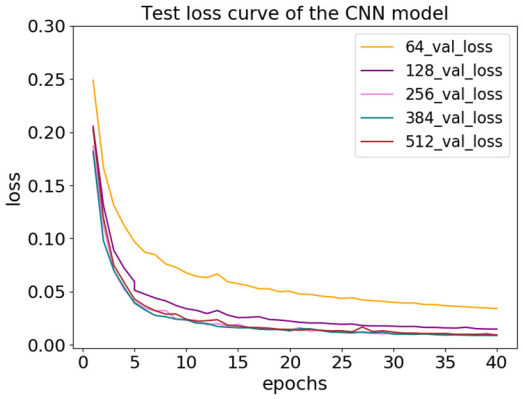
(b) CRNN coding model

FIGURE 8. Test accuracy-loss curve of CNN/CRNN models under different coding lengths.

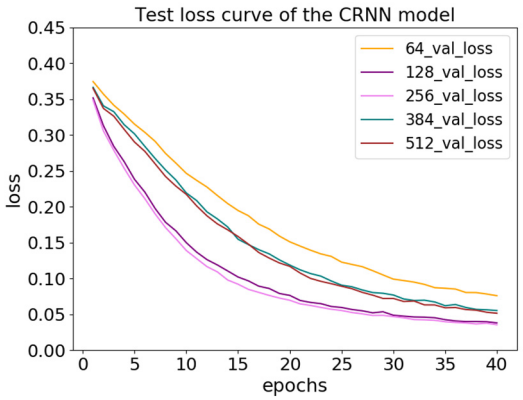
large, in other lengths, the loss value of the model is very small and approaching 0, and has high accuracy. In the CRNN model, when the hash code length is 128/256, the loss value of the model is smaller than other lengths, and the accuracy is relatively high.

C. mAP ANALYSIS

In order to further test the performance of the proposed model, this paper uses the mean Average Precision (mAP) to evaluate and analyze. At the same time, in order to test the robustness and collision resistance of the deep hashing binary code generated by the proposed coding model, before the test experiment, four content preserving operations (CPOs), such as amplitude reduction (−3dB), amplitude increase (+3dB), MP3 compression (MP3) and resampling 8-16kbps (R.Q) are performed on the test speech, and obtain a total of 4,000 speech data. The experiment uses the deep hashing binary code generated by the speech data processed by CPOs for testing, firstly, the proposed CNN/CRNN coding model is used to encode the speech that processed by various CPOs and calculate its AP value, and then the mean Average Precision (mAP) is calculated by the AP value. Table 1 shows



(a) CNN coding model



(b) CRNN coding model

FIGURE 9. Test loss curve of CNN/CRNN models under different coding lengths.

TABLE 1. mAP of CNN/CRNN models under different hash code lengths.

Models	Hash code length				
	64	128	256	384	512
CNN	0.8817	0.9092	0.9322	0.9494	0.9542
CRNN	0.8833	0.9103	0.9436	0.9556	0.9594

the mAP values obtained by the CNN/CRNN model under different coding lengths. The mAP calculation formula is shown in (17) and (18).

$$AP(q) = \frac{\sum_{k=1}^{n} (P(k) \times rel(k))}{m}, \quad rel(k) \in \{0, 1\} \tag{17}$$

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{18}$$

where n is the total number of speech in the database, m is the total number of speech related to the query speech, Q is the total number of the queries, $rel(k)$ indicates whether the speech at position k is related to the query speech, the correlation is 1, and the irrelevance is 0.

It can be seen from Table 1, CNN and CRNN coding models have achieved better mean average precision under different hash code lengths, which indicates that the two proposed deep hashing coding models have good coding ability for input speech data and can ensure well query performance. In general, the mAP value increases as the length of the hash code increases, because the longer the hash code, the better the feature representation ability of the input speech data, so the query precision is higher. In contrast, under the same hash code length, the mAP value of the CRNN coding model is slightly larger than that of the CNN coding model, which is because CRNN coding model has the advantage of processing temporal features and has better representation ability for temporal features. The experiment balances the model test accuracy, mean Average Precision (mAP), and system retrieval efficiency, the CNN/CRNN coding model with the hash code length of 384 is used for subsequent test evaluation.

D. SYSTEM RETRIEVAL PERFORMANCE ANALYSIS

In order to evaluate the retrieval performance of the proposed method, the experiment uses recall rate (R), precision rate (P) and F1 score (F_1) as evaluation indicators to test and analyze. Recall rate (R) represents the proportion of successful retrieved samples among all samples related to the query, its mathematical expression is:

$$R = \frac{TP}{TP + FN} \times 100\% \tag{19}$$

where TP represents the number of retrieved samples related to the query, FN represents the number of missed samples related to the query.

The precision rate (P) represents the proportion of the samples that are actually related to the query among all the retrieved samples, its mathematical expression is:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{20}$$

where FP represents the number of retrieved samples that are not related to the query.

The experiment shows that there is an anti-dependency relationship between recall rate (R) and precision rate (P), if the recall rate increases, the precision rate will decrease, and vice versa. Therefore, in order to balance recall rate and precision rate, and further test the retrieval performance, this paper uses F1 score (F_1) for evaluation. F1 score is a weighted average of recall rate and precision rate, with the maximum value of 1 and the minimum value of 0. The calculation formula of F1 score is shown in (21):

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{21}$$

The recall rate (R), precision rate (P) and F1 score (F_1) of the proposed method are tested by using the CNN/CRNN coding model with hash code length of 384, and compared with the retrieval performance of existing methods [14]–[17], Table 2 shows the experimental results.

TABLE 2. Comparison of retrieval performance under different network Models.

Methods	Models	Recall rate (R)	Precision rate (P)	F1 score (F_1)
Ref. [14]	CNN	91.00%	91.00%	0.9100
	CRNN	91.00%	91.00%	0.9100
	Inception-v3 CNN	95.00%	95.00%	0.9500
	Inception-v3 CRNN	96.00%	96.00%	0.9600
Ref. [15]	Gated-CRNN	58.90%	56.5%	0.5770
Ref. [16]	FPNN	96.23%	95.75%	0.9599
Ref. [17]	CNN	-	97.53%	-
	SC_{fc2}-CNN	-	96.40%	-
	MSC_{conv4}-CNN	-	99.49%	-
	MSC_{conv5}-CNN	-	100%	-
Proposed Method	CNN	99.80%	100%	0.9990
	CRNN	100%	99.90%	0.9995

TABLE 3. Comparison results of recall rate (R) and precision rate (P) under different CPOs.

Methods		Recall rate (R)				Precision rate (P)			
		-3dB	+3dB	MP3	R.Q	-3dB	+3dB	MP3	R.Q
Ref. [2]		-	-	96.0%	98.0%	-	-	96.0%	97.0%
Ref. [3]		100%	100%	100%	100%	100%	100%	98.0%	100%
Ref. [25]	Log Mel	100%	100%	100%	100%	100%	100%	99.9%	100%
	MFCC	100%	100%	100%	99.9%	100%	100%	100%	100%
Proposed Method	CNN	99.6%	100%	99.8%	100%	100%	100%	100%	100%
	CRNN	99.8%	100%	100%	100%	100%	100%	99.5%	100%

As can be seen from Table 2, the experimental results of the two coding models (CNN/CRNN) proposed in this paper achieve the best recall rate, precision rate and F1 score, indicating that the proposed method has better retrieval performance. In contrast, except for the low recall rate, precision rate and F1 scores of [15], other comparative literatures have good performance, and their F1 scores are all above 0.90, especially under the MSC_{conv5}-CNN model in [17], the precision rate has reached 100%. Therefore, under the proposed two deep hashing coding models designed, we can get the deep hashing binary code with better semantics and discriminability, so that the whole retrieval system has better recall rate, precision rate and F1 score.

In addition, in order to test the robustness of the deep hashing binary code generated by the two proposed coding models (CNN/CRNN), this paper uses the deep hashing binary code generated by the speech data after the four CPOs operations, such as amplitude reduction (−3dB), amplitude increase (+3dB), MP3 compression (MP3) and resampling 8-16kbps (R.Q) for evaluation and analysis, and compared with the perceptual hash method [2], [3] and the deep hashing method [25] commonly used in the field of encrypted speech retrieval.

Table 3 shows the comparison results of recall rate (R) and precision rate (P).

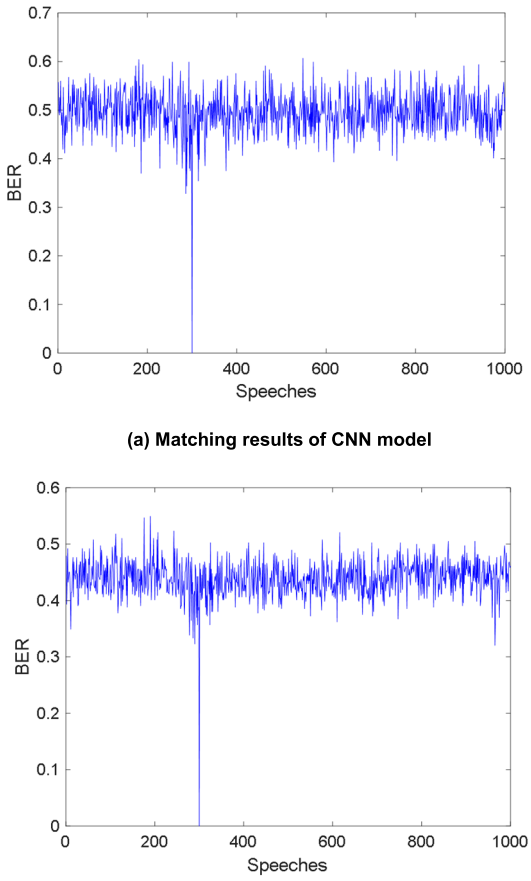
It can be seen from Table 3, the proposed two deep hashing coding models (CNN/CRNN) still have well recall rate and precision rate under four content preserving operations, which shows that the proposed method has good robustness and collision resistance. Compared with the perceptual hashing method [2,3] commonly used in the field of encrypted speech retrieval, the performance of the proposed method is slightly better, because the proposed method uses the deep neural network model (CNN/CRNN) to extract deep semantic features, which has strong semantic representation. At the same time, compared with the deep hashing method in [25], the performance of the proposed method is basically the same, in most cases, the test results of recall rate and precision rate have reached 100%. Therefore, the deep hashing binary code generated by the two proposed coding models (CNN/CRNN) has better robustness, and still guarantees the good recall rate and precision rate under various content preserving operations.

In addition, the experiment also tests the user query, the experiment randomly selects the 300th of the 1,000 test speeches as the query speech for matching retrieval, respectively obtains its deep hashing binary code under the CNN/CRNN model, and adopts the ''two-stage'' classification retrieval strategy to search, and matches with the hash sequence in the deep hashing index table. At the same time, the matching threshold of the two coding models is set to 0.26. Fig. 10 shows the matching results.

It can be seen from Fig. 10 that under the CNN/CRNN model, the normalized Hamming distance (BER) of the

TABLE 4. Comparison results of retrieval efficiency of different retrieval methods.

Hash lengths	Speech length(s)	Traversal retrieval (s)	"Two-stage" classification Retrieval (s)
64	10	0.0591	0.0085 = 0.0465
128	10	0.0720	0.0099 = 0.0479
256	10	0.0880	0.0380 + 0.0122 = 0.0502
384	10	0.1020	0.0164 = 0.0544
512	10	0.1201	0.0210 = 0.0590



(a) Matching results of CNN model

(b) Matching results of CRNN model

FIGURE 10. Matching results of the query speech.

300th speech is less than the set threshold 0.26, and the retrieval is successful; while the BER values of other speeches are all above 0.26, the distance is large and the matching fails. Therefore, the proposed method has better retrieval effect for user query.

E. SYSTEM RETRIEVAL EFFICIENCY ANALYSIS

In order to test the performance of the proposed ''two-stage'' classification retrieval method, the ''two-stage'' classification retrieval method is compared with the traditional traversal retrieval method under different hash code lengths. The experiment randomly selects 10,000 speeches in the THCHS-30 speech database for comparative analysis, they belong to 10 categories, and the number of speeches in each category is 1000. Table 4 shows the comparison of the retrieval efficiency between the traditional traversal retrieval method and the proposed ''two-stage'' classification retrieval method under different hash encoding length.

It can be seen from Table 4, under the same hash code length, the retrieval efficiency of the proposed ''two-stage'' classification retrieval method is higher than that of the traditional traversal retrieval method, and with the increase of hash code length, the difference is more and more obvious. In the retrieval schedule of the ''two-stage'' classification retrieval, the preceding 0.0380 represents the retrieval time of category hash *h1* from 10,000 speeches, that is, the time to retrieve the candidate set of the same category as the query. The following 0.0085, 0.0099, 0.0122, 0.0164 and 0.0210 represent the time used to retrieve the semantic feature hash *h2* from the obtained candidate set under different hash encoding lengths, that is, the time of semantic feature matching; The final 0.0465, 0.0479, 0.0502, 0.0544 and 0.0590 represent the total time of ''two-stage'' classification retrieval method under different hash code lengths. In summary, the proposed ''two-stage'' classification retrieval method is more efficient than the traditional traversal retrieval method, and with the increase of feature hash length, the efficiency improvement is more obvious.

In addition, in order to evaluate the retrieval efficiency of the proposed method, the experiment compares the average retrieval time (feature learning + retrieval matching) between the proposed method and the method in [2], [3], [25], in the proposed method, the CNN/CRNN coding model with the length of 384 bits is used for the experiment. Table 5 shows the comparison results of retrieval efficiency between the proposed method and the existing methods.

As can be seen from Table 5, the retrieval efficiency of the proposed method is higher than that of [2], [25], the average running time is 0.4394 seconds in the CNN coding model and 0.5274 seconds in the CRNN coding model, which is about 7.9 times of the method in [2] and 1.5 times of that in [25], this is because [2] first extracts manual features, and then constructs perceptual hash based on manual features, while the proposed method integrates feature learning and hash coding into a whole module, thereby reducing the hash code generation time. Compared with the [25], the proposed method adopts the ''two-stage'' classification retrieval strategy, which improves the efficiency of hash matching to a

TABLE 5. Comparison results of retrieval efficiency between proposed method and existing methods.

Methods		Frequency (GHz)	Speech length (s)	Average time (s)
Ref. [2]		3.20	4	4.2032
Ref. [3]		2.50	4	0.1814
Ref. [25]	Log Mel	2.20	10	0.7465
	MFCC	2.20	10	0.7964
Proposed Method	CNN	2.50	10	0.4394
	CRNN	2.50	10	0.5274

certain extent. Compared with the [3], the retrieval efficiency of the proposed method is slightly lower than that of the [3], this is because the [3] first classifies the speech data, and then compresses the generated hash sequence through the stroke length compression technology, thereby shortening the matching time, however, the final perceptual hash sequence generation process is more complicated than the deep hashing construction process in this paper, and its test speech length is shorter than this paper. Therefore, the CNN/CRNN coding model proposed in this paper has better retrieval efficiency and fully meets the retrieval requirements of the encrypted speech retrieval system.

To sum up, the proposed two deep hashing coding models (CNN/CRNN) have higher test accuracy and lower test loss, which shows that the proposed model can fit the input speech data well, and able to perform good deep hashing coding operation on the input speech. At the same time, the deep hashing binary code constructed in this paper has better recall rate, precision rate and F1 score in encrypted speech retrieval task, which is fully suitable for the application of retrieval task. Moreover, the proposed ''two-stage'' classification retrieval method has higher retrieval efficiency than the traditional traversal retrieval method, and with the increase of hash code length, the improvement of retrieval efficiency will be more obvious.

VI. CONCLUSION AND FUTURE WORK

In this paper, based on the powerful self-learning ability of the deep learning method and the advantages of fast retrieval speed and high precision of the deep hashing method, a classification retrieval method for encrypted speech based on DNN and deep hashing is proposed. The proposed method improves the retrieval efficiency and retrieval accuracy of the existing content-based encrypted speech retrieval system, and solves the problem of poor semantics of traditional manual perceptual features. The main work of this paper is as follows: 1) The proposed method designs two deep hashing coding models: CNN and CRNN coding model to perform advanced semantic feature extraction and deep hashing construction on input speech data, which can generate high-quality deep hashing binary codes, breaking through the limitations of traditional manual perceptual features; 2) By using the ''two-stage'' classification retrieval method, the retrieval efficiency and retrieval accuracy are further improved; 3) The Rossler chaotic map encryption method is used to ensure the security of cloud speech data. The experimental results show that the proposed two deep hashing coding models have better fitting degree to the input speech data, and the constructed deep hashing binary code has high recall rate, precision rate and retrieval efficiency in encrypted speech retrieval task.

The shortcoming of this paper is that although the ''two-stage'' classification retrieval strategy is adopted for retrieval, it reduces the candidate set and improves the retrieval efficiency, but in the case of large amount of data, the single-table query limits the retrieval efficiency to a certain extent, therefore, an efficient index structure is essential.

In future work, we will try to establish an efficient index structure, which can realize parallel query of multiple tables and achieve efficient encrypted speech retrieval task in the big data environment.

REFERENCES

[1] C. Glackin, G. Chollet, N. Dugan, N. Cannings, J. Wall, S. Tahir, I. G. Ray, and M. Rajarajan, ''Privacy preserving encrypted phonetic search of speech data,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 6414–6418.

[2] S. He and H. Zhao, ''A retrieval algorithm of encrypted speech based on syllable-level perceptual hashing,'' *Comput. Sci. Inf. Syst.*, vol. 14, no. 3, pp. 703–718, 2017.

[3] Q.-Y. Zhang, Z.-X. Ge, Y.-J. Hu, J. Bai, and Y.-B. Huang, ''An encrypted speech retrieval algorithm based on chirp-Z transform and perceptual hashing second feature extraction,'' *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 6337–6361, Mar. 2020.

[4] H. Zhao and S. He, ''A retrieval algorithm for encrypted speech based on perceptual hashing,'' in *Proc. 12th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Changsha, China, Aug. 2016, pp. 1840–1845.

[5] C. Shi, X. Li, and H. Wang, ''A novel integrity authentication algorithm based on perceptual speech hash and learned dictionaries,'' *IEEE Access*, vol. 8, pp. 22249–22265, Jan. 2020.

[6] B. Zhang and J. Lin, ''An efficient content based music retrieval algorithm,'' in *Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS)*, Xiamen, China, Jan. 2018, pp. 617–620.

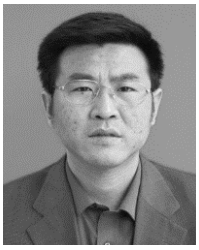
[7] F. Zeng, S. Hu, and K. Xiao, ''Deep hash for latent image retrieval,'' *Multimedia Tools Appl.*, vol. 78, no. 22, pp. 32419–32435, Nov. 2019.

[8] Q. Qin, Z. Wei, L. Huang, J. Nie, and X. Ji, ''A novel deep hashing method with top similarity for image retrieval,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 2067–2071.

[9] L. Weng, L. Ye, J. Tian, J. Cao, and J. Wang, ''Random VLAD based deep hashing for efficient image retrieval,'' 2020, *arXiv:2002.02333*. [Online]. Available: http://arxiv.org/abs/2002.02333

[10] Y. Li, L. Wan, T. Fu, and W. Hu, ''Piecewise supervised deep hashing for image retrieval,'' *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 24431–24451, Jan. 2019.

[11] L. Fan, Q.-Y. Jiang, Y.-Q. Yu, and W.-J. Li, "Deep hashing for speaker identification and retrieval," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2908–2912.
[12] Y. Shan, M. Liu, Q. Zhan, S. Du, J. Wang, and X. Xie, "Speech recognition based on deep tensor neural network and multifactor feature," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Lanzhou, China, Nov. 2019, pp. 650–654.
[13] E. ElMaghraby, A. Gody, and M. Farouk, "Noise-robust speech recognition system based on multimodal audio-visual approach using different deep learning classification techniques," *Egyptian J. Lang. Eng.*, vol. 7, no. 1, pp. 27–42, Apr. 2020.
[14] C. Bartz, T. Herold, H. Yang, and C. Meinel, "Language identification using deep convolutional recurrent neural networks," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, Oct. 2017, pp. 880–889.
[15] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 121–125.
[16] N. M. Patil and M. U. Nemade, "Content-based audio classification and retrieval using segmentation, feature extraction and neural network approach," *Adv. Comput. Commun. Comput. Sci.*, vol. 924, pp. 263–281, May 2019.
[17] C.-Y. Wang, T.-C. Tai, J.-C. Wang, A. Santoso, S. Mathulaprangsan, C.-C. Chiang, and C.-H. Wu, "Sound events recognition and retrieval using Multi-Convolutional-Channel sparse coding convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1875–1887, Jan. 2020.
[18] B. Kim and B. Pardo, "Improving content-based audio retrieval by vocal imitation feedback," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 4100–4104.
[19] Q.-Y. Zhang, L. Zhou, T. Zhang, and D.-H. Zhang, "A retrieval algorithm of encrypted speech based on short-term cross-correlation and perceptual hashing," *Multimedia Tools Appl.*, vol. 78, no. 13, pp. 17825–17846, Jan. 2019.
[20] P. Qin, J. Chen, K. Zhang, and R. Chai, "Convolutional neural networks and hash learning for feature extraction and of fast retrieval of pulmonary nodules," *Comput. Sci. Inf. Syst.*, vol. 15, no. 3, pp. 517–531, 2018.
[21] Y. Li, X. Kong, and H. Fu, "Exploring geometric information in CNN for image retrieval," *Multimedia Tools Appl.*, vol. 78, no. 21, pp. 30585–30598, Nov. 2019.
[22] J. Hung, J. S. Lin, and P. J. Wu, "Employing robust principal component analysis for noise-robust speech feature extraction in automatic speech recognition with the structure of a deep neural network," *Appl. Syst. Innov.*, vol. 1, no. 3, pp. 1–14, Aug. 2018.
[23] R. K. Nayyar, S. Nair, O. Patil, R. Pawar, and A. Lolage, "Content-based auto-tagging of audios using deep learning," in *Proc. Int. Conf. Big Data, IoT Data Sci. (BID)*, Pune, India, Dec. 2017, pp. 30–36.
[24] Dhiraj, R. Biswas, and N. Ghattamaraju, "An effective analysis of deep learning based approaches for audio based feature extraction and its visualization," *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 23949–23972, Sep. 2019.
[25] Q. Zhang, Y. Li, Y. Hu, and X. Zhao, "An encrypted speech retrieval method based on deep perceptual hashing and CNN-BiLSTM," *IEEE Access*, vol. 8, pp. 148556–148569, Aug. 2020.
[26] H. Shu, Y. Song, and H. Zhou, "Time-frequency performance study on urban sound classification with convolutional neural network," in *Proc. IEEE Region 10 Conf. (TENCON)*, Jeju, South Korea, Oct. 2018, pp. 1713–1717.
[27] J. Yu, K. Markov, and T. Matsui, "Articulatory and spectrum information fusion based on deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 742–752, Apr. 2019.
[28] X. Guo, L. F. Polanía, and K. E. Barner, "Audio-video emotion recognition in the wild using deep hybrid networks," 2020, *arXiv:2002.09023*. [Online]. Available: http://arxiv.org/abs/2002.09023
[29] M. Rahmani and F. Razzazi, "An LSTM auto-encoder for single-channel speaker attention system," in *Proc. 9th Int. Conf. Comput. Knowl. Eng. (ICCKE)*, Mashhad, Iran, Oct. 2019, pp. 110–115.
[30] J. Nan, "Study on coupled synchronization of Rossler chaotic system with single state variable," *J. Taiyuan Univ. (Natural Ed.)*, vol. 37, no. 3, pp. 13–17, Mar. 2019.
[31] D. Wang and X. Zhang, "THCHS-30: A free Chinese speech corpus," 2015, *arXiv:1512.01882*. [Online]. Available: http://arxiv.org/abs/1512.01882



QIUYU ZHANG (Member, IEEE) graduated from the Gansu University of Technology, in 1986. He is currently a Professor/Ph.D. Supervisor with the School of Computer and Communication, Lanzhou University of Technology. He is also the Vice Dean with the Gansu Manufacturing Information Engineering Research Center. His research interests include network and information security, information hiding and steganalysis, image understanding and recognition, and multimedia communication technology. He is a member of ACM and a Senior Member of CCF.

XUEJIAO ZHAO received the B.S. degree in digital media technology from the Lanzhou University of Arts and Science, Gansu, China, in 2018. She is currently pursuing the master's degree with the School of Computer and Communication, Lanzhou University of Technology. Her research interests include audio signal processing and application, information security, multimedia authentication, and retrieval techniques.

YINGJIE HU received the M.S. degree in computer software and theory from Lanzhou University, Lanzhou, China, in 2011. She is currently a Lecturer with the School of Computer and Communication, Lanzhou University of Technology. Her research interests include multimedia information processing and application, information security, multimedia authentication, and retrieval techniques.