VoIP Steganalysis Using Shallow Multiscale Convolution and Transformer

Jinghui Peng[1(✉)], Yi Liao[1], and Shanyu Tang[2]

[1] School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou 510665, Guangdong Province, China
jinghuipeng@gpnu.edu.cn

[2] Cybersecurity and Criminology Centre, University of West London, St. Mary's Road, W5 5RF London, UK

Abstract. Steganography is an effective method for transmitting secret information, but it can also be used for illegal activities such as terrorism, organized crime and data theft, etc. To solve the problem of steganography being used for malicious purposes, steganalysis technology has been developed. Steganalysis aims to detect whether the data has been steganography and identify whether it contains secret information, which is a kind of reverse process of steganography. VoIP data stream usually has high redundancy, which makes it an ideal carrier for steganography. In this paper, a Steganalysis Transformer (SAT) VoIP voice steganalysis method based on Transformer neural network is proposed with VoIP voice as the research object. The method first encodes the relative position of the features extracted from VoIP voice signals, combines the multi-scale convolution method to improve the local feature extraction to obtain more detailed feature information, transforms the high-dimensional sparse matrix into the low-dimensional dense features by mapping, and then realizes the steganalysis analysis through the feature extraction by the improved Transformer; the proposed SAT method is able to obtain the global features from the shallow layer and learn the high quality intermediate features. Experiments show that the SAT method proposed in this paper has superior performance, and the accuracy of VoIP steganalysis reaches 96.41%.

Keywords: Steganography · Steganalysis · Neural Network · Attention Mechanism · Transformer

1 Introduction

The way to protect the privacy and security of data is generally through the use of cryptographic techniques, by encrypting the content of the communication so that only the person who has the key can decrypt and view the communication. Cryptographic techniques have been widely used to encrypt plaintext data, transmit the ciphertext over the Internet, and decrypt the ciphertext at the receiving end to extract the plaintext [1]. However, since the ciphertext does not make much sense when interpreted, a hacker or intruder can easily detect that the information sent over the channel was encrypted

J. Cai et al. (Eds.): SPNCE 2023, LNICST 525, pp. 326–351, 2025.
https://doi.org/10.1007/978-3-031-73699-5_23

rather than plaintext, which naturally increases the curiosity of malicious hackers or intruders to cryptanalyze the ciphertext for attacks and makes them more likely to be targeted. Information hiding technology is to secretly embed the hidden information to be transmitted into the normal information by some specific methods, so that the secret information can be transmitted by the transmission of normal information at the same time without being detected, and it is an important research direction in the field of information security [2].

Steganography is a branch of information hiding technology, which aims to carry out secret communication without causing suspicion or detection by a third party, so as to ensure the security and privacy of communication [3]. Steganography can hide secret information in another information medium, which is generally called carrier, and the carrier can be images, voices, texts, videos, etc. There are two basic research methods for VoIP voice steganography: the steganography with the payload as the carrier in the voice stream of VoIP real-time transmission, and the steganography based on the network protocol. The payload-based VoIP voice steganography method exploits the redundancy of the voice stream itself and embeds secret information in the redundant bits of the carrier voice stream, which has better concealment and larger hidden capacity [4]. The payload-based VoIP voice steganography usually uses two basic steganographic algorithms, Least Significant Bit (LSB) steganography and Quantized Index Modulation (QIM) steganography. Steganography is difficult to detect and manipulate, and its correct use can ensure information security, but unfortunately, at present, steganography is mostly used in illegal activities, mainly involving terrorism, organized crime and data theft. For example, when spreading malicious software, hackers hide malicious software in ordinary files to steal users' personal information, control computers, and so on. The risk and impact of malicious use of steganography must be paid attention to, and there are huge security risks.

In order to solve the security risks caused by the malicious use of steganography, prevent the malicious use of steganography and protect information security, the most effective method is steganalysis, which can effectively detect steganography. It appears at the same time as steganography, which is antagonistic to each other. It was first used to evaluate the strength of steganography. The purpose of steganalysis is to analyze whether the carrier is steganographic, which belongs to the problem of pattern recognition or machine learning classification [5]. When the steganography algorithm embeds the secret data, the original data must be modified, and some statistical characteristics of the original data must also be modified. Therefore, we can distinguish whether it is steganographic by learning and analyzing the difference between the characteristics of the original data and the modified data. Steganalysis has a common pattern. Firstly, it extracts some feature information from the carrier of sample data, and then analyzes and classifies this feature information to detect whether the carrier is steganographic [6]. In addition, steganalysis has a very wide range of applications, such as network security, intelligence monitoring, legal forensics, etc. In the field of network security, steganalysis can be used to detect and prevent network attacks, prevent the spread of malicious software and data leakage; it is used to decrypt the secret information of terrorists and provide support for the fight against terrorism; it is used to crack the information of criminals and provide evidence for case investigation and evidence collection. Steganalysis is of great significance in

protecting network and data security, promoting the development of steganography, and contributing to social security and stability.

2 Relate Work

The research stage of steganalysis can be roughly divided into two stages: In the first stage, Rich Model (RM) [7] is usually used to manually extract the features of the best results; in the early stage, statistical strategies are usually used to manually classify the extracted statistical features; After machine learning is mature, binary classifiers are used for classification (the common ones are ensemble classifier (EC) [8], support vector machine (SVM) [9] and perceptron; In the second stage, with the rapid development of deep learning [10], steganalysis also began to apply the method of deep learning, which can automatically extract relevant features and learn to detect steganography. At present, the steganalysis method of deep learning has shown strong performance in accuracy, robustness and detection efficiency, and gradually replace the traditional method, which is the mainstream research trend.

According to the format, audio steganalysis can be divided into steganalysis technology for compressed format (such as MP3 and AAC) and steganalysis technology for uncompressed format (such as WAV). According to the correlation characteristics of split vector quantization (VQ) codewords with linear predictive coding filter coefficients changed after QIM steganography, a model called quantized codeword correlation network (QCCN) is constructed, which is based on the split VQ codewords from adjacent audio frames. A high-performance detector is constructed by using support vector machine (SVM) classifier. It can effectively detect steganography used in G.723.1 and G.729 low bit rate audio codec.

In terms of compressed audio, Jin et al. [11] proposed a steganalysis technique to detect mp3stego steganography in 2016. Mp3stego changes the quantized modified discrete cosine transform coefficient (QMDCT) during the compression process, which affects the correlation between adjacent QMDCT's of the audio coverage. Therefore, Markov features can be extracted from the carrier audio and steganography audio to describe the correlation between QMDCT, and then these features are crossed through the preprocessing steps to select the best features to train the support vector machine classifier. The experimental results show that this method has high detection accuracy in the case of low embedding rate. In 2020, Wang et al. [12] proposed another steganalysis technique for MP3, which extracts the steganalysis features of MP3 by calculating the QMDCT coefficient matrix of MP3, and uses rich high-pass filtering technique to improve the sensitivity of the technique to noise signals. The author found that each replacement of a QMDCT coefficient changes a Huffman codeword. For this reason, they proposed a correlation measurement module, which is used to measure the correlation between point and 2×2 and 4×4 detect any possible changes in the QMDCT coefficient matrix on blocks. To reduce the dimension of the features and to select the optimal feature, an empirical threshold is applied. For the classification task, the ensemble classifier is trained.

Li [13] proposed a steganalysis method in low bit rate coded voice stream to detect the steganography of VoIP in compressed domain. According to the correlation characteristics of split vector quantization (VQ) codewords with linear predictive coding

filter coefficients changed after QIM steganography, a model called quantized codeword correlation network (QCCN) is constructed based on the split VQ codewords of adjacent audio frames. A high-performance detector is constructed by using Support Vector Machine (SVM) classifier. It can effectively detect steganography used in G.723.1 and G.729 low bit rate audio codec.

For uncompressed format, it includes two methods: cooperative method and non-cooperative method. In the first method, the technique is based on the comparison between the estimated carrier signal and the steganography signal. There are many methods for estimating the carrier signal, including denoising, carrier linear basis, re-embedding and so on. However, it is also possible to estimate the steganalysis signal used for calibration, which was discovered by Ghasemzadeh et al. [14]. The authors proposed a general steganalysis technology based on calibration, which is a reliable audio steganalysis system based on Mel Frequency Cepstral Coefficient (R-MFCC), generating a model with the largest deviation from the HAS model and using genetic algorithm to optimize the dimension of features. In their technology, signals and random messages are embedded using re-embedding method. The energy feature is extracted, in which each signal and the re-embedding signal are divided into many blocks, and the energy of each block is calculated. Then, the energy of each block corresponding to the signal and its re-embedding is subtracted. Finally, the statistical properties of the energy features, including mean, skewness, standard deviation and kurtosis, are selected to train the SVM classifier. Their techniques have been evaluated with a variety of steganography techniques. The experimental results show that this method has a good detection effect in the case of pertinence and universality.

The non-cooperative method extracts feature directly from the audio signal based on the embedded feature domain. Han et al. [14] proposed a linear prediction method that extracts linear prediction LP features from segmented audio files. According to the experiment, the author found that LP parameters can significantly discriminate carrier and hidden information. Therefore, LP coefficients, LP residuals, LP spectra and LP Cepstral coefficients are extracted from time domain and frequency domain. The SVM classifier is trained based on the features extracted from the cover signal and the steganography signal. Extensive experiments have been conducted on different embedding rates and different steganography techniques have been tested. The results show that compared with the popular steganalysis technology at that time, the proposed technology is effective.

Traditional audio steganalysis schemes usually analyze the relevant feature information extracted manually or use traditional machine learning methods. The accuracy of the analysis results of traditional methods can usually be guaranteed at a high level. However, compared with deep learning, traditional audio steganalysis methods also have some problems. First, the algorithm is not universal enough to obtain features that are effective for most audio steganalysis algorithms; in terms of efficiency, it is far inferior to the currently mature deep learning method, and it cannot process data with large feature dimensions. Due to the concept of depth feature, neural networks have become a trend in deep learning and classification tasks in recent years. Both efficiency and accuracy reflect the power of deep learning. Neural networks also have better robustness

and effectiveness. Steganalysis based on deep learning is also increasingly applied and needs more research.

Under the influence of this environment, most of the steganalysis methods in recent years are deep learning. The experimental results also show that steganalysis based on deep learning method is the appropriate research direction. In 2015, steganalysis began to enter the field of deep learning, which is different from the traditional machine learning methods. Qian [15] first added the method of deep learning to the research of steganalysis. They proposed that the task of steganalysis can be regarded as a formulaic binary classification problem to distinguish cover objects and steganalysis objects, and constructed the detection model of steganalysis through two steps of feature extraction and classification. Regarding the development of convolutional neural networks for classification tasks, since the proposal of AlexNet, more and deeper neural networks have been proposed, such as VGG and GoogleNet, and RESNET has solved the problem of gradient disappearance caused by too deep network depth. In 2018, Mehdi Boroumand [16] and others built a steganalysis model based on the deep residual network inspired by the deep residual network. The experimental results show that it has been relatively improved in the area of JPEG images. On this basis, researchers have tried various improved and optimized neural network models for steganalysis, which have achieved good results.

Convolutional neural network in audio steganalysis, Chen [17] first proposed a audio steganalysis model (ChenNet) based on convolutional neural network (CNN) to detect LSB (least significant bit matching) steganalysis in time domain. Then, Lin [18] improved the convolutional neural network model and used the truncated linear unit and residual module, which were effective in image steganalysis, to optimize it. Experiments proved the effectiveness of the model optimization. For the steganalysis of MP3, Ren [19] proposed a universal audio steganalysis method of MP3 and AAC (Advanced Audio Coding) based on the deep residual network (RESNET) and used the audio spectrum as the network input. Taking the spectrogram as the input feature can effectively detect voice steganography based on ACC and MP3. At the same time, taking the spectrograms of different sizes as multi-scale input, a group of high-pass filters are designed to distinguish the spectral energy differences between time and frequency caused by different steganography methods. The classification results are obtained by using the deep residual network training, this scheme has relatively better performance than the scheme using quantized modified discrete cosine transform MDCT coefficient and Mel spectrum as feature input. Zhang [20] also proposed a audio steganalysis method based on time domain, improved the convolutional neural network and used the deep residual method to build the model.

Inspired by the inception module, Li [21] and others used different convolution kernels in their model to increase the width of CNN architecture, and then connected them and tried to use different activation modules dam to form a new CNN architecture for image steganalysis. The experimental results are better than the existing models. The multi-scale network structure is used for multi-scale feature fusion. Different sizes of convolution kernels are used to obtain different outputs, and then the depth superposition becomes a new output feature. In CNN, the receptive field of the high-level network is

different from that of the low-level network. The network extracts target features by layer-by-layer abstraction method. Different scale features have different effects on the results of the classification task. The model can be optimized by multiscale convolution kernel. The multi-scale model is derived from the proposed architecture of a deep convolutional neural network code named Inception [22].

Using the deep learning method to study VoIP steganalysis, Lin et al. [23] proposed an effective online steganalysis method to detect QIM steganalysis. This method is based on the code word correlation model of recurrent neural network (RNN), which can separate the relevant features into carrier audio and steganalysis audio, and then effectively detect QIM steganalysis. This method is the first deep learning network applied to the steganalysis task of network flow. In 2019, Yang [24] proposed a correct method that can combine the advantages of CNN and LSTM architecture, which uses bidirectional long-term and short-term memory recurrent neural network (BI-LSTM) to capture long-term context information in the carrier, and then uses CNN to capture local and global features and time carrier features to detect QIM-based steganalysis.

In the same year, Yang [25] analyzed the correlation of carriers and proposed a novel and very fast steganalysis method for VoIP streams. The vector quantization codewords are mapped into the semantic space, and only a hidden layer is used to extract the correlation between codewords. This method can quickly and accurately detect possible steganalysis in VoIP streams. In 2020, Yang et al. [26] designed a lightweight neural network called fast correlation extraction model (FCEM), which extracts features from VoIP frames only based on an attention variant called multi-head attention, and is significantly superior to the relatively complex recurrent neural network (RNN) and convolutional neural network (CNN) in terms of accuracy and time efficiency. In 2020, hu [27] designed a hierarchical representation network to solve the steganalysis problem of QIM steganography in low bit rate audio signals, and applied the three-level attention mechanism to different convolution blocks, so that it can pay different attention to contents of different importance in audio frames.

In 2022, Yang [28] developed a multi-channel convolutional sliding window (CSW) to analyze the correlation between a given frame and adjacent frames in VoIP signals, using two different channels to extract high-level features and low-level features, respectively, and analyzed the classification after linear layer fusion. This method has good performance in the steganography scheme with low embedding rate, and has greatly improved the detection efficiency of the model. It almost realizes the real-time detection in the process of VoIP communication. The above methods have limitations on the generality of many audio steganalysis methods. This method is mainly used to detect the steganography of QIM, and the effect of other steganography needs to be further studied.

Li [29] proposes a general frame-level steganalysis method for low bit-rate compressed audio. It uses the dual-domain representation method in time domain and compression domain to extract rich features from audio frames, and introduces an adaptive local correlation enhancement module to effectively model local features, which compensates for the shortcomings of the traditional transformer-based model. In 2022, Tian [30] proposed that in VoIP steganalysis, integer and fractional pitch delays are used as inputs, and a subframe splicing module is designed to organically integrate the integer

and fractional pitch delays of subframes for real-time detection. A spatial fusion module based on pre-activated residual convolution is designed to extract pitch spatial features and gradually increase its dimension to find more subtle steganographic distortion to improve the detection effect, in which the group extrusion weighted block is introduced to reduce the information loss in the process of increasing the feature dimension. A time fusion module is designed that uses stacked LSTM to extract pitch time features, and a gated feedforward network is introduced to learn the interaction between different feature maps while suppressing useless features for detection.

At present, in the field of steganalysis, the methods based on deep learning have gradually replaced the traditional methods. Researchers have focused on deep learning. At present, the steganalysis methods based on deep learning have been continuously improved in accuracy and other aspects, and have caught up with and surpassed the traditional methods. At the same time, they are absolutely ahead in efficiency. But at present, audio steganalysis based on deep learning is still in its infancy, and there is a lot of research space. Many problems, such as the accuracy of low embedding rate, need to be further improved.

According to the current research on audio steganalysis, although the traditional audio steganalysis methods can usually achieve detection, the efficiency is far less than that of audio steganalysis based on deep learning. In this paper, the work of VoIP steganalysis mainly focuses on the neural network based on multi-scale and attention mechanism, and improves the accuracy and efficiency of the model by optimizing and adjusting the depth model and algorithm. The main contents of this paper are as follows: A voice steganalysis method of VoIP based on transformer is proposed. In order to solve the problem of high resource consumption in the optimization depth network, combining the respective advantages of transformer and CNN, the model uses the relative position coding method to obtain the location information missing in the model, multi-scale convolution operation is introduced to obtain the local feature information, which is complementary to transformer, and the SAT neural network model is constructed for VoIP steganalysis. The multi-head self-attention mechanism improves the processing ability of long sequence audio data, enhances the generalization ability of the model, and realizes the high efficiency and high accuracy of VoIP steganalysis.

3 Proposed Model for VoIP Steganalysis

Transformer is a self-attentive neural network model originally used in natural language processing. It can effectively process long sequence data. Compared with CNN and RNN, the computational complexity of each layer is lower, but the local information acquisition ability is not as good as CNN and RNN [31]. In order to give full play to the respective advantages of CNN and transformer, form a benign complement, and reduce the computational resources and training time of the model under the premise of ensuring the detection accuracy. Our proposes a steganalysis transformer (SAT) method based on transformer neural network to analyze VOIP voice data, so as to recognize the steganalysis information embedded in VoIP voice signal. The technologies including relative position coding [32], multi-scale feature fusion method, mapping method, self-attention mechanism, residual connection and normalization are adopted. The relative

position coding is used to obtain the accurate position information in the self-attention calculation process, the multi-scale feature fusion method is used to improve the local feature extraction to obtain the detailed information, and the mapping method is used to adjust the sample shape. The self-attention mechanism reduces the dependence on external information, it can better capture the internal correlation of features and solve the problem of long dependence. The residual connection can prevent the gradient from disappearing, and the normalization can improve the generalization ability and effect of the model. Through the experimental verification of the model, it is proved that the SAT method can train faster and effectively detect the hidden information in VoIP audio signal.

We propose a VoIP steganalysis method based on satellite neural network, design a low-cost neural network model, and ensure the accuracy of hidden information detection at the same time. We use multi-scale feature fusion method and multi-head self-attention mechanism to improve the performance. The voice steganalysis method of VoIP based on SAT is proposed. By combining transformer and CNN for modeling, the advantages of both can be complemented, the generalization ability of the model can be improved, and the modeling ability of spatial and temporal features of the model can be enhanced. At the same time, the acquisition of global and local features is taken into account. The parallel computation of transformer greatly reduces the computational cost, and the strong feature extraction ability of CNN ensures the recognition accuracy of the model.

As shown in the Fig. 1, in our method to detect whether the VoIP voice carries secret information, we first preprocess the original VoIP voice data to extract the linear prediction parameters and adaptive codebook parameters in the voice signal, and then divide the two sets of features into training set and test set. The training set is used for training the SAT neural network model, and the test set is used for experimental testing. In the model training, the audio data samples are coded by relative position, and then the features are fused after multi-scale convolution, and then the fused features are mapped. The improved transformer module is used to compute self-attention, and multiple are superimposed to form multi-head self-attention. Combined with residual and normalization operations, the multi-layer perceptron classifier is used for classification, The steganalysis of VoIP audio is completed by distinguishing carrier samples from carrier samples. After the error of the model converges, the SAT model is stored and tested online with the test set, and the output results are used to determine whether the audio signal carries secret information.

3.1 Pre-processing VoIP Data

Before entering neural network model training, the data must be pre-processed. This process can control the data quality, standardize the data and optimize the algorithm. In the process of VoIP covert communication, it is necessary to sample the original VoIP voice signal and convert the analog voice signal to digital signal before it can be transmitted by the network. Usually, the WAV format is used as a lossless voice file format to store high quality voice data and protect the steganography information from damage. Next, the wav format must be converted into a pure PCM voice file. The PCM voice file does not contain any other metadata. After that, the pure voice data

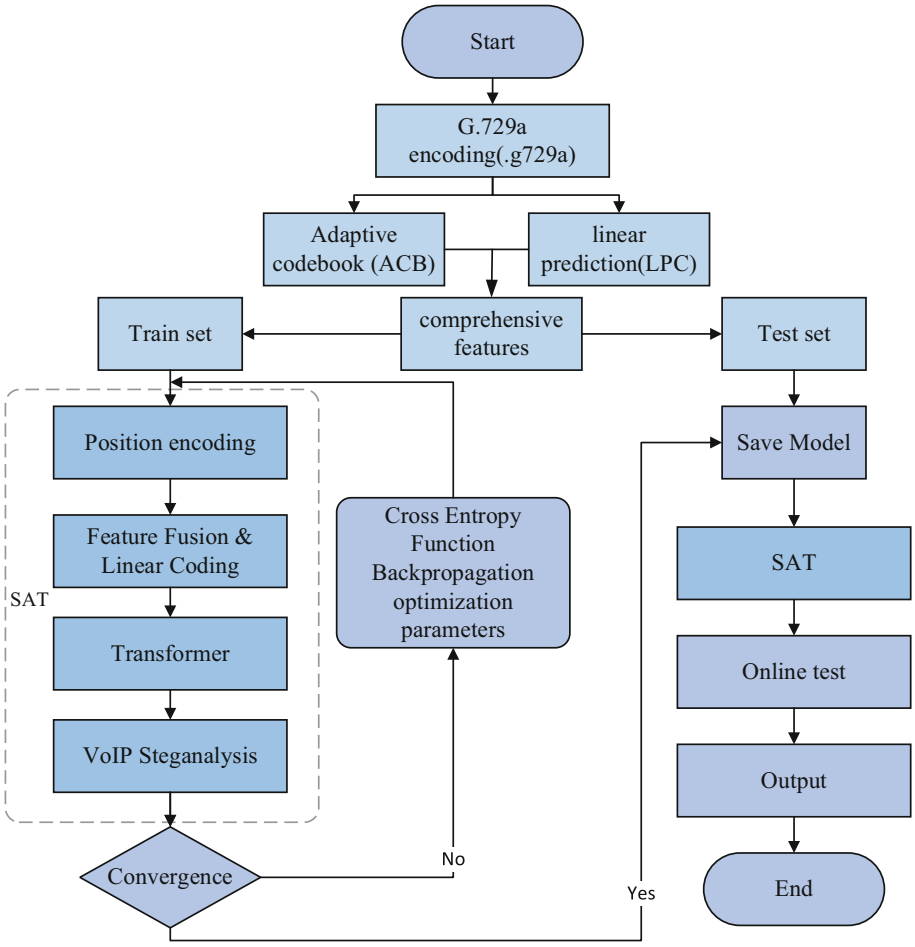

Fig. 1. Flow chart of SAT steganalysis algorithm

needs to be encoded into the PCM format. The encoders commonly used in VoIP voice communication include G.729, G.723.1, AMR, G.729A, etc.

According to the principle of VoIP communication, analog voice signals are converted into digital voice signals and compressed to reduce the bandwidth and delay of data transmission. Then, after negotiation, the compressed digital voice signal is divided into multiple packets for packet transmission, and header information is added to each packet. Unused fields of the protocol header are usually used to hide information; after receiving the digital voice signal data packet, the receiver reconstructs the header information to obtain a complete digital voice signal. Finally, the digital voice signal is restored to an analog voice signal and played back to complete the VoIP voice communication. Steganography is usually embedded in the compression coding process, so we extract the corresponding feature parameters in the compression coding for steganalysis.

The audio sample we use uses the coding format of G.729A. G.729A can provide high quality, high bit rate audio communication, and can compress the 16kHz, 16-bit precision PCM audio signal to a bit rate of 8kbps, ensuring high audio quality and low delay. Generally, the steganography of VoIP payload is concentrated in this compression process, so the corresponding decoder is used to obtain the corresponding features. The steganography method used in VoIP payload usually affects LPC parameters and ACB parameters, so extracting these two audio parameters as the input of neural network can effectively detect VoIP steganography.

We divided the G.729A compressed audio samples with sampling time of 1s into 100time frames, and extracted 7 LPC parameters and ACB parameters from each time frame. In order to meet the operation requirements of the subsequent neural network models, the combined features added one dimension. Finally, the data set is divided, the training set is used for model training, and the test set is used for evaluation test. Then, the position coding information is added, and the relative position coding is used to add the dominant position information to the sample.

3.2 Modeling and Training

Build a SAT model for training, and save the model after convergence for subsequent testing. The constructed model mainly consists of four main parts: location coding, feature fusion, linear coding and transformer. Location coding solves the missing location information in the computation of the attention mechanism. Feature fusion obtains local and global information. Linear coding is realized by mapping. The features within the method are used by transformer to compute self-attention and solve the problem of long-distance feature detection of time series.

3.3 Online Test

After the model is saved, the model structure and the saved optimal parameters are reproduced. The test set is put into the model for evaluation and testing, and the predicted label is obtained. The results are obtained after comparing with the real label, and whether the secret information is carried in the output carrier is obtained.

The overall structure of the SAT model is shown in Fig. 2. There are four parts in total. Figure 2(a) is the overall structure of the model, including data preprocessing part, position coding part, multi-scale feature fusion part, and improved transformer part. Figure 2(b) is an enhanced transform module, where norm represents the normalization operation. Here, layer normalization is used to reduce the gradient disappearance problem. X is an abstract multi-head self-attention module. MLP is a multilayer perceptron composed of a full connection layer, which also combines the operation of residual connection. Figure 2(c) is a multi-scale feature fusion module, which includes convolution operations and pooling of different scales. Figure 2(d) is the multi-head self-attention calculation module, which is the core of the transformer. The samples after linear coding calculate self-attention, and the superimposed h heads are fused to form multi-head self-attention.

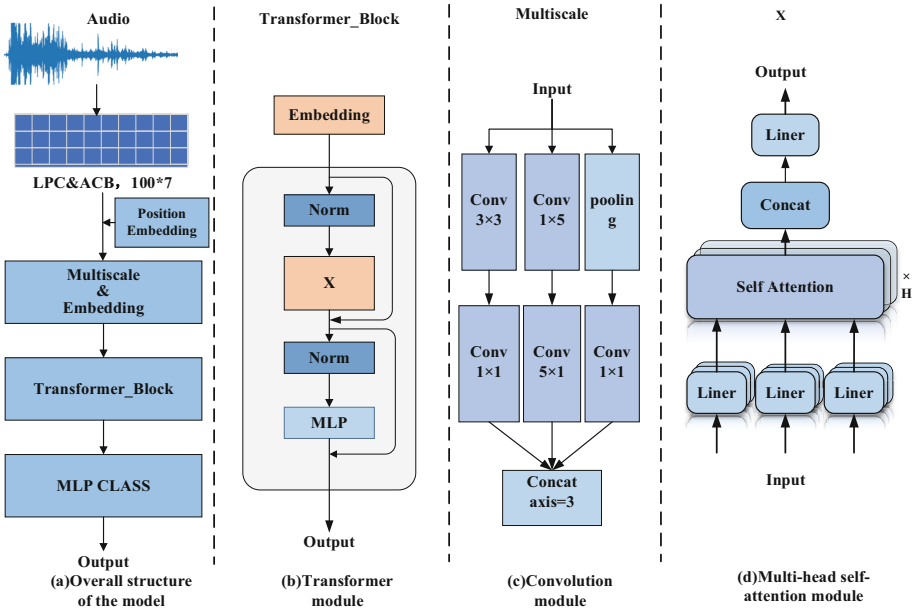

Fig. 2. SAT model structure

3.4 Relative Position Encoding

Position encoding describes the location and relationship of entities in the sequence and gives a unique representation to each element in the sequence. The self-attention used in the transformer neural network is a global operation that is insensitive to location information. Changing the location coding between two elements will not affect the result, and it does not have the ability to learn the word order of RNN. Therefore, it is necessary to add position encoding information. The traditional CNN model, due to its local processing mechanism, that is, it only works on several neighbors of the target element, so it will be limited in dealing with the problem of long sequence, but the relative position between elements can be noticed. Some previous work has also shown that CNN can learn some position information through padding, but how to express the position explicitly is still a problem. So, although there are some CNN operations in the model, this chapter marks the position information by adding position encoding to the encoding process.

The commonly used position encoding methods are absolute position encoding and relative position encoding. Absolute position coding generally uses learnable absolute position encoding, that is, the representation of each position is added directly to the representation of the token as a learnable vector. There are also problems with this method: (1) The position encoding itself needs to be learned by a large amount of data. If the length of the training set exceeds the maximum length, the model cannot learn the position information; (2) the relative relationship between position vectors is not exploited. Relative position encoding requires only a limited number of position codes that can express the relative position of any length, so it can process any length of data.

Relative position coding refers to the direct consideration of the relative position between two tokens when calculating the attention score. The output calculation elements z_i are the weighted sum of the input elements of the linear transformation, and the relative position information is taken into account, which can be expressed as Formula 1.

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + a_{ij}^V) \tag{1}$$

Calculate the weight coefficient using the Softmax activation function:

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \tag{2}$$

Compare the compatibility functions of two input elements to calculate e_{ij}, and add the relative position calculation:

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K + a_{ij}^K)^T}{\sqrt{d_z}} \tag{3}$$

3.5 Feature Fusion and Linear Encoding

CNN can capture local dependencies well, and transformer can obtain global dependencies by using multi-head self-attention mechanism. The two methods complement each other. At the same time, to further verify the effectiveness of the multiscale method proposed in the previous chapter, the multiscale feature fusion method proposed in the previous chapter is added to the SAT model to introduce the convolution operation.

The various operations of VIT [33] and Swin Transformer [34] in patch embedding are all for the purpose of adjusting the size of the input. A large input size is too expensive for the transformer to process, so the same processing method as for text is adopted for the image to ensure that the transformer can process it, so we can also process the audio samples accordingly. Patch embedding in vit is used to convert two-dimensional images into a series of one-dimensional patch embeddings. The input size is (H × W × C), According to Patch Size to get the size of P × P, The number of image blocks of is N = H × W/ P2. Where n can be understood as the sequence length, and the dimension Dim (size P2 × C) of each element in each sequence, It is called patch embedding.

The linear encoding in our model is realized by the mapping layer, and the size of the audio sample after preprocessing and extracting the feature parameters is 100 × 7. Since the audio samples are divided according to the time frame during feature extraction, we take the length of the time frame 100 as the sequence length, and the 7-dimensional features are mapped to the corresponding sample dimension. The final dimension is to maintain the same size of the original input data and the element relationship of the original sequence.

3.6 Transformer Module

Our model improves the transformer module, maintains the overall structure of the model, abstracts the core part, and reduces the residual operation, making it more suitable for VoIP voice steganalysis. At the same time, it does not overlap the module many times to reduce the complexity of the model. The performance advantage of the transformer is mainly due to its structure. When we abstract the multi-head self-attention of the encoder part, we get the module Meta-former [35], and use the multi-head self-attention method in the abstract block. Self-attention can reduce the dependence on external information and pay more attention to the internal correlation of data, so as to solve the problem of long dependence. Using multi head self-attention can make the model focus on multiple key information in the whole situation, so as to solve the problem of self-attention over focus on itself. Therefore, the transformer architecture is maintained in the model construction to verify the effectiveness of the attention mechanism in VoIP voice steganalysis.

$$X = InputEmbedding(x) \tag{4}$$

where, x is the input of the sample, and X is the input of the transformer after the embedding.

The two residual connections can be expressed as:

$$X = Attention(LN(x)) + x \tag{5}$$

where, x is input, Attention is multi-head self-attention, LN is layer normalization, and batch normalization can also be attempted.

$$X = \sigma(LN(x)W_1)W_2 + x \tag{6}$$

where it is mainly composed of a two-layer MLP with nonlinear activation, x is the output of the first remaining connection, and the activation function here is ReLU (or GELU). The two-layer MLP with nonlinear activation has two full connection layers. After passing through a full connection layer, it is activated by the ReLU function and then connected to a full connection layer to get the output.

The structure of the MLP is shown in Fig. 3. There are two complete connection layers and a ReLU activation function.

Self-attention in Transformer is the core of it. Self-attention can be thought of as learning a relationship, the relationship between the current element and other elements in a sequence, and the dependency can be computed directly regardless of the distance. The attention function can be described as mapping a query and a set of key-value pairs to the output, which can be obtained by calculating the weighted sum of the values.

Attention function in the model through the matrix way to achieve parallel computing. First calculate the point multiplication result of query and all keys, and then multiply by the scaling factor 1/k d to prevent the product result from being too large. Send the above calculation results to the Softmax function to obtain the weight corresponding to the value. According to this weight, you can configure the value vector to get the final output. The process of calculating the output of attention relationship can be simply expressed as the function attention (Q, K, V), and the calculation formula is:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

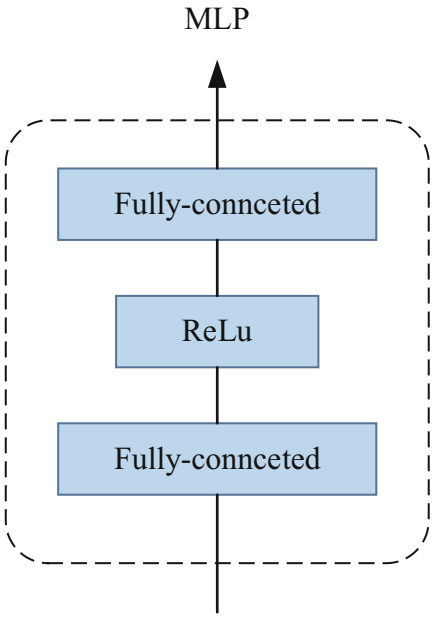

Fig. 3. MLP

The multi-head attention module is shown in Fig. 2(d). Multi-head self-attention uses multiple queries to compute and select multiple information from the input information in parallel. Each attention focuses on different parts of the input information, initializes multiple groups of different (Q, K, V) matrices by different coding forms to compute multiple attention, and then connects them. Its calculation formula is expressed as:

$$\begin{aligned} MultiHead(Q, K, V) =& Conact(head_1, \ldots, head_n)W^o head_1 \\ =& Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{8}$$

Finally, MLP is also used to achieve classification, and sigmoid function is used to output the samples that distinguish steganography from non-steganography. Or add a classification label in the embedding to output the same classification effect.

4 Results and Discussion

4.1 Data Set

The dataset we use is from VStego800k, which is a large dataset for voice steganalysis over VoIP, and is a mixture of different steganography algorithms, embedding rate, and quality factors. The duration of all samples in the data set is uniformly shortened to 1 s, and the collected audio signals are mixed with male and female speakers, Chinese and English. This dataset contains 814592 streaming audio clips, which are divided into 50000 samples as the test set and the remaining samples as the training

set. VStego800k uses two typical streaming audio steganography algorithms for each steganographic sample, and randomly selects the embedding rate of 10% -40% to embed secret information.

The audio sample data set used in the experiment is a total of 310000 Chinese audio samples, including steganographic and non-steganographic samples, with a ratio of 1:1. In this chapter, the data set is divided according to the ratio of 8:2, and gets 248000 samples as the training set and 62000 samples as the test set. The sample steganography method is randomly selected from CNV-QIM [36] pitch steganography method [37], and the embedding rate is randomly selected from 10%, 20%, 30%, and 40%.

Each segment of the sample is divided into 100 time frames. In each frame, a total of 7 data including LPC features and ACB features are extracted. Finally, the shape of the feature sample in the input model is 100×7. For the feature parameter selection, it is based on the steganography of LPC parameters. It can be observed that the relevant special effects of split vector quantization codeword of linear predictive coding filter coefficients have changed after QIM steganography. In audio coding, the LPC coefficients of each frame are extracted and converted into line spectral frequency (LSF) coefficients. Then the LSF coefficients are coded using the VQ format. Low bit rate audio coders typically use split VQ. For example, in the G.729 or G.723.1 standards, the quantized LSF coefficients are described by the quantized codeword set $C = \{C1, C2, C3\}$, where C1, C2, and C3 are codewords selected from codebooks C1, C2, and C3, respectively. The QIM steganography scheme hides the secret information in the VQ process. This method is easy to implement and has low computational complexity. Therefore, this method is suitable for establishing covert channel in VoIP. Compared with MFCC Mel Cepstral Coefficient or LPL Linear Perceptual Prediction, which are commonly used by mainstream audio recognition systems in the past, LPC used in this chapter is based on the characteristics of human voice mechanism, which mainly distinguishes different vocals and vowels by the formant distribution position. Formants are areas where the energy of sound is relatively centralized in the spectrum. The frequency and bandwidth of formant can be well calculated by using LPC parameters to characterize the sound. The steganography method in the example is based on the fact that the LPC parameters will also change to a certain extent after steganography, so the extracted LPC features can be used as learning parameters.

ACB parameter is a parameter extracted from pitch feature. Pitch parameter is an important feature parameter in audio signal processing, which is used to describe the pitch periodicity of audio signal. Pitch period refers to the time of sound wave vibration repetition in each cycle of audio signal, and it is one of the most periodic parts of a audio signal. Pitch extraction algorithm is usually used to extract pitch period from audio signal and represent it as pitch parameter. Pitch frequency refers to the number of repeated pitch cycles per second in the audio signal, which is also the reciprocal of the pitch cycle. It can reflect the speaker's voice pitch, tone change and other information. Pitch intensity refers to the energy of the pitch part in the audio signal. It can reflect the speaker's emotional state, speaking posture and other information. Pitch phase refers to the difference between the start position and end position of the pitch part in a audio signal. It can reflect the speaker's speaking speed, voice mode, and other information.

Pitch-based steganography uses the pitch periodicity of audio signals to hide information. The embedding process of hidden information: convert the information to be hidden into binary code, and select different embedding schemes according to the parity of pitch periods. For example, when the pitch period is even, "0" and "1" can be represented by positive and negative cadence, while when the pitch period is odd, "0" and "1" can be represented by high and low rise and fall. In the process of pitch steganography, small amplitude change determines the value of pitch period, so the extracted pitch feature information can effectively enable the model to learn and classify. LPC parameters and ACB parameters can be steganography using either LSB or QIM. Therefore, selecting these two feature parameters can perform most of the VoIP voice steganalysis to improve the universality of the model.

4.2 Evaluation Index

The common evaluation indicators of steganalysis include accuracy, detection time, false detection rate and missing detection rate. Detection accuracy is the most important. The model training time is also a reference index. The detection time can reflect the detection efficiency of the model, but it is affected by the computer configuration. Steganalysis is a binary classification task in deep learning, so we can refer to the evaluation index of binary classification. There are two kinds of steganalysis samples, including cover and stego. Set stego as positive sample P = Positive and cover as negative sample N = Negative. There are:

$$acc = \frac{X_{tp} + X_{tn}}{X_{tp} + X_{tn} + X_{fp} + X_{fn}} \tag{9}$$

where: acc is the accuracy, X_{tp} is the number of samples correctly classified as steganographic, X_{tn} is the number of samples correctly classified as non-steganographic, X_{fp} is the number of non-steganographic samples incorrectly classified as steganographic, X_{fn} classifies the number of steganographic samples as non-steganographic samples for errors [38]. At the same time, the following indicators can reflect the performance of the model:

The detection sample duration is used as an evaluation index to assess the performance of steganography detection.

$$\overline{T} = \frac{T_{test}}{T_{Sam}} \tag{10}$$

where: $\overline{T}$ is the average detection duration, T_{test} is the total duration, T_{Sam} is the total length of the sample.

4.3 Experimental Environment

The model designed in this paper is written in Python language based on Tensorflow2.4 deep learning framework, and runs on win10 system with Intel ® Core™ I7-12700kf, GPU adopts NVIDIA GeForce RTX4090. The main parameters of the training process are set as follows: the learning rate is 0.001 by default, the batch size is set to 64, and the optimizer uses Adam.

4.4 Experimental Results

Table 2 shows the experimental results of VoIP steganalysis based on the SAT model. It can be seen that the final accuracy of the experimental results is at a high level. In the total of 62000 test samples, 1637 cover samples were detected as stego samples, and only 785 stego samples were misjudged as cover samples, that is, stego samples were not correctly detected. The test recall rate reached 97.46%, the accuracy rate of the test results reached 96.41%, and the comprehensive evaluation index F1 score was 96.45%, which can prove the effectiveness of the proposed method for VoIP steganalysis. Figures 4 and 5 shows the confusion matrix drawn based on the experimental results. From the confusion matrix, we can intuitively see that the performance of the VoIP steganalysis model is good in detecting stego-like positive samples, and can detect positive samples with secret information (Table 1).

Table 1. Sample processing method

Encoder	Steganographic algorithms	Parameters	Sample length	Embedding rates
G.729A	CNV-QIM	LPC	1s	0–40%
G.729A	PMS	ACB	1s	0–40%

Table 2. Experimental result

TP	TN	FP	FN	FPR	FNR	Precision	Recall	F1	ACC
30251	29523	1570	656	5.04%	2.12%	95.06%	97.78%	96.45%	96.41%

Figure 4 describes the process that the accuracy of the model of the training set and the test set continues to improve with the number of iterations, where the number of iterations is set to 20, and the number of iterations refers to the process of updating the weight parameters each time during the training process. Typically, we need to train the model repeatedly until the performance of the model tends to be stable. In Figs. 4, 5 and 6, the number of iterations is set to 20, and the model has improved significantly within this number of iterations. However, it should be noted that too few iterations can lead to under-fitting of the model, while too many iterations can lead to over-fitting. ACC reflects the model's accuracy of detection; ACC refers to the model's accuracy of detection, that is, the proportion of predicted results that agree with the actual results. The higher the ACC, the higher the accuracy of the model. Therefore, we hope to evaluate the performance of the model and adjust and optimize it by monitoring the ACC on the training set and the test set. From the training process diagram of the experiment, it can be seen that the accuracy of the training set increases and converges with the number of iterations, and the test validation set gradually increases with the number of iterations and finally tends to be stable. The accuracy of the training set with 20 iterations is 99.4%, and the accuracy of the test set is 96.41%.

Figure 5 describes the change in the loss function, which is an indicator to measure the difference between the predicted value and the actual value of the model. The loss value is the value calculated by the model according to the loss function in the training process. The loss value essentially reflects the performance of the model in the training process. The larger it is, the closer the predicted result of the model is to the real value, i.e., the better the performance of the model. The training loss gradually decreased and converged, and the loss value of the test set also tended to be stable after falling to a certain level. The loss value has two main functions: on the one hand, it is used to evaluate the performance of the model. During the training process, the super-parameters of the model can be adjusted according to the change of the loss value to achieve better performance; on the other hand, it is used to optimize the model. During back propagation, we need to use the loss value to update the model parameters by gradient descent to continuously optimize the model.

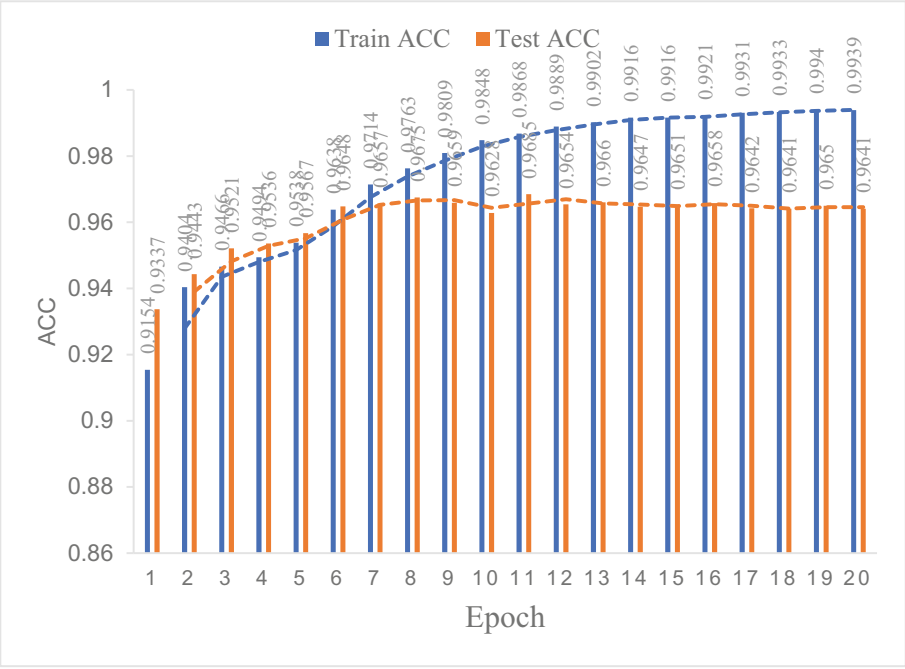

Fig. 4. Visualization of the training process

4.5 Comparative Analysis

In order to evaluate the performance of the proposed model, we compare the performance of the algorithm from many aspects, including comparing the different structures of the model and the existing VoIP steganalysis methods.

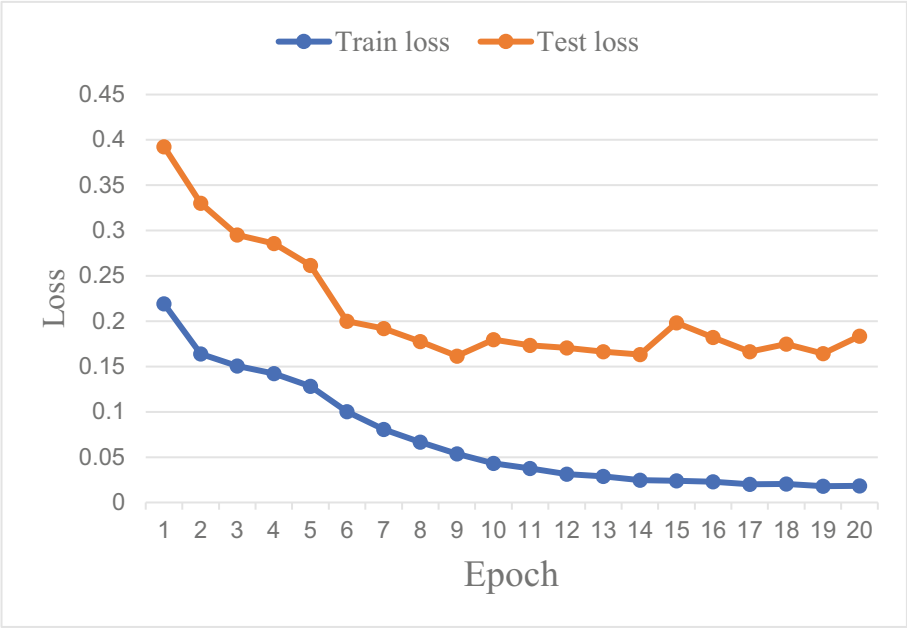

Fig. 5. Loss change curve

The model structure was adjusted and tested accordingly, as shown in Table 3. When we removed the multi-scale module and the self-attention module from the model, the model could not learn the effective parameters, and the final detection accuracy was only 50.21%, which could not detect the secret information in the VoIP voice signal. After removing the multi-scale module, multi head self-attention module and position coding module, the accuracy of the whole model decreased by 2.28%, 2.31% and 1.18%, respectively. Finally, the multi-head self-attention in the transformer structure is abstractly replaced, and the channel attention mechanism used in the table is replaced. It is found that the decrease in detection accuracy is not very large, which is 0.9% lower than the full model in this chapter. The results show that the attention mechanism can effectively improve the voice steganalysis performance of VoIP, and the structure of the transformer can give full play to its advantages of global modeling. Therefore, it is proved that multi-scale feature fusion method and attention mechanism are effective methods to improve the accuracy of voice steganalysis in VoIP.

In order to verify the VoIP voice detection performance of the proposed method, the detection performance of the proposed method is compared with several existing methods in mixed random samples. The mixed samples are randomly selected by steganography method, and the embedding rate is 0.1bps-0.4bps. As shown in Table 4, the comparison methods are SS-QCCN [39], CCN [40], RSM [23], FSM [25], and SFFN [27], and all the methods are tested on the mixed sample test set VStego800K. It can be seen from the table that some existing VoIP steganalysis methods do not perform well in the mixed samples with low embedding rate and short sampling time, and the detection accuracy is less than 90%, while the SAT method proposed in this paper has an accuracy, precision, recall

Table 3. Comparing different model structures

Model structure	Accuracy	Effect
Remove multi-scale and self-attention modules	50.21%	Decrease 46.20%
Remove multi-scale modules	94.13%	Decrease 2.28%
Remove multiple self-attention modules	94.10%	Decrease 2.31%
Remove position encoding modules	95.23%	Decrease 1.18%
Replace self-attention modules	95.51%	Decrease 0.9%
Proposed full model	96.41%	-

and F1 score of more than 95%. CCN and SS-QCCN are traditional machine learning methods based on manual features and classifiers. It can be clearly seen that the proposed method is superior to the traditional machine learning method in the case of complex test samples. RSM-SM, FSM and SFFN are some existing VoIP steganalysis methods based on deep learning. From the perspective of detection accuracy, this method is superior to these methods. Compared with the FSM with the best results among the existing methods, the accuracy of the methods in this chapter has been improved by 7.16%. The above experimental results show that the proposed algorithm has good performance in VoIP steganalysis, and can effectively detect the steganography in VoIP audio signals.

Table 4. Compare existing methods

Steganalysis method	ACC	Precision	Recall	F1 Score
PROPOSED	0.9641	0.9506	0.9778	0.9645
SS-QCCN	0.6117	0.6595	0.4617	0.5432
CCN	0.5542	0.5544	0.5517	0.5531
RSM-SM	0.5174	0.5103	0.8605	0.6407
FSM	0.8925	0.8885	0.8115	0.8100
SFFN	0.7048	0.7206	0.6689	0.6938

To verify the efficiency of the model proposed in this paper, we compared the training time and detection time of the model. The training time and detection time of the model are greatly affected by the hardware equipment. The total number of samples in one round of training is 24000, and the average detection time is the average detection time of each sample in the test. As shown in Table 5, the method proposed in this paper compared with the method of multi-scale and attention mechanism has been shortened to one third under both indicators. Under the condition of detecting the VoIP voice samples with a length of 1s, the detection time has been reduced by 0.68ms, but the detection accuracy has been slightly reduced by 0.94%. The experimental results show that the method proposed in this paper has been greatly improved in terms of efficiency. Under the premise of ensuring

the detection accuracy, it reduces the computational resources and training time of the model, and improves the efficiency of detecting VoIP voice steganography. The average time required by several different methods to detect the steganography information in the audio signal with the sampling time of 1 s is compared. Sat method and MAS method are the average detection time tested in the experimental environment. The detection sampling time of 1 s is 0.32 ms and 1 ms, respectively. Other method data are from FCEM [26]. The detection time of RSM-SM [23] and RCNN [24] methods takes longer than other methods, 4.165 ms and 3.754 ms respectively, HRN [39] method takes 0.512 ms, and FCEM method takes only 0.281 ms to complete the detection.

Table 5. Comparison of 1-s sample detection time

Steganalysis method	Average detection time	Average training time per step
PROPOSED	0.32 ms	21 ms
MAS	1 ms	64 ms
RSM-SM	4.165 ms	-
RCNN	3.754 ms	-
HRN	0.512 ms	-
FCEM	0.281 ms	-

Compare the FCEM method based on multi-head self-attention with the method based on CNN [41]. FCEM is a neural network constructed by using only multi-head self-attention, which realizes fast detection of VoIP steganography. Compared with the traditional audio steganalysis method, the CNN method has higher detection accuracy, stronger robustness and higher computational efficiency.

The FCEM model and the CNN model are reproduced and tested in the same environment. Figures 6 and 7 compare the accuracy, false positive rate and false negative rate of these methods, respectively. Sat is the method of this paper, in which the accuracy of this method is improved by 2.21%, the false positive rate is reduced by 1.4%, and the false negative rate is reduced by 3.13% compared with the FCEM method using only multi-head self-attention; compared with CNN, the accuracy of this method is increased by 4.61%, the false positive rate is reduced by 4.93%, and the false negative rate is reduced by 2.78%.

The experimental results in Figs. 6 and 7 show that the SAT method used in this paper can better detect the steganographic information in the mixed samples, and can achieve effective VoIP steganalysis. Compared with the FCEM model, it shows that the convolution operation introduced by the multi-scale module used in this chapter can compensate for the problem of local information certainty in the transformer, and the combination of the two is benign and complementary. Compared with the CNN model, the attention mechanism can obtain more global features, which proves that the SAT neural network proposed in this paper is an effective voice steganalysis method for VoIP.

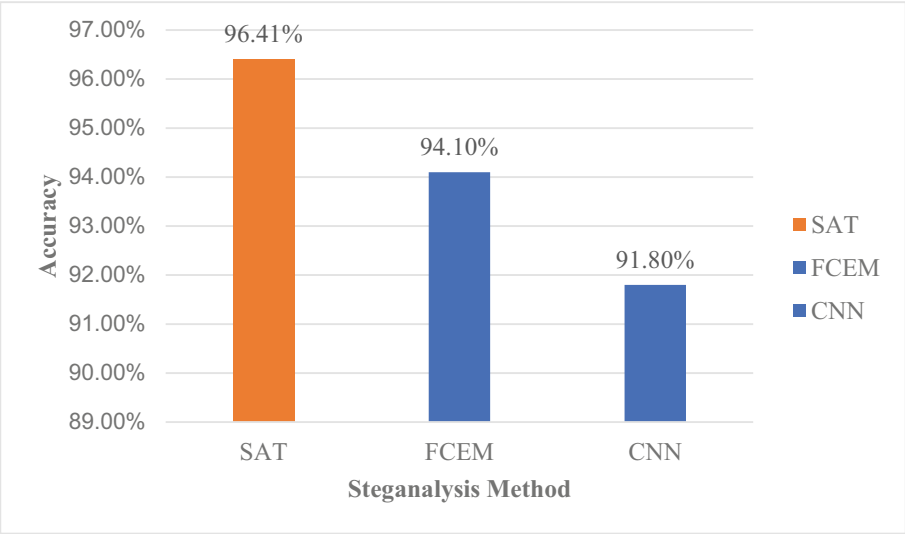

Fig. 6. Comparison of Accuracy with CNN and FCEM

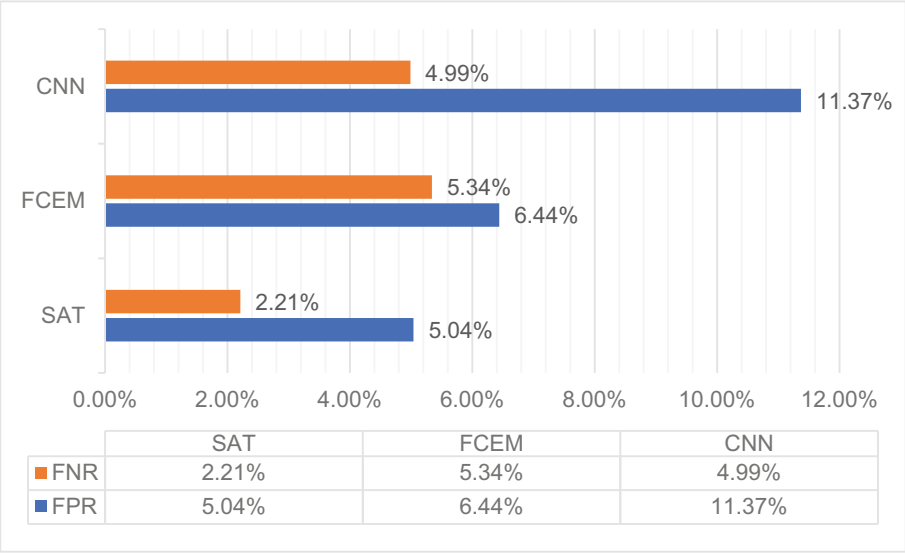

Fig. 7. Comparison FPR and FNR

5 Conclusion

The wide application and characteristics of VoIP network streaming media make it an excellent carrier for covert communication. Research on VoIP-based voice steganography and steganalysis has become a research hotspot. The malicious use of covert communication technology will make VoIP information security more serious. At present, the VoIP steganalysis technology is generally not mature, so it needs to further improve the theoretical system and technical framework. This paper proposes a SAT model based on transformer neural network for VoIP voice steganalysis. This method uses relative position coding to effectively obtain the relationship between sequence elements; combining the multiscale of the previous method, the convolution operation is introduced to enhance the capture of local information, and the dimension of the feature map is reduced by mapping; optimizing the transformer architecture, solving the problem of learning distance dependence with self-attention, and improving the processing ability of long audio sequences to achieve higher detection efficiency and accuracy. Experimental results show that the SAT model can detect steganography with low embedding rate, and the detection accuracy is 96.41%. At the same time, the detection time is also better than most existing methods, which can effectively detect malicious steganography. At the same time, the research also has some limitations. For the field of audio steganalysis, there is currently no high-quality data set in the field, and researchers often use self-made data sets for model experiments. Second, most of the audio steganalysis based on deep learning is actually "semi-blind" steganalysis, which usually extracts different feature information based on the steganography used in the data set for learning analysis, and does not achieve the real universal steganalysis.

At this stage of audio steganalysis, the method based on deep learning has basically replaced the traditional manual method. Future research should also focus on the deep learning method. In deep learning, the size of the data has a great impact on the performance. Therefore, in terms of data preprocessing, it is necessary to study methods that can extract more feature information. Since deep learning is used in the field of steganalysis, researchers usually use the neural network model that performs well in classification learning tasks at the current stage to apply it to steganalysis tasks, and some researchers optimize steganalysis by studying the feature parameters used for learning. There are still many things to consider according to the characteristics of steganalysis in neural networks. Many deep steganalysis models only graft models that perform well in classification. At the same time, there are many excellent methods in neural networks that have not been tested. Steganalysis tasks are also very different from general classification tasks. For general classification tasks, there are large differences between different types, while the steganalysis samples are usually slightly different from the original samples by a small amount of modification, which is a place that needs special attention. At the same time, considering the timeliness of real-time streaming media, we need to balance the accuracy of the model and the detection efficiency of the model. For the transformer-based method, this method has been proved to have excellent performance in various fields, but there is not much research in the field of steganalysis at present, so we need to verify the effectiveness of this method from more angles. In future research on steganalysis, it should be noted that unlike other deep learning classification tasks,

steganalysis itself pays more attention to details, and its characteristics should be taken into account in the research.

Acknowledgments. This work was supported in part by the Education Department of Guangdong Province under Grant 2021KTSCX063, Special topic of basic and applied basic research in Guangzhou under Grant SL2023A04J01043, Guangdong Regional Joint Fund under Grant 2022A1515110693and GPNU Science Foundation under Grant 2021SDKYA 025.

References

1. Priscilla, C.V., Hemamalini, V.: Steganalysis techniques: a systematic review. J. Surv. Fisher. Sci. **10**(2S), 244–263 (2023)
2. Dalal, M., Juneja, M.: Steganography and Steganalysis (in digital forensics): a cybersecurity guide. Multimedia Tools Appl. **80**(4), 5723–5771 (2020)
3. Peng, J., Jiang, Y., Tang, S., et al.: Security of streaming media communications with logistic map and self-adaptive detection-based steganography. IEEE Trans. Depend. Sec. Comput. **18**(4), 1962–1973 (2019)
4. Yi, X., Yang, K., Zhao, X., et al.: AHCM: adaptive Huffman code mapping for audio steganography based on psychoacoustic model. IEEE Trans. Inf. Forensics Secur. **14**(8), 2217–2231 (2019)
5. Peng, J., Tang, S.: Covert communication over VoIP streaming media with dynamic key distribution and authentication. IEEE Trans. Industr. Electron. **68**(4), 3619–3628 (2020)
6. Chaharlang, J., Mosleh, M., Rasouli-Heikalabad, S.: A novel quantum steganography-Steganalysis system for audio signals. Multimedia Tools Appl. **79**(25–26), 17551–17577 (2020)
7. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. **7**(3), 868–882 (2012)
8. Kodovsky, J., Fridrich, J., Holub, V.: Ensemble classifiers for steganalysis of digital media. IEEE Trans. Inf. Forensics Secur. **7**(2), 432–444 (2011)
9. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
10. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
11. Jin, C., Wang, R., Yan, D.: Steganalysis of MP3Stego with low embedding-rate using Markov feature. Multimedia Tools Appl. **76**(5), 6143–6158 (2017)
12. Wang, Y., Yi, X., Zhao, X.: MP3 steganalysis based on joint point-wise and block-wise correlations. Inf. Sci. **512**, 1118–1133 (2020)
13. Li, S., Jia, Y., Kuo, C.C.J.: Steganalysis of QIM steganography in low-bit-rate speech signals. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(5), 1011–1022 (2017)
14. Ghasemzadeh, H., Tajik Khass, M., Khalil Arjmandi, M.: Audio steganalysis based on reversed psychoacoustic model of human hearing. Dig. Sig. Process. **51**, 133–141 (2016)
15. Qian, Y., Dong, J., Wang, W., et al.: Deep learning for steganalysis via convolutional neural networks. In: Media Watermarking, Security, and Forensics 2015. SPIE, vol. 9409, pp. 171–180 (2015)
16. Boroumand, M., Chen, M., Fridrich, J.: Deep residual network for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. **14**(5), 1181–1193 (2018)
17. Chen, B., Luo, W., Li, H.: Audio steganalysis with convolutional neural network. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 85–90 (2017)

18. Lin, Y., Wang, R., Yan, D., et al.: Audio steganalysis with improved convolutional neural network. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, pp. 210–215 (2019)
19. Ren, Y., Liu, D., Xiong, Q., et al.: Spec-resnet: a general audio steganalysis scheme based on deep residual network of spectrogram. arXiv preprint arXiv:190106838 (2019)
20. Zhang, Z., Yi, X., Zhao, X.: Improving audio steganalysis using deep residual networks. In: Wang, H., Zhao, X., Shi, Y., Kim, H., Piva, A. (eds.) Digital Forensics and Watermarking. IWDW 2019. Lecture Notes in Computer Science(), vol. 12022. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43575-2_5
21. Li, B., Wei, W., Ferreira, A., et al.: ReST-Net: diverse activation modules and parallel subnets-based CNN for spatial image steganalysis. IEEE Signal Process. Lett. **25**(5), 650–654 (2018)
22. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
23. Lin, Z., Huang, Y., Wang, J.: RNN-SM: fast steganalysis of VoIP streams using recurrent neural network. IEEE Trans. Inf. Forensics Secur. **13**(7), 1854–1868 (2018)
24. Yang, H., Yang, Z., Huang, Y.: Steganalysis of VoIP streams with CNN-LSTM network. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, pp. 204–209 (2019)
25. Yang, H., Yang, Z., Bao, Y., et al.: Fast steganalysis method for VoIP streams. IEEE Signal Process. Lett. **27**, 286–290 (2019)
26. Yang, H., Yang. Z., Bao, Y., et al.: Fcem: a novel fast correlation extract model for real time steganalysis of VOIP stream via multi-head attention. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2822–2826. IEEE (2020)
27. Hu, Y., Huang, Y., Yang, Z., et al.: Detection of heterogeneous parallel steganography for low bit-rate VoIP speech streams. Neurocomputing **419**, 70–79 (2021)
28. Yang, Z., Yang, H., Chang, C.-C., et al.: Real-time steganalysis for streaming media based on multi-channel convolutional sliding windows. Knowl.-Based Syst. **237**, 107561 (2022)
29. Li, S., Wang, J., Liu, P.: General frame-wise steganalysis of compressed speech based on dual-domain representation and intra-frame correlation leaching. IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 2025–2035 (2022)
30. Tian, H., Qiu, Y., Mazurczyk, W., et al.: STFF-SM: steganalysis model based on spatial and temporal feature fusion for speech streams. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 277–289 (2022)
31. Han, K., Xiao, A., Wu, E., et al.: Transformer in transformer. Adv. Neural. Inf. Process. Syst. **34**, 15908–15919 (2021)
32. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
33. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
34. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
35. Yu, W., Luo, M., Zhou, P., et al.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10819–10829 (2022)
36. Xiao, B., Huang, Y., Tang, S.: An approach to information hiding in low bit-rate speech stream. In: IEEE GLOBECOM 2008–2008 IEEE Global Telecommunications Conference, pp. 1–5. IEEE (2008)

37. Huang, Y., Liu, C., Tang, S., et al.: Steganography integration into a low-bit rate speech codec. IEEE Trans. Inf. Forensics Secur. **7**(6), 1865–1875 (2012)
38. Gupta, A., Chhikara, R., Sharma, P.: A review on deep learning solutions for steganalysis. Int. J. Comput. Dig. Syst. (2023)
39. Yang, H., Yang, Z., Bao, Y., Huang, Y.: Hierarchical representation network for steganalysis of QIM steganography in low-bit-rate speech signals. In: Zhou, J., Luo, X., Shen, Q., Xu, Z. (eds.) Information and Communications Security. ICICS 2019. Lecture Notes in Computer Science(), vol. 11999. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-41579-2_45
40. Li, S.-B., Jia, Y.-Z., Fu, J.Y., et al.: Detection of pitch modulation information hiding based on codebook correlation network. Chin. J. Comput. **37**(10), 2107–2116 (2014)
41. Wang, Y., Yi, X., Zhao, X., et al.: RHFCN: fully CNN-based steganalysis of MP3 with rich high-pass filtering. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2627–2631. IEEE (2019)