A Secure Opus Pulse Steganographic Scheme Based on Message Transform

Yanzhen Ren*, Shan Zhong*, Weiping Tu† and Lina Wang*
* Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
† National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University

***Abstract*—This paper proposes a secure steganography scheme in Opus pulse domain. Opus is a versatile audio codec designed for interactive Internet applications and is widely used in the Internet communication applications. In Opus pulse domain, there exists a large embedding space, which makes it an ideal carrier for steganography. The challenge is that the generation of pulse samples depends on the value of previous pulse samples, so the embedding operation must be carried out at a single sample, otherwise it will lead to distortion drift. This paper proposes a scheme with message transform and rate-distortion evaluation to improve statistical security. The experiment results show that the proposed scheme achieves good auditory quality and statistical security with a large embedding capacity.**

I. Introduction

Steganography is a technique for embedding secret message into digital carriers, such as text, image, audio and so on [1], [2]. Its main purpose is to ensure that the hidden message does not attract others attention except by the intended receiver. As an attack technique of steganography, steganalysis is aimed at detecting the existence of those hidden message to avoid the illegal use of steganography [3]. In recent years, with the rise of social application, all kinds of audio files are flooding our lives, which greatly enriches the diversity of steganographic carriers. As a totally open, royalty-free, highly versatile audio codec, Opus [4] is widely used as the voice-over-IP (VoIP) codec in applications such as WhatsApp, the PlayStation 4. What's more, Opus is standardized by the Internet Engineering Task Force (IETF) and can handle a wide range of audio applications, including Voice over IP, videoconferencing, in-game chat, and even remote live music performances.

In view of the wide range of application scenarios, Opus could be a novel ideal carrier for audio steganography. To the best of our knowledge, there are very few steganographic schemes for Opus. Due to the fact that Opus incorporated technology from Skype's SILK codec [5] and Xiph.Org's CELT codec [6], steganography for those codecs which have the same or similar encoding principles works for Opus. In the case of speech coding, most codecs are under the framework of Algebraic Code Excited Linear Prediction (ACELP). In the existing steganographic schemes, there are three main types of embedding domains: pitch delay [7], [8], Linear Prediction Coefficient (LPC) [9], [10] and pulse [11], [12]. Pitch delay and LPC are the main information of speech signal, representing long-term prediction coefficients and short-term prediction coefficients, respectively. Pulse represents the excitation signal, which is the residual after long-term prediction and short-term prediction and plays an important role in auditory quality. In pitch domain, Gong *et al.* [7] considered the statistical distribution of adjacent subframes for designing the distortion function to achieve higher level of security. Ren *et al.* [8] found that the pitch delay sequences of AMR unvoiced speech do not have short-term relatively stability and proposed an adaptive steganographic scheme in AMR unvoiced speech segment. In LPC domain, Liu *et al.* [9] proposed a steganography for low bit-rate speech codec based on Matrix Embedding (ME) and constructed the mapping table according to the criterion of minimum distance of LPC vectors. Ren *et al.* [10] implemented the embedding operation by modifying the linear spectral frequency (LSF) quantization indices based on the statistical features of LSF codebook. In pulse domain, Miao *et al.* [11] embedded secret message by changing the position of the last fixed codebook (FCB) parameters in each track and the secret message can be extracted according to the relationship of the parameters of the same track. Ren *et al.* [12] proposed an adaptive steganographic scheme (AFA) by designing an additive distortion function on the basis of the optimal probability of pulse and the pulse correlation of the same track.

It can be seen from above steganography that those steganographic schemes are designed according to the characteristics of corresponding embedding domain. Among the three embedding domains of Opus, pulse domain is the most ideal embedding space for the reason that pulse occupies the largest number of bits in the bitstream of Opus. At 16KHz sampling rate, 320 pulse samples can be extracted from each 20 ms frame, while there are only 4 pitch lags and 12 or 16 LPC coefficients per frame. Because of the huge embedding space in pulse domain, steganography for Opus pulse domain is very practical and with great promise. Besides, the modification of Opus pulse has only a slightly effect on the quality of speech. However, the pulse domain in Opus has some different characteristics due to different technical details, which should be taken into consideration when designing the steganographic algorithm for Opus pulse. The most striking feature of Opus pulse is that in noise shaping quantization (NSQ), the quantization for each residual sample depends on previous quantization decisions [13], which means that all pulse samples are not independent and the modification in the quantization process of previous residual samples will change the quantization

results of subsequent residual samples . Thus, steganographic operation on pulse samples must follow the principle of NSQ to avoid a sharp decline in auditory quality.

Due to the great practical advantages of Opus pulse domain and the fact that existing steganography can not apply to this domain, this paper proposes a novel steganographic scheme in Opus pulse domain, which has a large embedding capacity. Given the huge limitations of steganographic operations, the proposed scheme is designed mainly from two aspects. On the one hand, the distribution of embedded message's bitstream is preprocessed by huffman decoding to reduce the modification rate of the steganography. On the other hand, rate-distortion is used to evaluate the modification direction of each pulse sample and that with less rate-distortion is adopted to embed secret message. The main contributions of this paper are summarized as follows:

1) a novel embedding domain with large capacity is discovered. For the reason that pulse represent residual signal and occupies a large number of bits in the bitstream, steganography in Opus pulse domain can achieve a large embedding capacity and maintain good speech quality.

2) limitations in Opus Pulse domain are obtained and a secure steganography for Opus pulse is proposed. There exists a strong relationship between previous pulse samples and current pulse sample. Thus, the steganographic schemes with STCs [14] are not suitable for this domain. The proposed scheme combines message transform and rate-distortion evaluation, which effectively reduces the modification rate and maintains good speech quality.

The rest of this paper is organized as follows. In Section II, Opus codec is briefly introduced and the pulse distribution characteristics of Opus are analyzed. In Section III, a novel steganography for Opus pulse is proposed. The experimental results of the proposed scheme are presented in Section IV. Finally, conclusions and further work are discussed in Section V.

II. Pulse Distribution of Opus

A. Overview of Opus

Opus is standardized by IETF in RFC 6716 and designed for interactive Internet applications. Fig. 1 shows the overview of Opus encoder. As shown in the figure, Opus incorporated technology from SILK and CELT. SILK encoder is based on Code-Excited Linear Prediction (CELP) and used to encode low frequency speech. CELT encoder is based on the Modified Discrete Cosine Transform (MDCT) and designed for music. Obviously, Opus with two techniques can adapt to a wider range of situations and achieve better quality than either SILK or CELT, which makes it a good carrier for steganography. In this paper, we mainly focus on the voice part of Opus, and when we refer to Opus, it represents Opus in the voice mode only. In the voice mode, the pulse domain for steganography is generated in the process of noise shaping quantization and the proposed scheme is designed according to its characteristics.

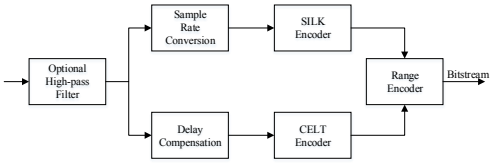

Fig. 1. Overview of Opus encoder [4].

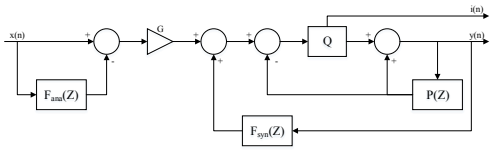

Fig. 2. NSQ block diagram [13].

B. Generation of Pulse in Opus

Pulse represents the excitation signal. In NSQ module, the residual signal is quantized and thereby the excitation signal is obtained. A simplified block diagram of NSQ is shown in Fig. 2. In this figure, $F_{ana}(z)$ and $F_{syn}(z)$ are the analysis and synthesis noise shaping filters. *P(z)* does LPC and LTP. The quantizer *Q* quantifies the residual signal in a series of operations to obtain the value of pulse, that is, the excitation signal. And *i(n)* denotes the quantized excitation indices, which point to the generated pulse value. The residual signal *r* is given by:

$$r = sign(dither) * (X(n) - LPC\_pred - LTP\_pred + NSF) \quad (1)$$

Where *LPC_pred* and *LTP_pred* represent the output of short-term prediction and long-term prediction. And *NSF* is the output of noise shape feedback. *dither* is generated by a pseudorandom generator, implemented with linear congruent recursions on previous quantization decisions within the same frame.

Then, two quantization level candidates are obtained based on *r*. And the final candidates is determined by the measurement of rate-distortion. Finally, the final candidate is quantified to the pulse value. In general, the larger the final candidate, the larger the pulse value. As shown in Fig. 3, the functional relationship between them is roughly a stepped piecewise function.

Because of the implementation mechanism of *Q* and the generation principle of *dither*, quantization error of each residual sample depends on previous quantization decisions. To improve rate-distortion performance, Opus adopts a Viterbi delayed decision mechanism [15]. The number of samples delay and Viterbi states change with the complexity settings. The higher the complexity, the better the performance of rate-distortion. At the lowest complexity level, the number of samples delay is zero. After the quantization process, those generated pulse samples are coded losslessly into the bitstream by entropy coding. In the process of decoding,

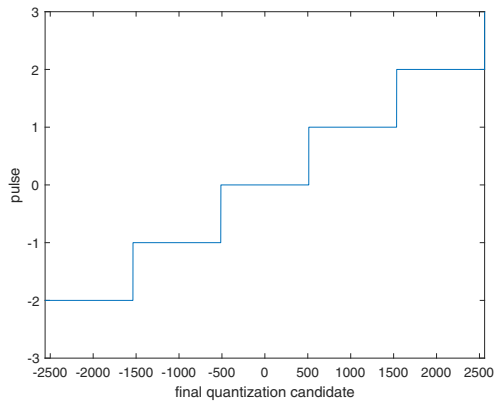

Fig. 3. Simplified function curve of the quantization of the final candidate.

through the inverse process of quantization, speech signal can be reconstructed by the pulse samples and other prediction coefficients.

Through the analysis of generation process of pulse in Opus, several aspects that need to be taken into consideration for the design of steganographic scheme in Opus pulse domain are obtained.

1) All pulse samples are not independent. Thus, the modification of pulse for embedding message must be made one after another during the quantization process. If the pulse samples are changed independently without the information in NSQ and ignoring the bad effect on subsequent pulse samples, such as the modification method in STCs framework, extra noise will appear in the process of decoding and the auditory quality is poor.

2) At high level compexity settings, because of the Viterbi delayed decision mechanism, the modification path of pulse is uncertain. Viterbi states produces a series of candidate quantization results. Combined with the number of samples delay, those candidates constitute multiple paths. The final quantization result is determined by the sum of the rate-distortion performance of subsequent delayed samples. The challenge is that which Viterbi state needs to be modified to embed message is uncertain at current sample. What's worse, it can't be delayed for the charateristic that the modification must be made one after another referring to the current state of NSQ for some side information. One way to solve that is to modify all the current Viterbi States, which introduces unnecessary distortions for extra modifications.

C. Characteristics of Opus Pulse

As there exists huge limitations in Opus pulse domain, a detailed analysis of the characteristics of pulse domain is needed for a secure steganographic scheme. In this section, the capacity, histogram feature and markov feature will be analysed.

a) Capacity

In Opus codec, the number of pulses is equal to the number of samples. In the voice mode of Opus, the internal sampling frequency supports 8KHz, 12KHz and 16KHz. Thus, under the sampling frequency of 16KHz, the number of pulse samples is 16000 per second. It can be seen that in Opus domain the embedding capacity is comparable with that in time domain. The characteristic of the large embedding capacity not only ensures the practicability to some degree, but also enhances the flexibility of the design of secure steganographic algorithms.

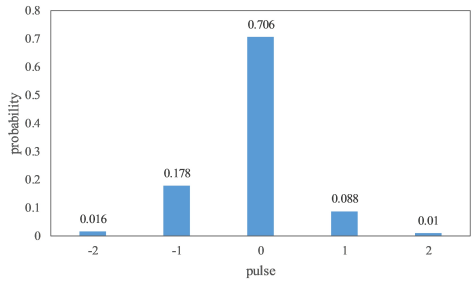

Fig. 4. The histogram of pulse samples. Those pulse samples which are not in the range of [-2,2] are not shown in this histogram for too less quantity.

b) Histogram feature

To analyse first-order statistical characteristics of pulse domain, 2000 audio samples are used to extract pulse samples and then the number of every pulse value is counted and averaged. Fig. 4 shows the histogram of pulse samples. The horizontal axis represents the value of pulse samples, that is, the quantized value of residual signal. The vertical axis is the statistical probability of corresponding pulse sample. The following conclusions can be drawn from this figure:

1) More than 95% of the pulse samples are distributed in [-1,1], with 0 accounting for about 70%. As all pulse samples are coded in the bitstream and can be extracted losslessly in the decoder, pulse samples occupy a large proportion of the compressed audio. To reduce the number of bits consumed by pulse samples, the residual signal are quantified to small integers. The subtlety of the quantification depends on bitrate.

2) There are more negative values than positive ones. The distribution of pulse is not symmetrical. In the design of steganographic scheme for pulse domain, changes in this characteristic should be avoided.

c) Markov feature

Markov feature reflects the relationship of adjacent pulse samples. Fig. 5 shows the Markov transition probability feature of Opus pulse. It can be seen that the values of the central pulse sample pairs are larger than those surrounding ones, which is consistent with the trend of histogram feature. As the quantity of pulse samples out of the range of [-1,1] is too little, the corresponding Markov transition probability is very low. In general, the Markov feature and histogram feature shows similar characteristics to some extent, that is, the distribution of pulse samples is mainly concentrated in the range of small values. In addition, from the diagonal point of view, Markov transition probability is roughly symmetrical.

Through the analysis of pulse domain, it can be found that the main challenge of steganographic schemes for Opus pulse

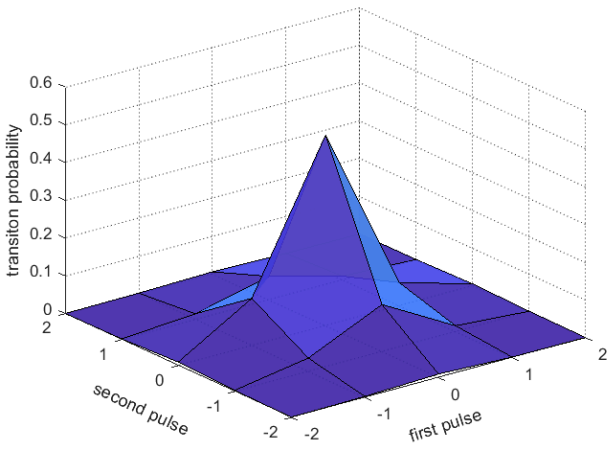

Fig. 5. Markov transition probability in Opus pulse domain.

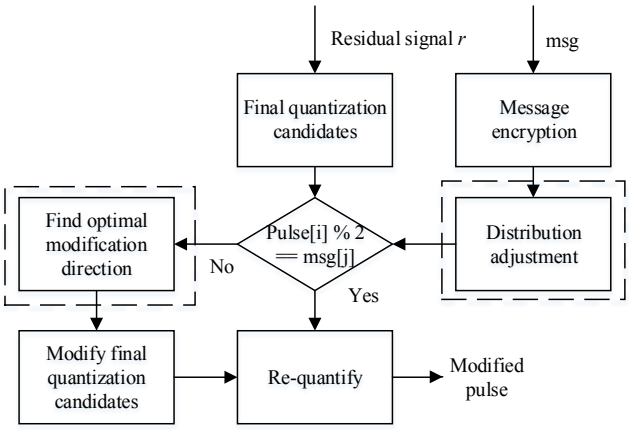

Fig. 6. The embedding process of the proposed scheme.

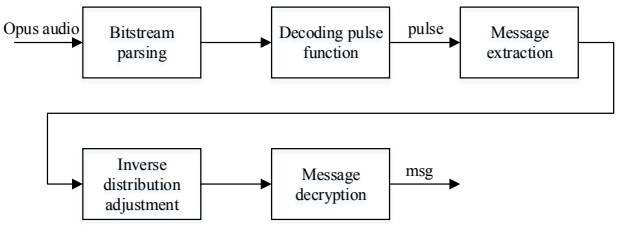

Fig. 7. The extracting process of the proposed scheme

lies in the dependency between pulse samples. Besides, the maintenance of statistical features is still a difficult job in designing secure steganographic algorithms for Opus pulse domain.

III. A SECURE STEGANOGRAPHIC SCHEME FOR OPUS PULSE

The aim of a secure steganography algorithm is to make stego and cover consistent in distribution. Actually, it's hard to achieve this as modifications to cover usually alters various statistical properties. However, by reducing the number of modifications, changes of those statistical features are less likely to arouse suspicion [16]. Based on this idea, this paper proposes a steganographic scheme based on message transform to reduce the number of modifications while maintaining the same embedding capaccity. Besides, in order to ensure the concealment of the proposed scheme, a rate-distortion measurement scheme is adopted to optimize the modification direction so that the impact on auditory quality is reduced.

This section first gives the overall framework of the proposed scheme, and then describes the core modules in detail.

A. The Embedding and Extracting Framework

Fig. 6 and Fig. 7 show the embedding and extracting process of the proposed scheme, respectively. In the embedding module, the parity of pulse value is used to hide secret information, with even values representing bit 0 and odd values representing bit 1. By modifying the final quantization candidate value, the parity of pulse is changed indirectly, thus the embedding of secret information is completed. As shown in Fig. 6, the core modules of the embedding process are marked with dotted box, that is, message transform and modification direction optimization. The main purpose of message transform is to reduce the number of modifications in embedding process by adjusting the distribution of secret message bitstream. In addition, in order to avoid introducing too much noise and reducing the coding efficiency and quality, the same evaluation criteria of the optimal quantization candidates are used to optimize the modification direction. In the extracting module, through the function of decoding quantization indices of excitation, pulse samples can be extracted exactly. Then, the embedded bitstream is obtained according to the parity of pulse values. Finally, after the inverse operation of distribution adjustment and decryption, the secret message is extracted correctly.

B. Message Transform

Usually, when steganography is used in practice, secret message is compressed and encrypted before the embedding operation. Therefore, it can be considered that the proportion of bit 0 and bit 1 in the embedded bitstream is 0.5 respectively and they are randomly distributed. However, in Opus pulse domain, as seen in Fig. 4, the ratio of even pulse samples to odd pulse samples is roughly 7:3. If encrypted message is directly embedded into the pulse samples, it is easy to know that the number of pulse samples which need to be modified is about half the number of bits in embedded bitstream. But if the distribution of bit 0 and bit 1 in secret message is adjusted to make the proportion of bit 0 larger, intuitively, the number of pulse samples that need to be modified will be reduced. Let the proportions of bit 0 and bit 1 in secret message be p_{b0} and p_{b1} respectively, and for pulse samples, the proportions of even and odd are represented as p_{even} and p_{odd}. Then, if the bitstream and pulse samples are randomly distributed

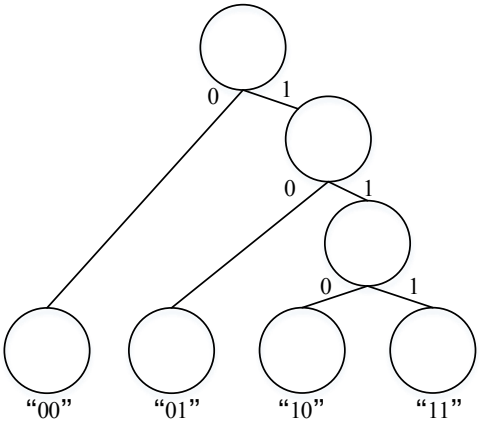

Fig. 8. A binary tree to construct prefix codes. "00", "01", "10" and "11" represent four symbols in the bitstream of message.

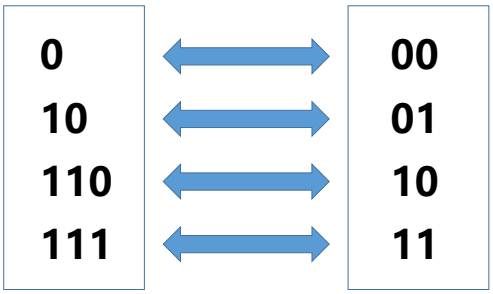

Fig. 9. The mapping table in message transform module. In the embedding process, symbols in the bitstream are transformed according to the mapping relationship from left to right.The extraction process is reversed.

according to above probability, , the modification rate m_{rate} in the embedding process is given by:

$$m_{rate} = p_{b0} * p_{odd} + p_{b1} * p_{even} \tag{2}$$

In this section, we tries to find a way to reduce m_{rate} under the constraints of p_{b0}+p_{b1}=*1* and p_{even}+p_{odd}=*1*. When p_{even} is roughly set to 0.7, Equation 2 is simplified to a decreasing function with p_{b0} as the independent variable. Therefore, in order to reduce m_{rate}, the distribution of secret message should be adjusted to make p_{b0} as large as possible. Through some kind of message transform, p_{b0} can be increased at the cost of adding extra bits to the message bitstream. Let the number of message bits before and after message transform be $msglen_{before}$ and $msglen_{after}$ respectively. As long as $msglen_{after}$ * m_{rate} <$msglen_{before}$ * *0.5*, the total modifications in the embedding process are reduced. Based on this idea, the proposed scheme adopts binary prefix codes to perform message transform.

Prefix codes [17] means that when encoding a character set, it requires the code of any character in the character set is not the prefix of the codes of other characters. Prefix code can be obtained from a binary tree. For example, in Fig. 8, set the left child nodes represent code 0 and the right child nodes represent code 1. Then, the codes on the path from the root node to the leaf node are the prefix codes for the leaf node. A set of prefix codes (0, 10, 110, 111) can be obtained from Fig. 8.

In the bitstream of secret message, we can divide the bits sequence into four symbols: {"00", "01", "10", "11"}. Through the prefix codes in Fig. 8, a mapping relationship can be obtained, as shown in Fig. 9. As the probabilities of those symbols in message bitstream are different, replacing the bits sequence with the corresponding symbols in the mapping table will change the proportions of bit 0 and bit 1.

In the message transform module, the core part is the design of the mapping relationship. To increase the proportion of bit 0, symbols with high probability should be mapped to those where bit 0 accounts for higher proportions. However, it's worth noting that the number of bits will be increased slightly after the replacement. Therefore, the mapping relationship must ensure that $msglen_{after}$ * m_{rate} <$msglen_{before}$ * *0.5*.

C. Rate-distortion Evaluation Scheme

In the voice mode of Opus, the quantizer of the residual signal is a trellis quantizer, implemented as a uniform scalar quantizer with a variable offset. The rate-distortion evaluation scheme is used to find the optimal quantization level of each residual signal by minimizing the rate and distortion. The number of candidates determine the amount of computation. The following is a brief introduction to the determination of quantization candidates and the measurement of rate-distortion, ignoring the tedious details.

In the NSQ module, there are two quantization level candidates named as *q1* and *q2*. The value of *q1* is calculated by substracting an offset from the residual *r* and then is fine-tuned by a series of related parameters. As for *q2*, it is obtained according to the value of *q1*. In general, the bias between *q1* and *q2* is one pulse. The relationship of *q1*, *q2* and *r* is given by:

$$q1 < r < q2 \tag{3}$$

Through the measurement of rate-distortion, a better quantization candidate is selected. Equation 4 shows the general idea of rate-distortion measurement. The left part of Equation 4 represents the trend of the number of bits required for pulse samples. The larger the absolute value of the quantization candidate, the more bits are required for quantized value. λ represents the rate/distortion tradeoff, relating with speech activity, speeech signal type, quantization offset, the number of states for delayed decision and so on. The right part of Equation 4 shows the distortion cost. It is calculated by mean square error. By comparing rate-distortion value *rd*, the final quantization candidate is chosen with the lower *rd*. At high complexity setting, the *rd* value of different Viterbi states is accumulated and a Viterbi delayed decison mechanism is achieved to improve the performance of rate-distortion.

$$rd = \lambda * |q| + (r - q)^2 \tag{4}$$

In the proposed scheme, if the parity of pulse is different from the secret bit, the final quantization candidate is modified to make the quantized value +1 or -1. In order to minimize the impact on rate-distortion, there are two candidates for modification as well, representing two modification directions. Suppose v_i represents the actual pulse value at current sample, then the candidates for modification are obtained through the inverse quantization process of v_i+1 and v_i-1. Similarly, calculating the *rd* according to Equation 4 and selecting the one with lower *rd*. In the case of high complexity settings, as the final quantization candidate is not determined at current sample, all quantization candidates in each Viterbi state are modified to ensure the parity of the final quantized value consistent with the secret bit. The experimental result shows that this scheme is better than the way of randomly selecting a modification direction.

IV. Experiments

A. Experimental Database And Metrics

To evaluate the performance of the proposed steganographic scheme, 2000 wav audio samples are selected as cover media to conduct experiments. 1000 audios of them are downloaded from Internet, and another 1000 audios are recorded with CoolEdit. By removing 44 bytes from the header, 2000 PCM audio clips are obtained from those wav audio samples. And each PCM audio clip is characterized as 16KHz, 16-bit quantization and mono. The duration of most clips is about 30s. In addition, the content of those clips is varied, including different languages, such as English, Chinese and Russian.

In the experiments, the performance of the proposed scheme will be evaluated from the three aspects: embedding capacity, concealment and statistical security. The metrics for evaluation are as follows:

1) Embedding capacity represents the number of bits embedded in the audio carrier per second, the unit of which is kbps (kbits per second). The maximum embedding capacity reflects not only the size of the embedding domain, but also the embedding ability of the steganography algorithm.

2) concealment indicates the ability that the carrier which is embedded with secret message will not be perceived to be abnormal. For the audio carrier, the change of auditory quality before and after steganography reflects the concealment. In the experiments, the perceptual evaluation of speech quality (PESQ) [18] is used to measure auditory quality. It is based on auditory perception model to measure the similarity of two audio clips. The value range of PESQ is [1, 4.5]. The higher the PESQ value, the better the auditory quality, indicating that steganographic scheme has less damage to the auditory quality.

3) Statistical security refers to the ability of steganography algorithms to keep the statistical characteristics of cover unchanged. In general, the design of steganalysis algorithm is based on those statistical features. And the detectability of those steganalytic schemes reflects the statistical security.

B. Experimental Setup

To the best of our knowledge, there is still not any steganographic scheme or steganalytic scheme for Opus pulse domain. In order to verify the effectiveness of the algorithm, four groups of comparative schemes are designed to illustrate the effect of the core module. As shown in Fig. 6, the core modules of the proposed scheme are message transform module (MT-module) and the modification direction selection module (DS-module). Thus, by removing either or both of the two modules, three comparative schemes are obtained. For sake of simplicity, those schemes (MT+DS, MT, DS, NONE) are named after the core modules they contain, where "NONE" refers to the scheme containing neither MT-module nor DS-module. The above four schemes will be compared mainly from the aspects of concealment and statistical security.

The following experiments are conducted. First of all, in order to obtain the practical embedding capacity,, the average PESQ value of 2000 audio clips which are embedded with random message in the MT+DS scheme under different embedding rate are tested. Then, the concealment of the above four methods are compared by average PESQ values. Finally, the statistical security of the four schemes are compared by a two-class linear classifier and a two-class Gaussian kernel classifier.

In the parameter settings of Opus codec, the encoding mode is "voip" and the bitrate is 25000 bits per second. As for the complexity setting of Opus codec, due to the limitations in high complexity settings mentioned in Section II, the complexity is set to 0.

C. Experimental Results

1) embedding capacity analysis

As described in Section II, the number of pulse samples depends on sampling rate. In the experiments, the sampling rate of the audio clip is 16KHz. Thus, the maximum embedding capacity of the proposed scheme is 16 kbps. However, it is impractical to embed message with such a large embedding capacity, because there are a great deal of 0 values in pulse domain and too many modifications will ruin the process of NSQ and lead to a rapid decline in auditory quality.

Table I shows the average PESQ value under different embedding capacities. Considering that the auditory quality is very poor under high embedding capacities, the following experiments are carried out under the embedding capacity of no more than 8 kbps.

2) concealment analysis

In order to verify the role of the MT-module and DS-module in maintaining auditory quality, the average PESQ value of 2000 stego audio samples generated under different embedding capacities is obtained to compare the concealment of the four schemes. The results are shown in Table II. It can be seen that the scheme with both MT-module and DS-module performs best under all embedding capacities. To make the experimental results more intuitive, the results of 100 stego audio samples under the embedding capacity of 8 kbps are displayed in Fig. 10. It is obvious that the PESQ value of the stego audio

TABLE I
THE PESQ VALUE OF MT+DS UNDER DIFFERENT EMBEDDING CAPACITIES

Embedding Capacity(kbps)	0	1.6	3.2	4.8	6.4	8	9.6	11.2	12.8	14.4	16
Average PESQ	4.110	4.042	3.973	3.896	3.812	3.718	3.613	3.492	3.353	3.180	2.948

TABLE II
THE AVERAGE PESQ VALUE UNDER DIFFERENT EMBEDDING CAPACITIES

	1.6kbps	3.2kbps	4.8kbps	6.4kbps	8kbps
$MT+DS$	**4.042**	**3.973**	**3.896**	**3.812**	**3.718**
MT	3.98	3.831	3.655	3.444	3.193
DS	4.022	3.923	3.795	3.629	3.395
$NONE$	3.937	3.713	3.403	2.964	2.308

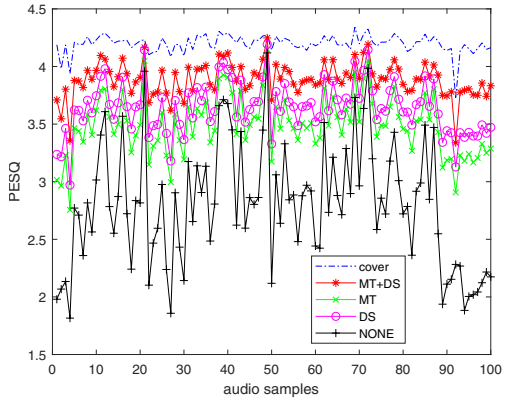

Fig. 10. The PESQ value of 100 cover audio samples and 100 stego audio samples generated by MT+DS, MT, DS, and None under the embedding capacity of 8 kbps.

samples produced by MT+DS scheme is higher than those of other schemes. In addition, as shown in Fig. 10, the DS scheme performs better than MT scheme in PESQ test and the NONE scheme performs worst. Therefore, through the comparative analysis of the experiments, it can be concluded that DS-module plays a major role in the maintenance of auditory quality, while MT-module also plays a secondary role because of reducing the modification rate.

3) statistical security analysis

In this experiment, the histogram features shown in Fig. 4 of the above schemes under different embedding capacities are used to evaluate the ability of maintaining statistical characteristics of the MT-module and DS-module. After normalization, those features are used to train a linear classification model and a Gaussian kernel classification model. Then, the accuracy of the two classification models in the test set can be used to evaluate statistical security of the four schemes. Each group of the experiments includes 4000 samples (2000 cover audio samples and 2000 stego audio samples). And the accuracy is calculated by a 10-fold cross-validation method. The higher the accuracy, the weaker the anti-detection ability of the steganographic scheme.

Table III shows the results of the linear classification module

TABLE III
THE ACCURACY OF THE LINEAR CLASSIFICATION MODULE UNDER DIFFERENT EMBEDDING CAPACITIES

	1.6kbps	3.2kbps	4.8kbps	6.4kbps	8kbps
$MT+DS$	0.488	**0.4928**	**0.5088**	**0.5085**	**0.5113**
MT	**0.4825**	0.527	0.835	0.9115	0.936
DS	0.518	0.8482	0.9485	0.9798	0.9945
$NONE$	0.5068	0.5683	0.839	0.9928	1

TABLE IV
THE ACCURACY OF THE GAUSSIAN KERNEL CLASSIFICATION MODULE UNDER DIFFERENT EMBEDDING CAPACITIES

	1.6kbps	3.2kbps	4.8kbps	6.4kbps	8kbps
$MT+DS$	**0.5045**	**0.4987**	**0.5087**	**0.5157**	**0.5308**
MT	0.5065	0.7561	0.8831	0.9253	0.9504
DS	0.6087	0.8984	0.9557	0.9815	0.995
$NONE$	0.5249	0.7936	0.9575	0.9914	0.9998

and Table IV shows the results of the Gaussian kernel classification module. As can be seen in Table III and Table IV, neither linear classification model nor nonlinear classification model can accurately classify cover audio samples and stego audio samples generated by the MT+DS scheme. Through further comparison, it is found that the classification accuracy of DS scheme under the embedding capacity of no more than 3.2 kbps is the highest, which means that the DS-module is not conducive to the maintenance of statistical features when it is not combined with the MT-module. Besides, the classification accuracy of MT scheme is lower than either that of DS scheme and NONE scheme, which proves that MT-module also plays a role in maintaining statistical characteristics.

In order to intuitively illustrate the ability of maintaining statistical characteristics, Fig. 11 shows the changes of histogram statistical characteristics under different embedding capacities. It can be seen that with the increase of embedding capacity, the histogram feature changes more. Besides, the histogram feature of the stego samples produced by DS scheme changes most. In general, the trends reflected in Fig. 11 are consistent with the conclusions drawn from Table III and Table IV. The more the histogram feature changes, the higher the classification accuracy of the linear model and Gaussian kernel model.

Through the analysis of the above comparative experiments, it can be proved that the core modules (MT-module and DS-module) of the proposed scheme are effective. Between them, the main function of MT-module is to reduce the modification rate and maintain statistical characteristics; the role of DS-module is to reduce the impact of steganographic modifications on auditory quality and improve the concealment of steganographpy algorithm. And the combination of MT-module and DS-module performs best in concealment and

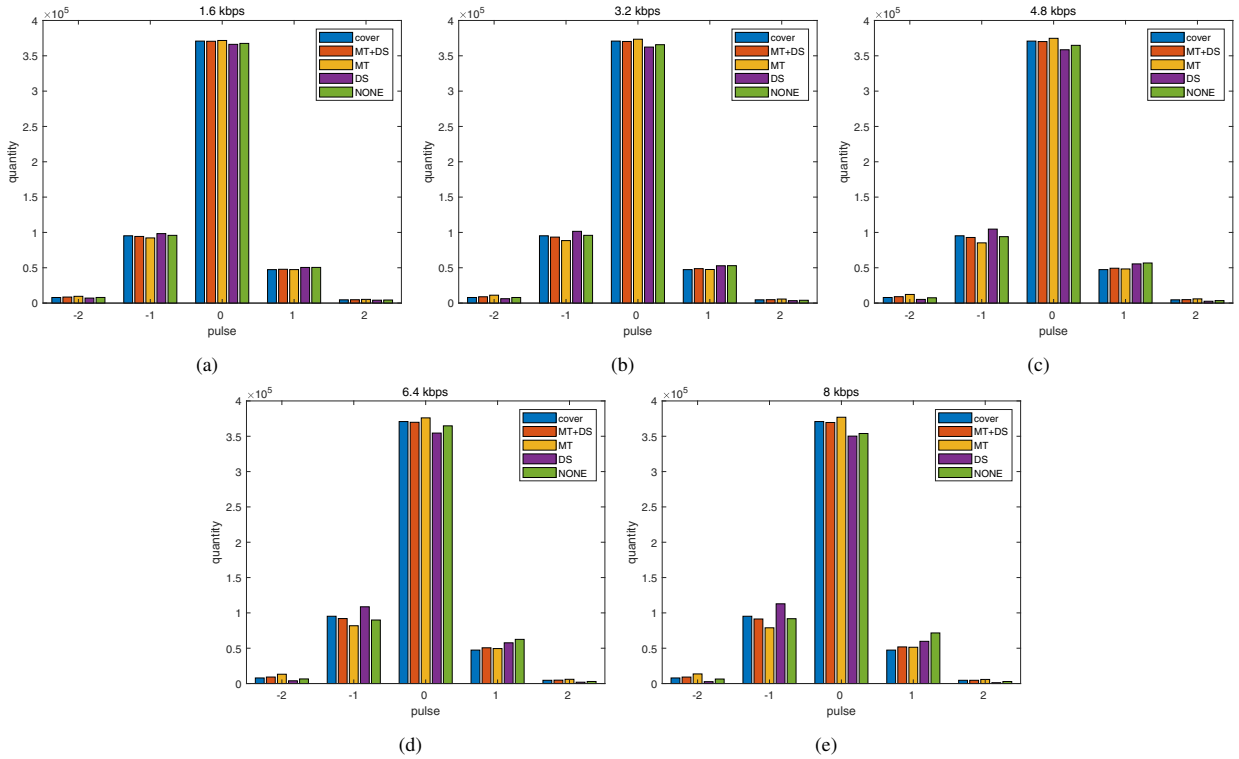

Fig. 11. The comparison of the histogram of the audio samples produced by MT+DS, MT, DS and NONE scheme under different embedding capacities.

statistical security.

V. Conclusions and Future Work

A secure steganographic scheme in Opus pulse domain is proposed in this paper. As Opus pulse domain takes a huge advantage in maximum embedding capacity compared to the embedding domain of other speech codec, it is an ideal and practical embedding domain for audio steganography. However, through the analysis of the pulse samples, it is found that there exists huge limitations in Opus pulse domain for the dependency between pulse samples. This paper puts forward a preliminary solution to this problem by message transform and finding the optimal modification direction. Besides, the comparative experiments show that the proposed scheme is effective and can maintain good auditory quality at a high embedding capacity. Meanwhile, the histogram characteristic can also be well maintained. In terms of embedding capacity, the proposed scheme can achieve 8 kbps at the sampling rate of 16 KHz while ensuring concealment and security to some degree.

As Opus pulse domain is a promising embedding domain for audio steganography, we will do a deeper research on the analysis of pulse domain and propose a better scheme to solve the matter of interdependence between pulse samples.

References

[1] M. Hussain, A. W. A. Wahab, Y. I. B. Idris, A. T. Ho, and K.-H. Jung, "Image steganography in spatial domain: A survey," *Signal Processing: Image Communication*, vol. 65, pp. 46–66, 2018.

[2] Z. Wei, B. Zhao, B. Liu, J. Su, L. Xu, and E. Xu, "A novel steganography approach for voice over ip," *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, no. 4, pp. 601–610, 2014.

[3] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevnỳ, "Moving steganography and steganalysis from the laboratory into the real world," in *Proceedings of the first ACM workshop on Information hiding and multimedia security*, 2013, pp. 45–58.

[4] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the opus audio codec," *IETF, September*, 2012.

[5] K. Vos, S. Jensen, and K. Soerensen, "Silk speech codec," *IETF draft*, 2010.

[6] J.-M. Valin, G. Maxwell, and T. B. Terriberry, "Celt: A low-latency, high-quality audio codec," *The Xiph. Org Foundation*, 2011.

[7] C. Gong, X. Yi, and X. Zhao, "Pitch delay based adaptive steganography for amr speech stream," in *International Workshop on Digital Watermarking*. Springer, 2018, pp. 275–289.

[8] Y. Ren, D. Liu, J. Yang, and L. Wang, "An amr adaptive steganographic scheme based on the pitch delay of unvoiced speech," *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 8091–8111, 2019.

[9] P. Liu, S. Li, and H. Wang, "Steganography integrated into linear predictive coding for low bit-rate speech codec," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2837–2859, 2017.

[10] Y. Ren, W. Zheng, and L. Wang, "Silk steganography scheme based on the distribution of lsf parameter," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 539–548.

[11] H. Miao, L. Huang, Z. Chen, W. Yang, and A. Al-Hawbani, "A new scheme for covert communication via 3g encoded speech," *Computers & Electrical Engineering*, vol. 38, no. 6, pp. 1490–1501, 2012.

[12] Y. Ren, H. Wu, and L. Wang, "An amr adaptive steganography algorithm based on minimizing distortion," *Multimedia Tools and Applications*, vol. 77, no. 10, pp. 12 095–12 110, 2018.

[13] K. Vos, K. V. Sørensen, S. S. Jensen, and J.-M. Valin, "Voice coding with opus," in *Audio Engineering Society Convention 135*. Audio Engineering Society, 2013.

[14] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion

in steganography using syndrome-trellis codes," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 920–935, 2011.

[15] E. Ayanoglu and R. Gray, "The design of predictive trellis waveform coders using the generalized lloyd algorithm," *IEEE transactions on communications*, vol. 34, no. 11, pp. 1073–1080, 1986.

[16] P. Sallee, "Model-based steganography," in *International workshop on digital watermarking*. Springer, 2003, pp. 154–167.

[17] D. S. Hirschberg and D. A. Lelewer, "Efficient decoding of prefix codes," *Communications of the ACM*, vol. 33, no. 4, pp. 449–459, 1990.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.