IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 June 2022, accepted 20 July 2022, date of publication 26 July 2022, date of current version 29 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3194051

RESEARCH ARTICLE

Steganalysis of Compressed Speech Based on Global and Local Correlation Mining

JIAWEI WANG, JIE YANG, FEIPENG GAO, AND PENG XU
Jiyang College, Zhejiang A & F University, Zhuji 311800, China

Corresponding author: Jie Yang (yangjie_work126@126.com)

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LQ20F020004, in part by the Scientific Research Startup Fund for Talent of Jiyang College under Grant JY2018RC04, and in part by the Zhejiang University Student Science and Technology Activity Plan and New Seedling Talent Plan under Grant 2021R412038.

ABSTRACT Most of the existing steganalysis methods are designed for specific steganography methods in low-bit-rate compressed speech stream and lack of generalization ability. In practical applications, the steganography methods in compressed speech are various and cannot be predicted in advance. We can only employ numerous possible steganalysis method to detect, which is laborious and time-consuming, and cannot achieve real-time detection. Therefore, it is necessary to develop a general steganalysis method that can detect multiple steganography methods simultaneously for compressed speech. To this end, a steganalysis method based on global and local correlation mining is proposed in this paper. Firstly, a codeword distributed embedding module is introduced to transform the compressed codewords into a compact feature representation. Then, global-guided correlation mining module and local-guided correlation mining module are used to extract the correlation change before and after steganography in the view of global and local. Finally, the detection results can be obtained by the full connection layers. Experimental results show that the proposed method can reach a better detection performance than the existing steganalysis methods at different embedding rates and speech lengths.

INDEX TERMS Correlation mining, compressed speech, deep learning, steganography, steganalysis.

I. INTRODUCTION

Information hiding, also known as steganography, is a technology that utilizes the sensory redundancy of human sense organ to embed secret information into the carrier in an undetected way. The choice of carriers is diverse, such as videos [1], [2], images [3], [4], audios [5], [6], speeches [7], [8], texts [9], [10], and so so. In recent years, with the rapid development of network information technology, voice over IP (VoIP) gradually plays an important role in daily communication of human beings. VoIP is highly dynamic, real-time and time-varying. It is an excellent carrier of information hiding. Based on the compressed voice in VoIP, we can build a covert communication channel conveniently. However, like every coin has two sides, steganography poses a huge threat to cyber security. Hackers and criminals can also transmit illegal message in the network through steganography based on compressed speech. To eliminate these security threats, scholars have carried out research on countermeasure technology, called steganalysis.

In practical applications, low-bit-rate speech codecs, such as the analysis-by-synthesis linear predictive coding (AbS-LPC), are widely used for compressed speech encoding. Therefore, most steganography methods based lon compressed speech utilize AbS-LPC low-bit-rate speech codecs to achieve covert communication [11]. According to the different embedding positions of secret information, these information hiding methods can be mainly summarized as three categories. The first category utilizes the pitch filter to perform information hiding [12]–[15]; the second category employs the LPC filter to embed secret information [16]–[19]; and the third category directly modifies the codeword in the compressed speech stream [20]–[25].

So far, researchers have studied the corresponding steganalysis methods for each kind of specific steganography methods, and achieved satisfactory detection performance.

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

However, the general detection of various steganography algorithms is a problem that has not been well solved. In practical applications, the possible steganography algorithms are diverse and cannot be predicted in advance. This determines that the application scenarios of steganalysis methods that can only detect a single specific steganography algorithm are relatively limited.

In addition, some researchers have proposed some steganalysis methods based on time-domain features, such as Mel-frequency cepstral coefficients (MFCC) [26]. Low-bit-rate compressed speech loses a lot of redundant information in speech coding, and the method based on time-domain feature extraction cannot well reflect the characteristics of compressed speech. Therefore, these detection methods have poor detection performance for low-bit-rate compressed speech. It is necessary to develop a general steganalysis methods which can detect multiple different steganography algorithms in compressed speech.

Human speech is usually a certain form of continuous language expression. Hence, there is correlation between compressed speech sequences. This correlation exists not only between adjacent words, but also between adjacent, and even sentences with a longer receptive span. Embedding secret information in compressed speech will affect the implicit correlation. The key of steganalysis is to capture the change of correlation before and after steganography. The target of this paper is to achieve general steganalysis, which determines that we must effectively capture all kinds of correlation changes. Therefore, a steganalysis method based on global and local correlation mining is proposed here. The contribution of this paper can be summarized as follows:

1) Analyzing the characteristic of steganography in compressed speech, we propose Global-guided Correlation Mining (GCM) and Local-guided Correlation Mining (LCM). These two proposed modules conduct steganography-sensitive correlation feature extraction in view of global and local respectively.
2) Based on the above modules, an efficient steganalysis network for steganography in compressed speech is proposed. For detecting multiple steganography methods, the proposed steganalysis method can reach an accuracy of 80%, outperforming the existing methods. The experimental results reflect on the effectiveness of our proposed method.

II. RELATED WORK

The target of AbS-LPC is minimizing the error between the synthetic speech signal and the original speech signal. To this end, two processes are performed: pitch synthesis filtering analysis and LPC synthesis filtering analysis. The purpose of pitch synthesis filtering analysis is to capture the long-term correlations of the speech signal, whereas that of LPC synthesis filtering analysis is to capture the short-term correlations. To a certain extent, steganography in low-bit-rate speech coding process will disturb the original correlations of the speech signal. Therefore, the steganalysis methods are designed based on the correlation changes before and after steganography.

In view of the steganalysis of the information hiding based on the pitch filter, many support vector machine-based (SVM-based) methods have been proposed. The key step of these methods is to construct an effective feature vector for SVM training. Li *et al.* [27] found that steganography through modulating pitch period search range [12] would inevitably change the pitch delay values of adjacent speech frames. Based on this change, a steganography-sensitive codebook correlation feature vector was obtained with the help of a codebook correlation network. In addition, symbiotic characteristics [28], calibrated second-order differential Markov transition probability feature [29], and calibrated probability distributions of the difference feature [30] were successively presented to conduct steganalysis.

For the steganalysis of the information hiding based on the LPC filter, Li *et al.* [31] found that the quantization index modulation (QIM) steganography [16] would change the LPC indexes. To quantify these correlated characteristics of the LPC indexes, the first-order Markov transition probabilities that change the most were used to form the categorical feature vector for each LPC index. In reference [32], they constructed a quantization codeword correlation network model based on the transition probabilities of intra-frame and inter-frame correlation of the crossed LPC indexes. In addition, Bayesian network-based (BN-based) method [33] and neural network (NN-based) methods [34]–[36] have also been put forward. In reference [33], a steganography-sensitive Codeword Bayesian Network (CBN) was proposed based on the correlation changes of codeword spatiotemporal transition, and Bayesian inference was used for classification. In reference [34], Lin *et al.* proposed a steganalysis network based on recurrent neural network (RNN). To reduce the time cost and improve the detection accuracy of the NN-based method, Yang *et al.* [35] mapped vector quantization codewords into a semantic space and utilized one hidden layer to extract the corrections between the codewords. Yang *et al.* [36] developed multi-channel convolutional sliding windows to analyze the correlations between a given frame and its neighboring frames.

In terms of the steganalysis of the information hiding based on codewords modification, Tian *et al.* [37], [38] proposed two SVM-based steganalysis methods. In reference [37], they firstly proved that the probabilities of all speech parameter values would tend to be equal along with the increase of the embedding rate. Then, they chose the best performing feature from four probabilistic features, i.e., histogram distribution, differential histogram distribution, Markov transition matrix and differential Markov transition matrix, to implement steganalysis. In reference [38], the zero-crossing count (ZCC) statistical features and average Mel-frequency cepstral coefficients feature of inactive speech frames were employed to construct the feature vector for SVM training. The ZCC statistical features consist of average ZCC of inactive frames,

the ratio between the average ZCC, the difference between the average ZCC and their calibrated versions.

Up to now, most of the steganalysis methods in compressed speech are designed for a specific kind of steganography methods. A few steganalysis methods used for detecting multiple steganography methods have been proposed in recent years. Yang *et al.* [39] proposed a NN-based common steganalysis method, in which a codeword Bayesian network (CBN) was constructed based on the whole codewords in speech stream. CBN employed Bayesian inference to implement classification. Hu *et al.* [40] proposed a novel deep model named as steganalysis feature fusion network (SFFN), which can detect the first two kinds of steganography methods simultaneously. Li *et al.* [41] proposed a common detection method based on codeword embedding, Bi-LSTM and CNN attention mechanisms, named CBCA. CBCA can detect the three kinds of steganography methods simultaneously with better performance than the previous two steganalysis methods. However, there is still room for improvement in detection accuracy. In this paper, a common steganalysis method is proposed for low-bit-rate compressed speech which takes full advantage of the global and local correlations in compressed speech.

III. METHOD

There are strong correlation patterns between codewords in the compressed codeword stream. These correlation patterns can be classified as intra-word correlation, intra-sentence (cross-word) correlation, and cross-sentence correlation. And these correlations are likely to be attenuated when hidden data is embedded in the original compressed codeword stream. Thus, it is inspired that these correlation patterns can be considered a promising indicator to extract features in the codeword stream for steganalysis. To this end, a steganalysis method based on global and local correlation mining is proposed, which will be comprehensively described in the rest of this section.

A. ARCHITECTURE

The overall architecture of our proposed steganalysis method is shown in Fig. 1. The model consists of four components, codeword distributed embedding, global-guided correlation mining, local-guided correlation mining and prediction module.

Compressed codeword is highly abstract. Before being fed into the network, codewords must be converted into a form that is conductive to feature extraction by the neural network. This can be realized by introducing the codeword distributed embedding module [41]. The compressed speech stream containing T frames with N codewords per frame can be denoted as a matrix X with a size of $T \times N$:

$$X = \begin{bmatrix} X_{1,1} & \cdots & X_{1,N} \\ \vdots & \ddots & \vdots \\ X_{T,1} & \cdots & X_{T,N} \end{bmatrix} \tag{1}$$

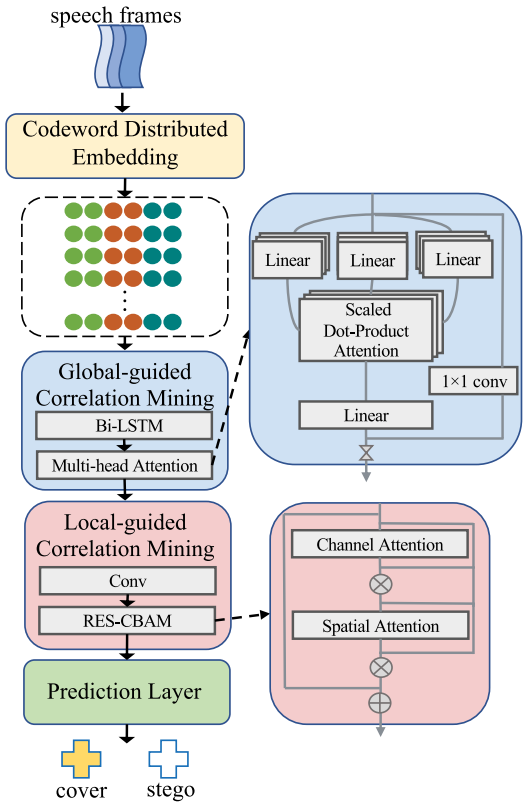

FIGURE 1. The overall architecture of our proposed steganalysis method. Four networks are built in our method: codeword distributed embeddding, global-guided correlation mining, local-guided correlation mining, and prediction layer.

The one-hot coding is used to map each codeword and then the one-hot vector of codewords in each frame are concatenated together to obtain a new matrix $\mathbf{M}$. Since $\mathbf{M}$ is a sparse matrix, we transform it to a more compact form. Three different mapping dictionaries are constructed to achieve this. The dimension of them are $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ respectively. By this, we can obtain three mapping matrices and concatenate them to get the final embedding matrix.

Based on the embedding matrix, we conduct global and local correlation mining. The global-guided correlation module consists of Bi-LSTM and multi-head self-attention. The Bi-LSTM structure is a near-ideal solution for capturing the global sequence correlation feature. In this paper, we utilize a two-layer Bi-LSTM network to mine the sequential correlation between words and sentences in compressed speech stream. To alleviate the gradient vanishing, we introduce a skip connection composed of 1×1 convolution structure. Besides, a multi-head self-attention mechanism is utilized to make the detection network focus on several different representation sub-spaces.

The local-guided correlation module consists of depthwise and pointwise convolution and convolutional block attention mechanism. Different from LSTM, convolution structure have access to modeling the local correlation by controlling a sliding window (also called convolution kernel). Thus,

we use a convolution-based structure to capture the correlation change before and after steganography. In an attempt to be able to improve the representation of key features and the stability of the network, we present a convolutional block attention mechanism with residual blocks.

Eventually, the final feature matrix R is input to the fully connected layer, and the obtained probability value F is used to determine whether the detected speech coded word stream contains steganographic information:

$$\text{Ruturn Result} = \begin{cases} \text{stege}, & F \geq 0.5 \\ \text{cover}, & F < 0.5 \end{cases} \tag{2}$$

B. CODEWORD DISTRIBUTED EMBEDDING

Inspired by natural language processing, we came up with the idea of constructing a mapping from compressed codewords to compact and powerful feature representations. By this, we can transform the codewords in compressed speech stream into a form that facilitates deep feature extraction with the help of neural-based models. Also, such a form is beneficial for mining the correlations in the view of global and local.

The approach of one-hot encoding is to use N-bit status registers to encode N states, each of which has its independent register bits and only one of which is valid. That is to say, only one bit is 1 and the rest are 0. If the value of the codeword is b and its encoding range is 0 to 2^{a-1}, then the unique thermal encoding of the codeword can be expressed as:

$$L_{\text{one - hot}}(X) = \left\{0 \leq i \leq 2^{a-1} | h_i == i ? 1 : 0\right\} \tag{3}$$

We stitch the codewords in the phonetic codeword matrix $\mathbf{X}$ after unique thermal encoding to obtain the matrix $\mathbf{M}$. $\mathbf{M}$ is the mapped sparse matrix, which can be represented as:

$$M = \begin{bmatrix} L_{\text{one - hot}}(X_{1,1}) & \cdots & L_{\text{one - hot}}(X_{1,N}) \\ \vdots & \ddots & \vdots \\ L_{\text{one - hot}}(X_{T,1}) & \cdots & L_{\text{one - hot}}(X_{T,N}) \end{bmatrix} \tag{4}$$

To obtain a compact representation of the codeword matrix, we create three empty matrices of different dimensions, $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$, map each frame in the $\mathbf{M}$ matrix to these three matrices to obtain the new embedding matrix, and stitch them together into a complete matrix $\mathbf{S}$ as the input to the following network.

C. GLOBAL-GUIDED CORRELATION MINING

The compressed speech stream is a typical time sequence signal. To extract the contextual feature of sequence signal, the Recurrent Neural Network (RNN) structure is considered as a prevailing solution. However, for the standard RNN, the vanishing gradient is an inevitable problem in practical applications. To solve this, the Long Short Term Memory network (LSTM) is commonly used. An LSTM structure consists of a set of recurrently connected blocks, also called memory blocks. Each block contains one or more recurrently connected memory cells and three multiplicative units, i.e., the input, output and forget gates. These units can provide

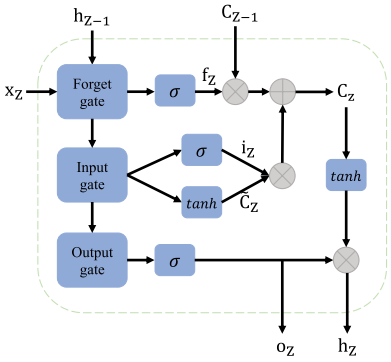

FIGURE 2. The structure of LSTM is composed of three gates, forget gate, input gate and output gate.

continuous analogs of write, read and reset operations for the cells. The net can only interact with the cells via the gates. The structure of an LSTM layer can be seen in Fig. 2 and the calculation of different units are as follows:

The calculation formulas for the forget gate $\mathbf{f}$ is as follows:

$$\mathbf{f}_Z = \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{Z-1}, \mathbf{x}_Z] + \mathbf{b}_f) \tag{5}$$

The input gate $\mathbf{i}$ of LSTM can be calculated by:

$$\mathbf{i}_Z = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{Z-1}, \mathbf{x}_Z] + b_i) \tag{6}$$

The alternative memory cell $\tilde{C}_Z$ will be generated as:

$$\tilde{\mathbf{C}}_Z = \tanh(\mathbf{W}_c \cdot [\mathbf{h}_{Z-1}, \mathbf{x}_Z] + b_c) \tag{7}$$

Then, the current memory cell C will be updated as:

$$\mathbf{C}_Z = \mathbf{f}_Z \cdot \mathbf{C}_{Z-1} + \mathbf{i}_Z \cdot \tilde{\mathbf{C}}_Z \tag{8}$$

Finally, the hidden layer $\mathbf{h}$ and the output layer o of the LSTM network can be obtained by:

$$\mathbf{f}_Z = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{Z-1}, \mathbf{x}_Z] + \mathbf{b}_o) \tag{9}$$

$$\mathbf{h}_Z = \mathbf{o}_Z \cdot \tanh(\mathbf{C}_Z) \tag{10}$$

The structure of Bi-LSTM is shown in Fig. 3. In the steganalysis process, after the codeword distributed embedding, the transposed matrix $\mathbf{X}_{embedding}$ is used as the input of the Bi-LSTM model. Each frame passes through the Bi-LSTM network to extract the context features of the compressed speech stream.

The attention mechanism can determine which part of the entire input needs more attention. Inspired by this, we introduce a multi-head attention mechanism to increase the weights of feature points that are helpful for steganalysis,, as shown in Fig. 4 and Fig. 5. The calculation process of the introduced multi-head attention is as follows. For the i-th head, the output of Bi-LSTM layer $\mathbf{L}$ passes through different linear layers to obtain $\mathbf{Q}^{(i)}$,$\mathbf{K}^{(i)}$ and $\mathbf{V}^{(i)}$ as follows:

$$\mathbf{Q}^{(i)} = \mathbf{L}^T \mathbf{W}_Q^{(i)} \tag{11}$$

$$\mathbf{K}^{(i)} = \mathbf{L}^T \mathbf{W}_K^{(i)} \tag{12}$$

$$\mathbf{V}^{(i)} = \mathbf{L}^T \mathbf{W}_V^{(i)} \tag{13}$$

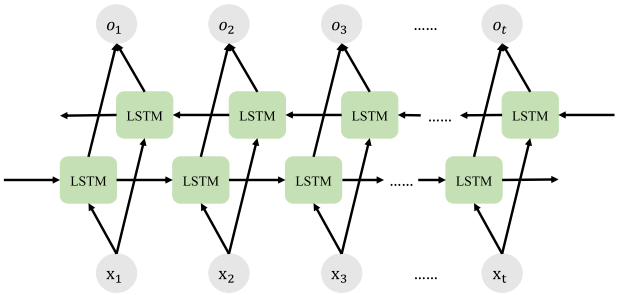

FIGURE 3. Composition structure of BI-LSTM: two Lstm models are spliced together before and after.

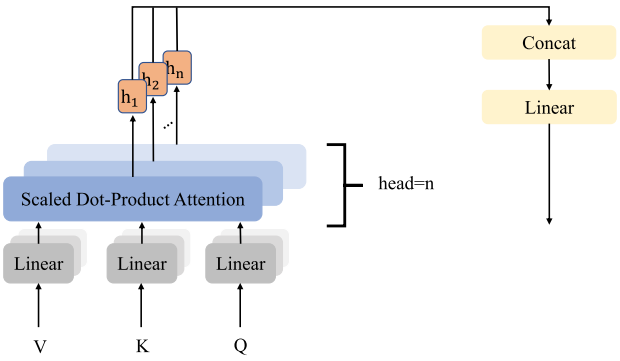

FIGURE 4. The multi-attentional process. By different linear layers, the input is converted to Q, K and V respectively. The output of different heads are concatenated through the linear layer.

where $\mathbf{W}_Q$,$\mathbf{W}_K$,$\mathbf{W}_V$ denote weight matrices. For the i-th heads $\mathbf{H}^{(i)}$,there is:

$$\mathbf{H}^{(i)} = \text{softmax}(\frac{\mathbf{Q}^{(i)}(\mathbf{K}^{(i)})'}{\sqrt{d_k}})\mathbf{V}^{(i)} \tag{14}$$

where $\frac{1}{\sqrt{d_k}}$ represents scaling factor. The results between the different heads will be concatenated together as the final output after linear transformation.

$$\mathbf{O}_{\text{MA}} = \text{concat}(\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \cdots, \mathbf{H}^{(N_h)})\mathbf{W}_{\text{ML}} \tag{15}$$

where N_h denotes the number of heads and $\mathbf{W}_{\text{ML}}$ represents the final linear transformation weight matrix. Besides, we introduce a skip connection composed of 1×1 convolution structure to alleviate the gradient vanishing. The global-guided feature $\mathbf{B}$ is obtained by concatenating $\mathbf{O}_{\mathbf{MA}}$ and the output of 1×1 convolution structure.

D. LOCAL-GUIDED CORRELATION MINING

In this section, we will explain the local-guided correlation mining module in detail. It consists of convolution blocks and Convolutional Block Attention Mechanism (CBAM). Firstly, we employ a depthwise convolution layer to convolve $\mathbf{B}$ in groups and extract the inra-word features in $\mathbf{B}$. Assuming that the convolution kernel has a weight of W_d and the dimension of input and output are i and a, respectively, the output can be

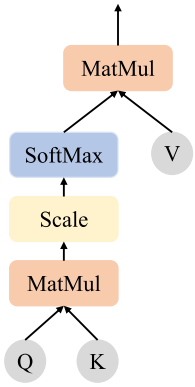

FIGURE 5. The structure of the scaling dot product attention model, the parameters W of Q, K and V for linear transformation are different, and the parameters of different heads are not shared.

represented as:

$$C_d(i,a) = \sum_{j=1}^{k} W_d(a,j) \cdot B^T([i+j-\frac{[k+1]}{2}], a) \tag{16}$$

where k is the convolution kernel size, d is the number of output channels. Furthermore, we use a 1×1 pointwise convolution kernel to extract deeper intra-word correlations, and fuse features. The convolution output O can be denoted as:

$$O(i,j) = \sum_{j=1}^{d} W_p(i,j) \cdot c_d(j,a) \tag{17}$$

To refine high-level features and suppress irrelevant noises, a convolutional block attention mechanism with residuals is introduced. It is able to focus on features on the spatial and channel perspectives, outperforming in classification tasks than methods that focus only on the channel perspective. The first component of CBAM is the channel attention module, and the architecture is shown in Fig. 6. The input feature map of channel attention is $O \in R^{C \times H \times W}$. To integrate the spatial information on each channel, two spatial features describing the channels are firstly generated by averaging pooling and maximum pooling. Then these two features are fed into a shared fully connected network, and the new spatial information obtained through the shared network is summed to obtain the channel attention feature map by activation function. The above process can be formulated as:

$$\begin{aligned} M_c(O) &= \sigma(\text{MLP}(\text{AvgPool}(O)) + \text{MLP}(\text{MaxPool}(O))) \\ &= \sigma(W_b(W_a(O^{\text{c}}_{\text{avg}})) + W_b(W_a(O^{\text{c}}_{\text{max}}))) \end{aligned} \tag{18}$$

Before entering the spatial attention module, we multiply the channel attention weight M_c with the feature map $\mathbf{O}$ to turn it into the weighted feature map $\mathbf{O}'$:

$$\mathbf{O}' = M_c(O) \otimes \mathbf{O} \tag{19}$$

The spatial attention module can explore the intrinsic relationship between the spatial dimensions of the feature map,

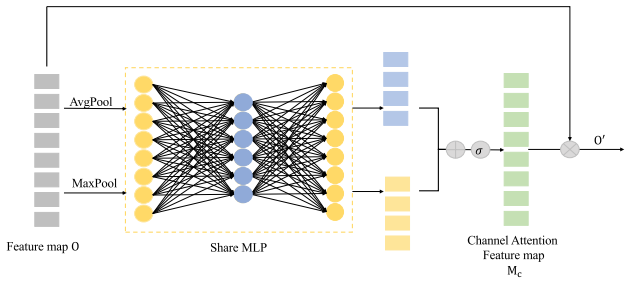

FIGURE 6. The architecture of channel attention module.

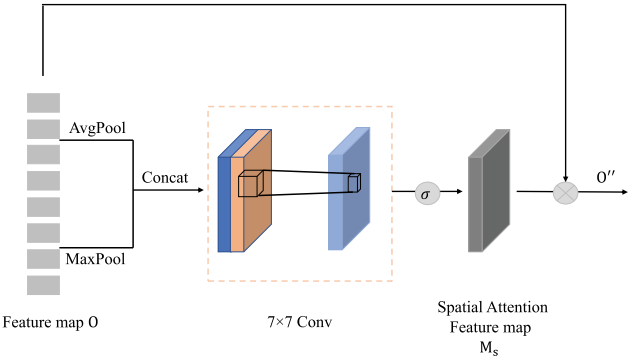

FIGURE 7. The architecture of spatial attention module.

i.e., noticing the effective features and ignoring the irrelevant noise. The architecture is shown in Fig. 7. In view of obtaining the information of the channel dimension in the feature map, we first conduct average pooling and maximum pooling and then concatenate the pooling results together. Next, a standard 7×7 convolution layer with sigmoid activation function can be used to obtain the spatial attention feature map. The above process can be formulated as:

$$\begin{aligned} M_s(O') &= \sigma(f_{7\times7}(\text{concat}(\text{AvgPool}(O'), \text{MaxPool}(O')))) \\ &= \sigma(f_{7\times7}(\text{concat}(O^s_{\text{avg}}, O^s_{\text{max}}))) \end{aligned} \tag{20}$$

Finally, we multiply the spatial attention weight with the feature map to turn it into the weighted feature map:

$$\mathbf{O}'' = M_s(O') \otimes \mathbf{O}' \tag{21}$$

To make the network more stable and to improve the de-fitting of the ideal state, a mechanism of residuals is used in the convolutional block attention. The ultimate feature space R of this model is obtained by adding the feature map to the input feature map of the convolution block attention, which can be expressed as:

$$\mathbf{R} = (\mathbf{O}'' + \mathbf{O}) \tag{22}$$

IV. EXPERIMENTS AND DISCUSSIONS

A. EXPERIMENTAL SETTINGS

We perform experiments on the speech dataset presented in [34]. The dataset contains 41 hours of Mandarin speeches and 72 hours of English speeches. The chosen low-bit-rate speech codec is G.723.1 (6.3 kbit/s). The encoded speeches constitute the dataset of cover (non-steganographic) speech. To comprehensively evaluate the performance of our proposed steganalysis method, five steganography methods, called ACL [12], CNV [16], HYF [20], SEC [17] and NPP [19] are used. Embedding hidden information in the cover speeches by these five steganography methods respectively, we get five different stego (steganographic) datasets. We compare our proposed method with four latest steganalysis methods, MFCC [26], CEBN [39], SFFN [40] and CBCA [41]. Both the training and testing stages were executed on GeForce GTX 3090 GPU with 24 G Graphics Memory. We use PyTorch to help implement the model and algorithm. In addition, in the process of training the neural network, we choose Adam as the optimizer with a learning rate of 0.001 and the cross entropy as the loss function. The maximal training epoch is 200, and the batch size in the training process is 128. The metric for evaluation is detection accuracy.

B. PERFORMANCE ANALYSIS UNDER DIFFERENT EMBEDDING RATES

The embedding rate is defined as the ratio of the number of embedded bits to the entire embedding capacity. Under ten different embedding rates (10%-100%) of 10 second speech length, five detection methods are compared. The results of the experiment are shown in Table 1 and Table 2. From the experimental results in the table, we can demonstrate that the detection accuracy of the five detection methods increases with the increase of the embedding rate. The lower the embedding rate is, the less the steganographic part of the speech is, and the harder it is to be detected. Usually, the quality of a model is evaluated according to its performance under the condition of low embedding rate.

In the Mandarin dataset, in the face of the three steganography methods CNV, HYF and SEC, the detection accuracies of MFCC, CEBN and SFFN are all lower than 60% when the embedding rate is 10%, while CBCA and our proposed method can achieve a satisfactory performance. Besides, the detection accuracy of our method is the highest, outperforming than CBCA. For ACL and NPP steganography methods, when the embedding rate is 10%, the detection accuracies of our proposed method are 89.88% and 93.26%, which is significantly better than other comparison steganalysis methods.

In the English dataset, the overall detection accuracy is better than that in the Mandarin dataset. However, in the face of HYF and SEC steganography methods, the detection accuracy of MFCC, CEBN and SFFN are still lower than 60% when the embedding rate is 10%. For CNV and NPP steganography methods, the proposed method is significantly better than other four detection methods, reaching an accuracy of 87.5% and 96.19%. For SEC steganography, which is the most difficult to detect, the detection accuracy of the proposed method can also achieve 92.35% when the embedding rate is 30%, and can exceed 99.3% for other steganography methods. This shows that our method can still maintain satisfactory detection performance in the face of some information hiding algorithms that are difficult to detect.

TABLE 1. The detection accuracies for the Mandarin dataset at different embedding rates.

Steganography Method	Detection Method	Embedding Rate (%)									
		10	20	30	40	50	60	70	80	90	100
ACL	MFCC	52.14	52.61	53.50	54.50	54.80	55.20	56.80	58.53	60.03	61.27
	CEBN	62.81	69.53	78.10	85.05	89.80	92.50	93.07	95.21	97.18	98.20
	SFFN	83.26	87.10	99.14	100	100	100	100	100	100	100
	CBCA	88.92	98.50	99.16	100	100	100	100	100	100	100
	Ours	**89.88**	**99.10**	**99.90**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
CNV	MFCC	54.21	56.24	57.16	58.62	62.08	63.80	65.07	66.54	67.51	70.89
	CEBN	57.51	65.49	72.68	82.18	86.98	90.32	91.99	94.13	98.21	98.75
	SFFN	59.28	70.02	77.74	85.04	89.10	92.14	93.97	96.99	98.07	98.73
	CBCA	82.76	95.00	98.21	99.52	99.90	100	100	100	100	100
	Ours	**83.46**	**95.66**	**98.88**	**99.82**	**99.98**	**100**	**100**	**100**	**100**	**100**
HYF	MFCC	55.07	56.13	57.52	58.80	60.02	60.09	61.94	62.76	65.00	68.50
	CEBN	56.61	62.45	71.82	77.50	81.07	84.13	85.04	87.91	90.03	91.82
	SFFN	56.60	62.32	66.06	72.50	77.56	79.12	82.50	87.00	90.90	92.17
	CBCA	85.00	95.04	97.99	99.02	99.60	99.99	100	100	100	100
	Ours	**86.00**	**95.65**	**98.65**	**99.56**	**99.97**	**100**	**100**	**100**	**100**	**100**
SEC	MFCC	50.02	51.70	52.51	54.89	56.09	56.98	57.40	58.82	61.23	68.04
	CEBN	54.20	57.39	62.11	63.98	70.00	72.80	78.00	80.94	82.12	85.41
	SFFN	55.21	60.02	65.22	68.79	75.00	77.91	80.94	85.00	85.93	90.00
	CBCA	67.27	80.40	89.12	94.50	97.18	97.50	98.00	99.04	99.48	100
	Ours	**67.38**	**81.02**	**89.73**	**94.53**	**97.48**	**99.01**	**99.54**	**99.85**	**99.97**	**100**
NPP	MFCC	55.61	56.20	56.34	57.04	57.50	58.07	58.98	59.90	60.23	62.26
	CEBN	61.20	69.50	73.54	76.32	82.11	85.00	89.50	90.88	92.48	95.00
	SFFN	87.50	95.12	96.81	99.73	99.79	99.90	99.90	100	100	100
	CBCA	93.06	98.00	99.10	**99.90**	**99.90**	99.90	99.90	100	100	100
	Ours	**93.26**	**98.39**	**99.48**	99.74	99.84	**99.95**	**99.97**	**100**	**100**	**100**

C. PERFORMANCE ANALYSIS UNDER DIFFERENT SPEECH LENGTHS

For steganalysis, speech length is also an important factor affecting the detection accuracy. Experiments analyze the detection performance of the model under different speech lengths. Under the condition of 100% embedding rate and speech length from 1s to 10s. we evaluate the performance of five steganalysis methods. The experimental results are shown in Table 3 and Table 4.

It can be seen from the figure that in the Mandarin dataset, for ACL, the method in this paper achieves 100% detection accuracy when the speech length is 1s, and for CNV and NPP, when the speech length is 1s, the accuracy reaches 99.36% and 99.74%. significantly better than other comparison methods. In the English dataset, for ACL and NPP, our method slightly outperforms SFFN and CBCA, and significantly outperforms MFCC and CEBN. For SEC, when the speech length is 1s, the detection accuracy of MFCC is only over 50%, while that of SFFN is over 75%. This method further improves the detection accuracy to 95.51%. For CNV and HYF, the detection accuracy of our method can exceed 99.4% even when the speech length is 1 s. The results show that the detection method in this paper still has good detection performance for low-length speech samples.

D. GENERAL PERFORMANCE ANALYSIS OF THE MODEL UNDER MIXED DATA SET

In practical application, facing a compressed speech stream, we will not have prior knowledge to know which steganographic method is adopted in advance. Once our detection model is trained, it should be able to detect various steganographic algorithms simultaneously. To this end, we construct a MIX dataset, including 1/5 of each of the five steganographic datasets, ACL, CNV, HYF, SEC and NPP, and then mix them. This dataset is used to evaluate the detection performance on whether the unknown compressed speech stream contains hidden information.

First, the performance of the model under different embedding rates (10%-100%) when the speech length is 10 seconds is tested. The higher the accuracy, the better the performance of the model in the face of unknown steganography method. The experimental results are shown in Fig. 9. Experimental results show that the performance of the proposed model is better than that of the four models compared. In the English

TABLE 2. The detection accuracies for the English dataset at different embedding rates.

Steganography Method	Detection Method	Embedding Rate (%)									
		10	20	30	40	50	60	70	80	90	100
ACL	MFCC	54.08	56.10	58.30	60.91	62.80	64.57	65.51	66.89	68.93	69.72
	CEBN	63.92	70.51	78.89	87.30	91.58	94.69	95.82	96.71	97.90	99.62
	SFFN	90.48	97.78	99.81	100	100	100	100	100	100	100
	CBCA	90.50	98.58	99.91	100	100	100	100	100	100	100
	Ours	**90.50**	**98.86**	**99.93**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
CNV	MFCC	55.33	57.28	58.59	60.53	63.30	64.50	66.42	67.39	68.84	72.10
	CEBN	58.81	67.62	76.10	84.73	87.68	91.84	93.67	94.50	95.92	96.09
	SFFN	64.80	73.29	82.18	87.81	93.82	95.58	96.70	97.10	97.74	98.43
	CBCA	85.62	96.09	98.10	99.80	100	100	100	100	100	100
	Ours	**87.50**	**97.05**	**99.46**	**99.95**	**100**	**100**	**100**	**100**	**100**	**100**
HYF	MFCC	54.30	57.71	58.43	60.72	62.40	62.55	64.08	65.24	66.10	68.45
	CEBN	57.18	63.71	74.69	78.34	83.79	85.09	85.80	87.88	92.21	94.10
	SFFN	56.53	63.81	66.89	74.88	78.10	77.93	86.21	90.29	93.60	95.26
	CBCA	89.41	96.90	98.42	98.90	99.08	100	100	100	100	100
	Ours	**90.36**	**97.68**	**99.30**	**99.89**	**99.93**	**100**	**100**	**100**	**100**	**100**
SEC	MFCC	50.28	54.34	55.10	56.82	57.93	58.33	60.78	61.24	63.37	65.13
	CEBN	54.92	60.39	64.71	66.08	72.30	75.06	80.26	82.43	84.70	86.66
	SFFN	56.19	64.34	64.85	75.30	76.14	83.72	86.20	86.70	87.78	92.15
	CBCA	72.13	84.56	91.89	95.81	96.67	96.70	98.32	98.81	99.40	100
	Ours	**72.58**	**84.56**	**92.35**	**96.30**	**98.46**	**99.34**	**99.75**	**99.92**	**99.99**	**100**
NPP	MFCC	57.31	59.19	59.90	61.21	63.30	64.07	64.92	64.97	66.19	69.16
	CEBN	63.68	72.77	75.08	78.30	83.87	86.67	89.92	93.60	94.75	96.81
	SFFN	89.20	92.78	98.15	98.69	99.11	99.74	100	100	100	100
	CBCA	95.11	94.59	98.69	99.60	99.71	100	100	100	100	100
	Ours	**96.19**	**99.04**	**99.75**	**99.94**	**99.98**	**100**	**100**	**100**	**100**	**100**

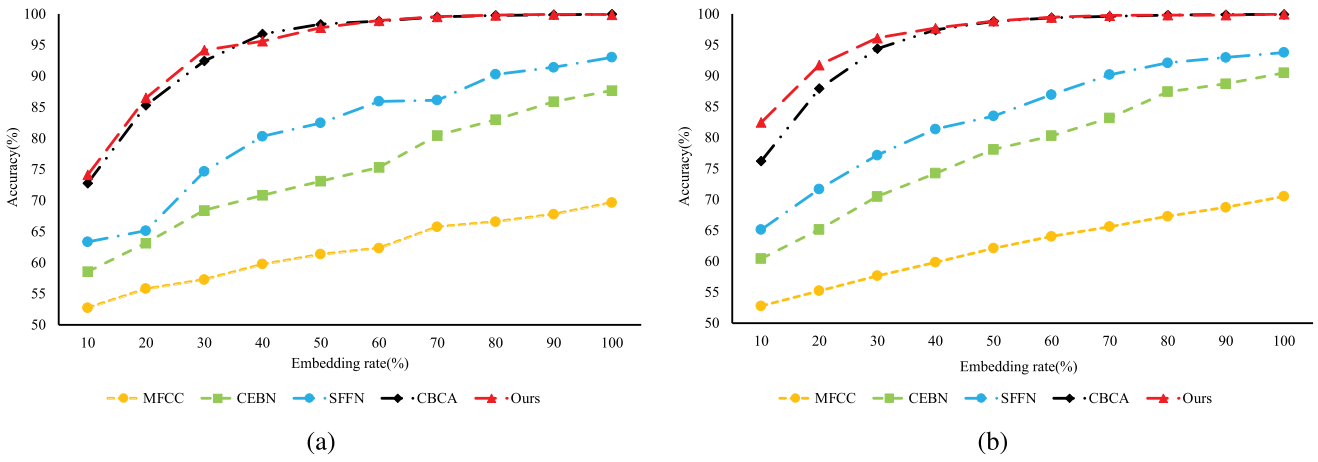

FIGURE 8. The detection accuracies for the Mix dateset at different embedding. (a) Mandarin Mix dataset. (b) English Mix dataset.

mixed data set, the model is obviously superior to MFCC, CEBN and SFFN models. In the detection tasks with low embedding rates of 10% and 20%, the model is 4%-6% higher than the previous excellent CBCA model, and the accuracy is up to 82.46% and 91.76%. When the embedding rate is 30%, the detection accuracy of the proposed model is higher than that of MFCC, CEBN and SFFN models when the embedding rate is 100%.

Next, the performance of the model at different voice durations (1s-10s) with 100% embedding rate was tested. The

TABLE 3. The detection accuracies for the Mandarin dataset at different speech lengths.

Steganography Method	Detection Method	Speech Lengths (s)									
		1	2	3	4	5	6	7	8	9	10
ACL	MFCC	51.89	52.00	53.21	54.04	55.01	56.02	58.49	59.47	60.03	61.04
	CEBN	87.39	89.54	90.01	91.87	93.91	94.06	95.09	96.13	97.22	98.01
	SFFN	100	100	100	100	100	100	100	100	100	100
	CBCA	100	100	100	100	100	100	100	100	100	100
	Ours	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
CNV	MFCC	55.23	56.01	57.11	58.47	60.52	62.45	64.04	67.39	69.01	70.80
	CEBN	79.71	81.87	83.84	86.03	89.12	91.89	94.91	94.99	95.63	96.59
	SFFN	85.02	87.81	91.02	92.90	95.04	96.01	96.49	97.57	97.92	99.03
	CBCA	98.12	99.04	99.09	99.17	99.57	99.89	100	100	100	100
	Ours	**99.36**	**99.93**	**99.98**	**99.99**	**100**	**100**	**100**	**100**	**100**	**100**
HYF	MFCC	55.03	56.01	56.46	58.59	60.09	62.43	64.92	66.04	67.49	68.68
	CEBN	77.37	79.01	82.28	84.02	85.78	87.01	87.65	88.31	90.48	91.68
	SFFN	73.54	75.01	82.03	83.78	86.52	86.47	87.39	90.02	91.25	92.16
	CBCA	98.49	98.89	98.96	99.01	99.12	99.52	100	100	100	100
	Ours	**99.19**	**99.78**	**99.96**	**99.98**	**99.99**	**100**	**100**	**100**	**100**	**100**
SEC	MFCC	50.82	51.21	52.48	53.87	56.02	58.01	59.78	61.49	63.04	63.47
	CEBN	69.49	72.21	73.04	74.23	76.38	78.01	80.48	83.04	83.51	85.54
	SFFN	72.51	75.58	80.02	80.54	83.01	84.28	86.19	87.72	89.01	90.04
	CBCA	93.52	96.31	98.78	99.02	99.91	99.95	99.96	100	100	100
	Ours	**93.59**	**97.52**	**98.93**	**99.54**	**99.91**	**99.95**	**99.97**	**100**	**100**	**100**
NPP	MFCC	52.01	53.37	53.79	55.02	56.14	57.38	58.59	60.01	61.27	62.12
	CEBN	82.31	84.12	85.24	86.47	88.86	90.04	91.51	93.35	94.02	95.01
	SFFN	98.09	99.27	99.95	99.97	100	100	100	100	100	100
	CBCA	99.01	99.91	99.95	99.97	100	100	100	100	100	100
	Ours	**99.74**	**99.91**	**99.95**	**99.97**	**100**	**100**	**100**	**100**	**100**	**100**

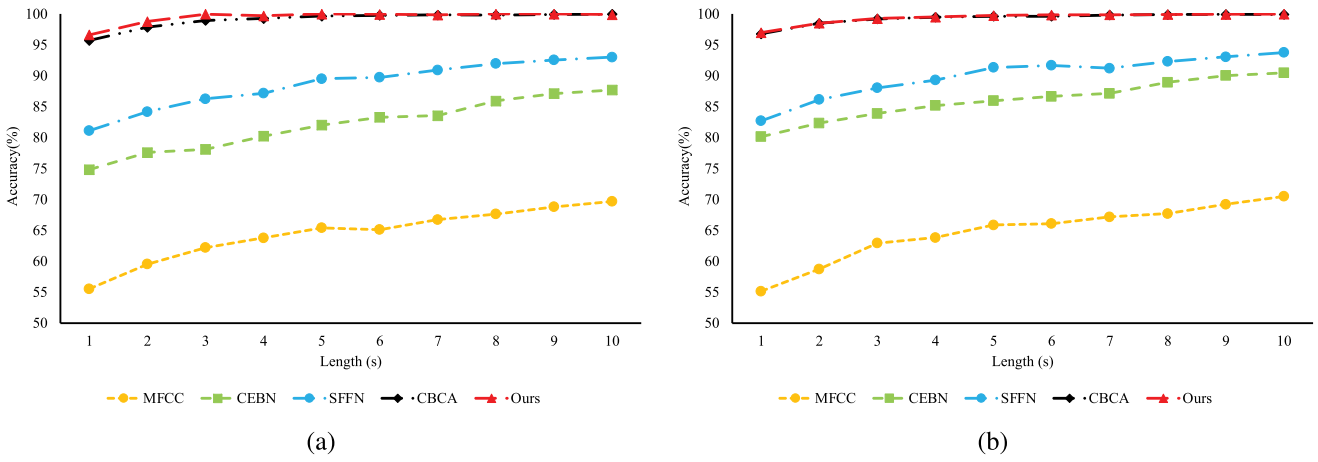

FIGURE 9. The detection accuracies for the Mix dateset at different speech length. (a) Mandarin Mix dataset. (b) English Mix dataset.

experimental results are shown in Fig. 10. According to the experimental results, in the face of different speech duration detection tasks, the proposed model still shows better performance than the previous four models. In the Mandarin mixed data set, when the speech duration is 5s, the detection accuracy of this model has reached 100%. In the speech detection task with low duration (1s), the detection accuracy of the proposed model can reach above 96%, showing good performance.

E. ANALYSIS OF THE INFLUENCE OF DIFFERENT MODEL STRUCTURES ON PERFORMANCE

In this section, the effects of different submodel structures on model performance are discussed. In order to reflect the

TABLE 4. The detection accuracies for the English dataset at different speech lengths.

Steganography Method	Detection Method	Speech Lengths (s)									
		1	2	3	4	5	6	7	8	9	10
ACL	MFCC	55.18	57.61	58.29	58.71	61.80	62.53	64.43	66.68	67.71	70.08
	CEBN	87.08	90.23	92.29	94.62	96.57	97.63	98.20	98.61	99.10	99.23
	SFFN	98.71	100	100	100	100	100	100	100	100	100
	CBCA	99.49	100	100	100	100	100	100	100	100	100
	Ours	**99.99**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
CNV	MFCC	56.22	59.41	62.20	63.69	64.32	64.87	66.81	68.20	68.79	72.31
	CEBN	83.57	86.21	87.68	90.16	91.71	94.51	95.58	96.82	97.40	97.91
	SFFN	87.32	91.56	94.23	95.71	95.88	96.30	97.52	98.23	98.81	98.89
	CBCA	98.06	98.09	98.13	99.52	100	100	100	100	100	100
	Ours	**99.56**	**99.97**	**99.99**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
HYF	MFCC	56.10	53.26	54.12	62.59	63.72	64.57	64.89	66.51	67.90	68.82
	CEBN	80.87	83.26	83.51	84.78	85.13	87.32	88.90	90.12	92.89	94.24
	SFFN	75.09	81.92	85.81	86.89	89.63	89.80	90.74	92.59	94.82	94.91
	CBCA	97.81	98.64	98.71	99.13	99.51	100	100	100	100	100
	Ours	**99.41**	**99.86**	**99.95**	**99.98**	**99.98**	**100**	**100**	**100**	**100**	**100**
SEC	MFCC	54.90	55.63	57.18	57.81	58.63	60.91	62.67	64.20	64.89	65.12
	CEBN	69.61	71.12	72.46	73.42	74.90	77.58	80.71	84.06	85.91	87.64
	SFFN	75.12	80.20	84.47	84.61	86.87	86.19	86.92	87.54	89.90	92.28
	CBCA	94.87	96.23	97.81	98.37	98.80	99.12	99.40	99.58	100	100
	Ours	**95.51**	**98.26**	**99.25**	**99.67**	**99.79**	**99.91**	**99.95**	**99.99**	**100**	**100**
NPP	MFCC	54.88	55.12	57.59	59.91	63.43	64.89	65.10	65.83	67.91	69.93
	CEBN	85.10	87.27	89.92	92.43	92.61	93.68	98.10	95.14	96.21	96.89
	SFFN	97.83	98.80	99.07	99.31	99.79	100	100	100	100	100
	CBCA	97.91	99.14	99.30	99.51	99.82	100	100	100	100	100
	Ours	**99.91**	**99.95**	**99.98**	**99.99**	**99.99**	**100**	**100**	**100**	**100**	**100**

TABLE 5. The influence of different model structures on performance.

Model substructure	Mandarin	English
CDE	65.54	67.61
CDE+GCM	71.66	80.59
CDE+GCM+LCM	74.14	82.46

TABLE 6. Comparsion of neural network models on time complexity.

Method	Test time ratio(‰)
SFFN	0.61
CBCA	1.33
Our	1.31

influence of different model substructures on model performance, a mixed data set with speech length of 10s and embedding rate of 10% was used for the experiment. The experimental results are shown in Table 5. It can be seen from the table that in unknown steganography analysis with low embedding rate, it is not enough to obtain the code word information with compact representation only by Codeword Distributed Embedding (CDE) structure model. The performance of the model was improved by adding GCM structure after CDE structure. This is because Bi-LSTM model and multi-head attention mechanism are introduced to pay attention to the information of before and after speech frames and different subspaces. LCM structure is added after the above model structure, and the convolutional neural network and CBCA model in LCM structure are used to supplement the spatial information in different frames of speech, so that the model performance is improved again.

F. TIME COMPLEXITY ANALYSIS

Network compressed voice detection task needs to check whether the voice is steganography in real time. Therefore, the model detection time of each speech sample should be as short as possible. In this section, the time complexity of the model will be analyzed. The speech of 1s to 10s is selected for detection, and the average detection time is used as the detection time of the model. In this part, a platform equipped with GTX 3090 GPU is used to conduct the test time experiment. Since MFCC and CEBN methods are not neural

network models and do not require GPU testing, this part is only compared with SFFN and CBCA. The experimental results are shown in Table 6.

As can be seen from the table, the test time is between CBCA and SFFN. When the tested speech length is 1s, the test time of the model in this paper is 1.31%of the total time, that is, it takes 1.31 ms to detect a 1s speech sample. It can be seen that the proposed model in this paper can achieve real-time performance in the network compressed speech detection task.

V. CONCLUSION

Aiming at the lack of an efficient general ABS-LPC steganalysis method for low rate compressed speech. A steganalysis method based on global association mining and local association mining is proposed in this paper. In practice, no matter which steganography method is used, the correlations in speech codeword stream will be changed. Therefore, our proposed method focuses on extracting relevant changes before and after steganography from global and local perspectives. Experimental results show that the proposed method has better performance than the existing general steganalysis methods. In the future, we hope to achieve a general detection method for unknown speech steganography methods with better performance under very low embedding rate and short duration.

ACKNOWLEDGMENT

(Jiawei Wang and Jie Yang contributed equally to this work.)

REFERENCES

[1] D. Xu and B. Guan, "An improved commutative encryption and data hiding scheme for HEVC video," *Multimedia Tools Appl.*, vol. 81, no. 13, pp. 18105–18127, May 2022.
[2] M. Kumar, J. Aggarwal, A. Rani, T. Stephan, A. Shankar, and S. Mirjalili, "Secure video communication using firefly optimization and visual cryptography," *Artif. Intell. Rev.*, vol. 55, no. 4, pp. 2997–3017, Apr. 2022.
[3] K. Yu, L. Chen, Y. Wang, and T. Lu, "A channel coding information hiding algorithm for images based on uniform cyclic shift," *Multimedia Tools Appl.*, vol. 81, no. 8, pp. 11279–11300, Mar. 2022.
[4] A. H. M. Almawgani, A. R. H. Alhawari, A. T. Hindi, W. H. Al-Arashi, and A. Y. Al-Ashwal, "Hybrid image steganography method using Lempel Ziv Welch and genetic algorithms for hiding confidential data," *Multidimensional Syst. Signal Process.*, vol. 33, no. 2, pp. 561–578, Jun. 2022.
[5] B. Ma, J.-C. Hou, C.-P. Wang, X.-M. Wu, and Y.-Q. Shi, "A reversible data hiding algorithm for audio files based on code division multiplexing," *Multimedia Tools Appl.*, vol. 80, no. 12, pp. 17569–17581, Feb. 2021.
[6] X. Sun, K. Wang, and S. Li, "Audio steganography with less modification to the optimal matching CNV-QIM path with the minimal Hamming distance expected value to a secret," *Multimedia Syst.*, vol. 27, no. 3, pp. 341–352, Jun. 2021.
[7] C. O. Mawalim and M. Unoki, "Speech watermarking method using McAdams coefficient based on random forest learning," *Entropy*, vol. 23, no. 10, p. 1246, Sep. 2021.
[8] J. Wen, H. Zeng, Y. Wang, S. Liu, and Y. Xue, "An SVD-based adaptive robust speech steganography using MDCT coefficient," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2517–2536, Jan. 2021.
[9] D.-C. Wu and Y.-T. Hsu, "Authentication of LINE chat history files by information hiding," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 1, pp. 1–23, Jan. 2022.
[10] J. Qin, Z. Zhou, Y. Tan, X. Xiang, and Z. He, "A big data text coverless information hiding based on topic distribution and TF-IDF," *Int. J. Digit. Crime Forensics*, vol. 13, no. 4, pp. 40–56, Jul. 2021.
[11] Z. Wu, J. Guo, C. Zhang, and C. Li, "Steganography and steganalysis in voice over IP: A review," *Sensors*, vol. 21, no. 4, p. 1032, Feb. 2021.
[12] Y. Huang, C. Liu, S. Tang, and S. Bai, "Steganography integration into a low-bit rate speech codec," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1865–1875, Dec. 2012.
[13] A. Janicki, "Pitch-based steganography for speex voice codec," *Secur. Commun. Netw.*, vol. 9, pp. 2923–2933, Feb. 2016.
[14] Y. Ren, D. Liu, J. Yang, and L. Wang, "An AMR adaptive steganographic scheme based on the pitch delay of unvoiced speech," *Multimedia Tools Appl.*, vol. 78, no. 7, pp. 8091–8111, Apr. 2019.
[15] X. Liu, H. Tian, Y. Huang, and J. Lu, "A novel steganographic method for algebraic-code-excited-linear-prediction speech streams based on fractional pitch delay search," *Multimedia Tools Appl.*, vol. 78, no. 7, pp. 8461–8847, Apr. 2019.
[16] B. Xiao, Y. Huang, and S. Tang, "An approach to information hiding in low bit-rate speech stream," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2008, pp. 1–5.
[17] H. Tian, J. Liu, and S. Li, "Improving security of quantization-index-modulation steganography in low bit-rate speech streams," *Multimedia Syst.*, vol. 20, no. 2, pp. 143–154, 2014.
[18] P. Liu, S. Li, and H. Wang, "Steganography integrated into linear predictive coding for low bit-rate speech codec," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2837–2859, Jan. 2017.
[19] P. Liu, S. Li, and H. Wang, "Steganography in vector quantization process of linear predictive coding for low-bit-rate speech codec," *Multimedia Syst.*, vol. 23, no. 4, pp. 485–497, Jul. 2017.
[20] Y. F. Huang, S. Tang, and J. Yuan, "Steganography in inactive frames of VoIP streams encoded by source codec," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 296–306, Jun. 2011.
[21] Z. Wu, H. Cao, and D. Li, "An approach of steganography in G.729 bitstream based on matrix coding and interleaving," *Chin. J. Electron.*, vol. 24, no. 1, pp. 157–165, Jan. 2015.
[22] Z. Yang, X. Peng, Y. Huang, and C. Cheng, "A novel method of speech information hiding basedon 3D-magic matrix," *J. Internet Technol.*, vol. 20, no. 4, pp. 1167–1175, Sep. 2019.
[23] L. Zhang, X. Hu, W. Rasheed, T. Huang, and C. Zhao, "An enhanced steganographic code and its application in voice-over-IP steganography," *IEEE Access*, vol. 7, pp. 97181–97195, 2019.
[24] F. Li, B. Li, Y. Huang, Y. Feng, L. Peng, and N. Zhou, "Research on covert communication channel based on modulation of common compressed speech codec," *Neural Comput. Appl.*, vol. 809, pp. 1–14, Apr. 2020.
[25] Z. Wu, R. Li, and C. Li, "Adaptive speech information hiding method based on K-means," *IEEE Access*, vol. 8, pp. 23308–23316, 2020.
[26] Q. Liu, A. H. Sung, and M. Qiao, "Temporal derivative-based spectrum and mel-cepstrum audio steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 3, pp. 359–368, Sep. 2009.
[27] S. Li, Y. Jia, J. Fu, and Q. Dai, "Pitch modulation information hiding detection based on codebook association network," *Chin. J. Comput.*, vol. 37, no. 10, pp. 2107–2117, Oct. 2014.
[28] Y. Jia, S. Li, Y. Jiang, Q. Dai, and H. Deng, "Detection of G.729A pitch modulation information hiding based on symbiotic characteristics," *Acta Electronica Sinica*, vol. 43, no. 8, pp. 1513–1517, Aug. 2015.
[29] Y. Ren, J. Yang, J. Wang, and L. Wang, "AMR steganalysis based on second-order difference of pitch delay," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 6, pp. 1345–1357, Jun. 2017.
[30] Y. Wu, H. Zhang, Y. Sun, and M. Chen, "Steganalysis of AMR based on statistical features of pitch delay," *Int. J. Digit. Crime Forensics*, vol. 11, no. 4, pp. 66–81, Oct. 2019.
[31] S.-B. Li, H.-Z. Tao, and Y.-F. Huang, "Detection of quantization index modulation steganography in G.723.1 bit stream based on quantization index sequence analysis," *J. Zhejiang Univ. Sci. C*, vol. 13, no. 8, pp. 624–634, Aug. 2012.
[32] S. Li, Y. Jia, and C.-C. J. Kuo, "Steganalysis of QIM steganography in low-bit-rate speech signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 1011–1022, May 2017.
[33] J. Yang and S. Li, "Steganalysis of joint codeword quantization index modulation steganography based on codeword Bayesian network," *Neurocomputing*, vol. 313, no. 3, pp. 316–323, Nov. 2018.
[34] Z. Lin, Y. Huang, and J. Wang, "RNN-SM: Fast steganalysis of VoIP streams using recurrent neural network," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 7, pp. 1854–1868, Jul. 2018.
[35] H. Yang, Z. Yang, Y. Bao, S. Liu, and Y. Huang, "Fast steganalysis method for VoIP streams," *IEEE Signal Process. Lett.*, vol. 27, pp. 286–290, 2020.

[36] Z. Yang, H. Yang, C.-C. Chang, Y. Huang, and C.-C. Chang, "Real-time steganalysis for streaming media based on multi-channel convolutional sliding windows," *Knowl.-Based Syst.*, vol. 237, Feb. 2022, Art. no. 107561.
[37] H. Tian, Y. Wu, Y. Cai, Y. Huang, J. Liu, T. Wang, Y. Chen, and J. Lu, "Distributed steganalysis of compressed speech," *Soft Comput.*, vol. 21, no. 3, pp. 795–804, Feb. 2017.
[38] H. Tian, J. Liu, C.-C. Chang, Y. Huang, and Y. Cai, "Detecting steganography in inactive voice-over-IP frames based on statistic characteristics of fundamental frequency," *IEEE Access*, vol. 8, pp. 6117–6129, 2020.
[39] J. Yang, P. Liu, and S. Li, "A common method for detecting multiple steganographies in low-bit-rate compressed speech based on Bayesian inference," *IEEE Access*, vol. 7, pp. 128313–128324, 2019.
[40] Y. Hu, Y. Huang, Z. Yang, and Y. Huang, "Detection of heterogeneous parallel steganography for low bit-rate VoIP speech streams," *Neurocomputing*, vol. 419, pp. 70–79, Jan. 2021.
[41] S. Li, J. Wang, P. Liu, M. Wei, and Q. Yan, "Detection of multiple steganography methods in compressed speech based on code element embedding, Bi-LSTM and CNN with attention mechanisms," *IEEE/ACM Trans. Audio, Speech, Language Process*, vol. 29, pp. 1556–1569, 2021.

JIAWEI WANG is currently pursuing the B.S. degree with the Jiyang College, Zhejiang A&F University, Zhuji, China.

His current research interests include deep learning and information hiding.

JIE YANG received the B.S. degree in electronic science and technology from Beijing Normal University, Beijing, China, in 2013, and the Ph.D. degree from the University of Chinese Academy of Sciences, in 2018, Beijing.

He has been an Associate Professor with the Jiyang College, Zhejiang A&F University, Zhuji, China, since 2021. His current research interests include machine learning, multimedia signal processing, and data hiding.

FEIPENG GAO received the B.S. degree in computer science and technology from the Jiyang College, Zhejiang A&F University, Zhuji, China, in 2022. He is currently pursuing the M.S. degree with the Zhejiang A&F University, Linan, China.

His research interests include machine learning and information hiding.

PENG XU received the Ph.D. degree in optical engineering from Zhejiang University, Hangzhou, China, in 2017.

He was a Visiting Researcher at the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada, from 2019 to 2021. He is currently a Lecturer with the Jiyang College of Zhejiang A&F University, Zhuji, China. His current research interests include multimedia applications and machine learning.