International Journal of Digital Crime and Forensics
Volume 11 • Issue 4 • October-December 2019

Steganalysis of AMR Based on Statistical Features of Pitch Delay

Yanpeng Wu, Xiamen Meiya Pico Information Co., Ltd., Xiamen, China
Huiji Zhang, Xiamen Meiya Pico Information Co., Ltd., Xiamen, China
Yi Sun, Xiamen Meiya Pico Information Co., Ltd., Xiamen, China
Minghui Chen, Xiamen Meiya Pico Information Co., Ltd., Xiamen, China

ABSTRACT

The calibrated matrix of the second-order difference of the pitch delay (C-MSDPD) feature has been proven to be effective in detecting steganography based on pitch delay. In this article, a new steganalysis scheme based on multiple statistical features of pitch delay is present. Analyzing the principle of the adaptive multi-rate (AMR) codec, the pitch delay values in the same frame is divided into groups, in each of which, a pitch delay has a closer correlation with the other ones. To depict the characteristic of the pitch delay, two new types of statistical features are adopted in this article. The new features and C-MSDPD feature are together employed to train a classifier based on support vector machine (SVM). The experimental result shows that, the proposed scheme outperforms the existing one at different embedding bit rates and with different speech lengths.

KEYWORDS

Adaptive Multi-Rate Speech, C-MSDPD, C-PDDPD, Markov Transition Probability, MDPD, Pitch Delay, Speech Steganalysis, Speech Steganography, Support Vector Machine

1. INTRODUCTION

Steganography is a security technique that utilizes digital files or network protocols to embed secret messages (Provos & Honeyman, 2003). Compared with traditional security technology, steganography has the advantage of concealment, which will make it undetectable for attackers. Accordingly, steganography can be applied to covert communication.

The research of steganography is mainly concentrated in images. Content-adaptive steganographic methods are the most secure schemes in recent years. Compare with traditional steganographic methods, content-adaptive steganographic methods can provide better security to resist the statistical detection. Filler, Judas and Fridrich (2010) developed a framework with Syndrome-Trellis Codes (STCs), which could be used for minimizing additive distortion between cover and stego images. There are many algorithms implemented by STCs, such as highly undetectable stego (HUGO) method (Bas, 2010), spatial-universal wavelet relative distortion (S-UNIWARD) method (Holub and Fridrich, 2013) et al. To enhance the security of covert communication, Sedighi, Cogranne and Fridrich (2016) proposed a method by using an estimated multivariate Gaussian cover image model to minimize the statistical detect ability. Content-adaptive image steganographic methods increase the difficulty of detection, but steganalysis technologies also make some progress in these years.

DOI: 10.4018/IJDCF.2019100105

Rich-model based steganalysis is the modern methods for stego images detection. Fridrich and Kodovsky (2012) first design a rich-model based steganalysis method for images steganography. In their method, high dimensional features and ensemble classifier are employed to enhance the detection accuracy. Then Goljan, Fridrich and Cogranne (2014) designed an extension of the spatial rich model for color images. To detect the content-adaptive image steganographic methods, Denemark, Boroumand and Fridrich (2016) design some high order features by the knowledge of the selection channel. Luo et al. (2016) analysis the character of STCs and designed a steganalysis method for HUGO steganography. The method can not only detect the stego images but also extract the secret messages. Recently, Liu, Yang and Kang (2017) proposed a steganalysis method combines convolutional neural network with rich-models and ensemble classifiers. Experimental results show that the method has better performance than the state-of-the-art one. However, due to the structure and character differences between the parameters of image and speech, it is hard to directly employ the steganalysis methods on image to achieve effective detection for speech steganography.

In recent years, with the development of mobile network and smart phone, Voice over IP (VoIP) has become widely employed by mobile communication such as network telephone or instant message. Compared with other carriers for covert communications, VoIP has obvious advantages, for example, its large volume for embedding data could provide high covert bandwidth, and its instantaneity could provide real-time communication environment. Therefore, there are many works have been done for the steganography based on VoIP. As a standard of speech compression, AMR is widely employed by 3G, 4G systems or VoIP in speech services. Due to its great performance on speech compression, AMR is adopted as the file format for many communication applications such as intent message or speech recorder on smart phones. Therefore, the steganography of AMR speech codec has attracted extensive attention in recent years.

In general, the steganography based on VoIP can be divided into two classes (Mazurczyk, 2013): The first one carries out information hiding by modifying the protocol of VoIP (Huang, Yuan, Chen & Xiao, 2011; Jiang, Tang, Zhang, Xiong & Yip, 2016; Mazurczyk & Lubacz, 2010), the other one embeds secret messages by manipulating the parameters of speech codec during or after encoding process. The steganography based on parameter modification is the most common approach for covert communication based on VoIP. Because of the redundancy of compress speech, slightly change of parameter would not affect the speech quality obviously. Wu and Yang (2006) found that fixed codebook indices are ideally suitable for embedding secret message. They proposed an approach of steganography based on Analysis-by-Synthesis (ABS) by modifying the fixed codebook index parameter in the course of encoding process. To enhance the security of steganography, Wu, Cao, and Li (2015) adopted matrix coding in the modification of fixed codebook index afterward. Geiser and Vary (2008) proposed a steganography method based on an alternative search strategy of the fixed codebook. The experimental results demonstrate that the method causes a negligible effect on the subjective quality of the coded speech with high speed of secrete messages transmission. Miao et al. (2012) also choose fixed codebook indices as the carrier to embed messages, and they used an embedded factor to control the embedding capacity. The method can embed message with both high capacity and low speech distortion by adjusting the factor during the process of speech encoding.

Besides the fixed codebook indices, Liner Prediction Coefficient (LPC) is another feasible domain for steganography. Xiao, Huang, and Tang (2008) utilized an algorithm called complementary neighbor vertices (CNV) to divide the codebook into two parts. Quantization index modulation (QIM) is employed to embed bits into LPC during codebook searching. To enhance the security of QIM, Tian, Liu and Li (2014) introduced a novel steganographic method based on random position selection and matrix encoding strategy. The experimental results show that the approach has greater steganalysis resistance than the Xiao's one (Xiao et al. 2008). Liu, Tian, Lu and Chen (2015) utilized the matrix embedding strategy to hide secret information during the linear predictive coding process. The method has better performance for resist steganalysis and lower speech distortion.

Pitch delay is one of the most important parameters in speech codec. Some researchers have pointed out that the pitch period could not be accurately predicted (Hess & Shaughnessy, 1984). Thus, as the predict value of pitch period, pitch delay has the characteristic of considerable redundancy, which means it is a feasible position for steganography. Huang, Liu, Tang and Bai (2012) implemented information hiding by dividing search range of the pitch delay into two parts. One part is used only when the embedded secret message bit is 1 and the other one is employed while the bit is 0. During the speech encoding, the search range of pitch delay in each subframe is changed according to the value of each bit of secret message. Yan, Tang and Sun (2015) found that pitch delay parameters in second and fourth subframe of G.723.1 can achieve better performance than the other ones for steganography. They proposed an algorithm that embeds bits by using a double layer steganography method to reduce the distortion of speech.

Steganography can enhance the security of information transmission, but at the same time, it could be exploited to be engaged in illegal activities such as terrorist attacks or other crimes. Therefore, the countermeasure technology of steganography, steganalysis has become a hot research area in the late years. To detect the steganography based on fixed codebook indices, Ding and Ping (2010) extracted multiple features from the fixed codebook indices. SVM was employed as the classifier to distinguish the stego speech from the cover ones. The method can detect the steganographic speech produced by Wu's (2006) scheme, but it is hard to correctly classify stego speech generated by Geiser's (2008) method. Addition, Miao, Huang, Shen, Lu and Chen (2013) employed Markov transition matrix and two types of entropy as features to detect the steganography based on fixed codebook indices. The steganalysis approach can detect not only Wu's (2006) steganography method but also Geiser's (2008) one. To achieve higher accuracy, Ren, Cai, Tang and Wang (2015) further presented a steganalysis algorithm which employs the probabilities of the same pulse positions as features. Miao's (2013) and Ren's (2015) approach can detect the both steganography methods based on the fixed codebook indices. But their features are not enough to characterize the fixed codebook indices. Tian et al. (2017) introduced a novel steganalysis method based on more complete features. The experimental results show that the detection accuracy of Tian's (2017) method is higher than both Miao's (2013) and Ren's (2015) ones at any embedding bit rate or with any sample length.

To detect the steganography of LPC, Li, Tao, and Huang (2012) developed a steganalysis approach based on the analysis of quantization index sequence. With the feature of index distribution characteristics (IDC), the method can detect CNV-based steganography in some cases. As mentioned before, Tian et al. (2014) and Liu et al. (2015) presented a more secure steganographic method and make the IDC-based steganalysis inefficient. Because of that, recently, Li, Jia and Kuo (2017) presented a steganalysis based on a codeword correlation network which is constructed by splitting vector quantization codeword from adjacent speech frames. From the experimental results it can be seen that the steganalysis method can detect the steganography based on CNV-QIM with high accuracy even if the matrix embedding strategy is involved.

To detect the steganography based on pitch delay, some steganalysis schemes have been proposed in recent years. Li, Jia, Fu and Dai (2014) developed an algorithm based on codeword correlation network. The experimental results illustrate that the method can detect Huang's (2012) steganography method with high accuracy. It has been proven that this idea also works well on steganalysis of CNV-QIM afterwards (Li et al., 2017). Then Ren, Yang, J. Wang and L. Wang, (2017) found that the second-order difference of pitch delay could be used for steganalysis due to the stability of pitch delay would be affected by information hiding. They proposed a steganalysis scheme based on C-MSDPD feature. The experimental results show that Ren's (2017) approach outperforms Li's (2014) one in terms of correct detection rate at various embedding bit rates. Especially when the embedding bit rate is low, Ren's (2017) method has obvious advantages.

As a matter of fact, C-MSDPD is not enough for characterizing the steganography of short length speech. By analyzing the principle of adaptive codebook search, a new steganalysis scheme of the pitch delay is present. In this paper, we present two new types of features to describe the difference

between cover speech samples and stego ones. The two types of features and C-MSDPD feature are together employed for training a SVM-based classifier. The experimental results show that the proposed scheme can provide better detection performance than Ren's (2017) one at different embedding bit rates and with different speech length.

The rest of this paper is organized as follows. Section 2 describes the principle of adaptive codebook search at first, and then the steganalysis and steganography schemes of pitch delay are reviewed in detail. After the analysis of the characteristics of pitch delay, Section 3 introduces the features proposed in this paper and describes the proposed SVM-based steganalysis scheme in details. Section 4 shows the experimental evaluation and their results. The conclusion is given in Section 5 at last.

2. BACKGROUND AND RELATED WORK

In this section, the principle of adaptive codebook search is introduced first, and then the steganography approaches and the state-of-the-art steganalysis method based on pitch delay will be reviewed in detail.

2.1. Principle of AMR Codec

AMR is an audio compress format based on the code excited linear predictive (CELP) coding model (Ekudden, Hagen, Johansson and Svedberg, 1999). Figure 1 shows the principle of AMR encoding algorithm. In AMR encoder, each frame of 20 ms is divided into 4 subframes. Synthetic speech is constructed by feeding the excitation vectors chosen from adaptive and fixed codebook through a linear prediction synthesis filter (3GPP/ETSI, 2016). To choose the optimum vectors from the codebook, an algorithm named ABS is employed for adaptive and fixed codebook search. The main idea of ABS is to minimize the mean square error between the original and synthesized speech.

Pitch delay is an important parameter of AMR codec. As the predict value of pitch period, pitch delay is searched in adaptive codebook. The adaptive codebook search is implemented based on subframes and the search procedure mainly consists of two parts, open-loop pitch analysis and closed-loop pitch analysis. First, open-loop pitch analysis is employed to obtain open-loop estimated lags and the estimated lags will be utilized to control the search range of pitch delay. Then closed-loop pitch analysis is implemented to get the pitch delay and gains by ABS.

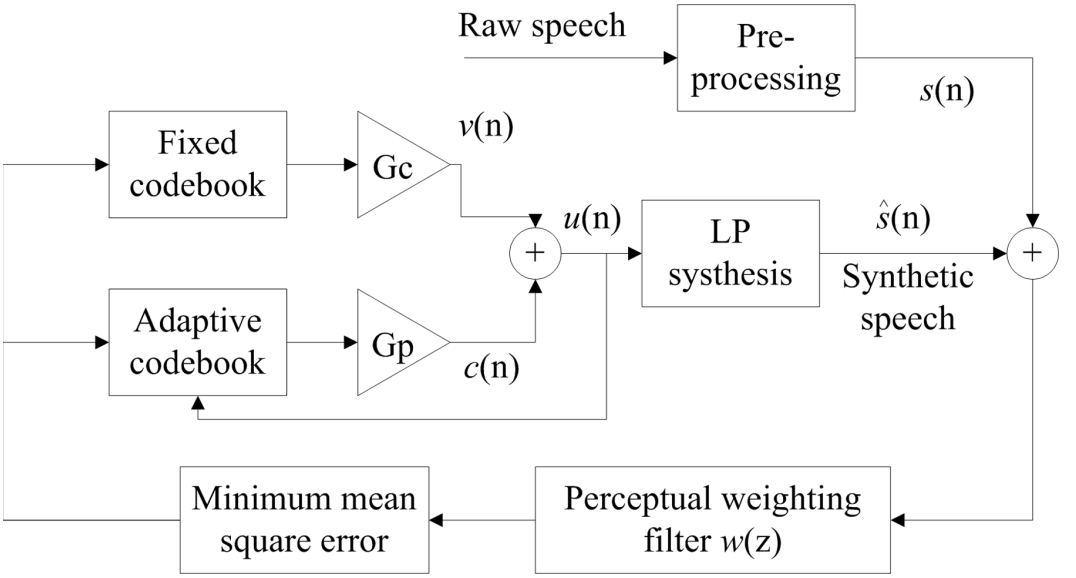
To introduce the principle of adaptive codebook search and the steganography method clearly, AMR-NB 12.2kb/s mode is taken as an example of the following text. At 12.2kb/s mode, open-loop pitch analysis is performed twice per frame which means there are two open-loop estimated lags will be obtained in each frame. One lag is applied for pitch delay search in the first subframe and the other one is employed for the search procedure in the third subframe. Pitch delay has two parts, integer pitch delay and fractional pitch delay. In the first (or third) subframe, the integer pitch delay is searched around the first (or second) open-loop estimated lag. The search range of the first integer pitch delay is determined as follows:

$$p_{0,i} = \begin{cases} \left[18,\ 24\right], & T_{0,i} < 21 \\ \left[T_{0,i} - 3, T_{0,i} + 3\right], & 21 \le T_{0,i} \le 140 \\ \left[137,143\right], & T_{0,i} > 140 \end{cases} \tag{1}$$

where $p_{0,\,i}$ is the integer pitch in the first subframe of i-th frame, and $T_{0,\,i}$ is the open-loop estimated lag for the first subframe.

Different to the first and third subframe, the search range of the integer pitch in the second and last subframe is determined by the integer pitch in the previous subframe. For the second integer pitch in the i-th frame $p_{1,\,i}$, the search range is shown as below:

Figure 1. Principle of AMR encoding algorithm



$$p_{1,i} = \begin{cases} \left[18,\ 27\right], & p_{0,i} < 23 \\ \left[p_{0,i} - 5, p_{0,i} + 4\right], & 23 \le p_{0,i} \le 139 \\ \left[134,\ 143\right], & p_{0,i} > 139 \end{cases} \tag{2}$$

The search is implemented by minimizing the mean square error between the original speech and synthesized speech. After the integer part is determined, the fractional part will be searched in the similar way in each frame. It is obvious that the search range of the optimum integer pitch delay in each subframe is different in most cases. Because of the particular determination strategy of the search range, pitch delay in the second and fourth subframe will have closer correlation with the previous one than the later one.

2.2. Principle of Steganography Based on Pitch Delay

From here we can see that the pitch delay obtained by closed-loop pitch analysis is suboptimal since the search range is only a small part of all potential values. Besides, it has been proven that it is hard to correctly predict pitch period (Hess & Shaughnessy 1984). Therefore, many scholars try to hide information into AMR codec by modifying the pitch delay. Adjusting the search range of pitch delay is a common way to embed secret messages.

In Huang's (2012) approach, each subframe can embed one bit. They divided the search range of pitch delay into two parts. One part only consists of even values and the other one only contains odd values. If the secret message bit is 0, only even values will be searched in closed-loop pitch analysis. On the contrary, only odd integer pitch delay will be chosen while the secret message bit is 1. Thus, the decoder can extract the secret message bit by using this formula:

$$\begin{aligned} b_{i+j\times 4} &= 0, \quad \text{if } \mathrm{mod}(p_{i,j}, 2) = 0 \\ b_{i+j\times 4} &= 1, \quad \text{if } \mathrm{mod}(p_{i,j}, 2) = 1 \end{aligned} \tag{3}$$

Where $p_{i,j}$ is the integer pitch delay of the i-th subframe in the j-th frame, and the secret message bit embedded in $p_{i,j}$ is $b_{i+j\times4}$.

Yan et al. (2015) found that the pitch delay in second and fourth subframe is more suitable for steganography. To reduce the distortion of speech and achieve high covert bandwidth, they designed a double layer steganography algorithm. In the first layer, the embedding method is similar with Huang's (2012) approach. The search range of pitch delay in the second and fourth subframe is adjusted to embed secret message bits. The extraction strategy of secret message bits is also based on (3). In order to embed more bits into the second and fourth subframes, the second layer steganography is implemented based on:

$$b_{2+k\times3} = \left\lfloor \frac{\mathrm{mod}\left(p_{1,k},\ 4\right)}{2} \right\rfloor \oplus \left\lfloor \frac{\mathrm{mod}\left(p_{3,k},\ 4\right)}{2} \right\rfloor \tag{4}$$

Where $b_{2+k\times3}$ is the third bit embedded in the k-th frame while $p_{1,k}$ and $p_{3,k}$ are the values of the pitch delay in the second and fourth subframe of the k-th frame respectively.

By using the double layer steganography, they can embed more bits into pitch delay without much distortion of speech. As a result of that, the covert communication based on the double layer steganography will be safer.

2.3. Principle of Steganalysis Based on Pitch Delay

Ren et al. (2017) found that steganography will make the pitch delay sequences of stego speech less stable than it of cover speech. They performed an experiment to compare the difference between the pitch delay of cover speech and stego speech. The experimental result shows that the second order difference of pitch delay is suitable to be the feature of steganalysis. Moreover, they designed a matrix called Markov transition probability matrix of the second-order difference of pitch delay (MSDPD) to show the difference between the original and steganographic speech. The matrix M_D2 is calculated by:

$$M_{D^2}(i,j) = \frac{\sum_{t=0}^{N-4} P\left(D^2(t) = i, D^2\left(t+1\right) = j\right)}{\sum_{t=0}^{N-4} P\left(D^2(t) = i\right)} \tag{5}$$

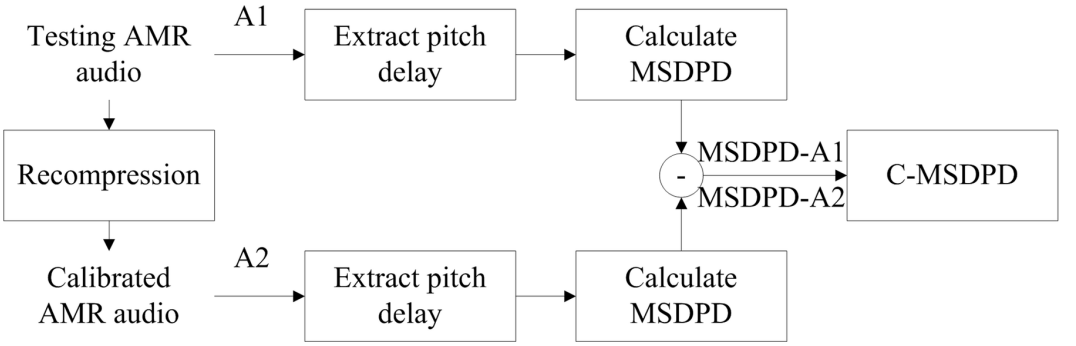
where $D^2(t)$ is the second order difference of pitch delay in the t-th subframe. $P(D^2(t) = i)$ is the probability that $D^2(t) = i$. $P(D^2(t) = i, D^2(t+1) = j)$ is the probability that $D^2(t) = i$ and $D^2(t+1) = j$ at the same time. $M_D2(i, j)$ is the transition probability that the second order difference of pitch delay in current subframe is j while the value in previous subframe is i.

To improve detection accuracy and reduce computational complexity, they set the threshold of $D^2(t)$ to [-6, 6]. Thus, the dimension of M_D2 in 12.2kb/s mode is 169. In order to further improve the performance of the classifier, they applied a calibration method to estimate the feature of cover speech. The feature extraction process is shown on Figure 2 where C-MSDPD is the feature sent into SVM for training. The calculation method of C-MSDPD is

$$\text{C-MSDPD} = \text{MSDPD-A1} - \text{MSDPD-A2} \tag{6}$$

where MSDPD-A1 is the MSDPD calculated by the tested speech, and MSDPD-A2 is the MSDPD extracted from the recompressed speech.

Figure 2. The extraction process Of C-MSDPD features



3. STEGANALYSIS SCHEME BASED ON STATISTICAL FEATURES OF PITCH DELAY

Although C-MSDPD has great performance on detecting steganography based on pitch delay, there is still room for improvement. In this section, we will introduce two new types of features for steganalysis of pitch delay. Based on the principle of adaptive codebook search, the pitch delay in the same frame is divided into some groups. The first feature is calibrated probability distributions of the difference of pitch delay in the same group (C-PDDPD). The other one is Markov transition probability matrix of the difference of pitch delay in same group (MDPD). The new features and C-MSDPD feature will be extracted from speech samples and sent together into SVM for training and test.

3.1. Calibrated Probability Distributions of The Difference Between Pitch Delay in The Same Group (C-PDDPD)

As mentioned before, the search range of each integer pitch delay is different in most cases. The search ranges of the first and third integer pitch delay are determined by the open-loop estimated lags while the second and last search range for optimum integer pitch are rely on the value of previous pitch delay. Thus it can be seen that, the correlation between the first and second pitch delay is stronger than it between the second and third pitch delay. Based on the above, we divide the four pitch delay in the same frame into two groups. One group consists of the first two pitch delay and the other one is composed of the last two pitch delay.

Short-term invariance is an important characteristic of speech signals, which means the signal of speech should be stable in short time duration. Because the duration of voiced phoneme is 30-50ms (Yan et al., 2015) and the length of each frame in AMR codec is only 20ms, the signal in the same frame could be seen as stable. The pitch delay sequence in the same frame should also be stable. The main idea of steganography based on pitch delay is to embed secret messages by adjusting the search range, thus the steganographic behavior will have a huge impact on the short-term invariance of pitch delay sequences, especially for the pitch delay in the same group.

To depict the impact, we adopt the probability distributions of the difference between pitch delay in the same group. Assume that the speech segment consists of N frames, there are $2N$ difference values can be obtained by the speech segment. For the k-th frame of the speech, the two difference values can be calculated as follows:

$$\begin{aligned} d_{0,k} &= p_{0,k} - p_{1,k} \\ d_{1,k} &= p_{2,k} - p_{3,k} \end{aligned} \tag{7}$$

where $d_{0,k}$ is the difference value between the first two pitch delay in the k-th frame and $d_{1,k}$ is the one between the last two pitch delay. The probability of the difference between pitch delay in the same group can be calculated as:

$$P(d) = \frac{\sum_{k=0}^{N-1} \delta(d_{0,k} = d) + \sum_{k=0}^{N-1} \delta(d_{1,k} = d)}{2N} \tag{8}$$

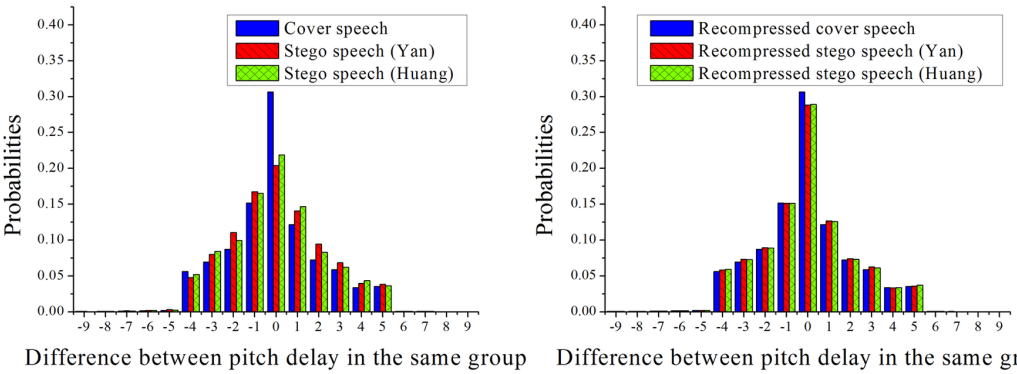
where $P(d)$ is the probability that the difference value between pitch delay in the same group is d. In 12.2kb/s mode, the range of d is [-9, 9]. δ $(x = y)$ is a characteristic function defined as:

$$. \delta(x{=}y) = \begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases}. \tag{9}$$

To verify the effectiveness of the distribution on steganalysis, 1000 speech segments last 10s are employed as speech samples. Figure 3(a) illustrates the average probability distribution of $P(d)$ calculated by the original speech and the steganographic one at 100% embedding bit rate. According to short-term invariance of pitch delay and the strong correction between the pitch delay in the same group, the distribution of $P(d)$ should be mainly concentrated in the range near 0. Steganography changes the distribution, $P(0)$ of the stego speech is much less than it calculated by the cover speech, and the distribution fluctuations of the cover speech are larger than the stego ones.

In addition, recompression is applied to both cover and stego samples. Figure 3(b) shows the average probability distributions of $P(d)$ calculated by recompressed cover and stego speech. Different to the distributions calculated by original stego speech, the distributions of both recompressed speech are similar to the ones of original cover speech. Thus the distributions of the recompressed speech can be used as the calibration to estimate the distributions of the original cover speech. If the difference between the distribution of the tested speech and its recompressed one is obvious, the tested speech could be seen as the speech which is carrying with secret message bits.

Figure 3. The average probability distribution of the difference between pitch delay in the same group



(a) Probability distributions of original speech

(b) Probability distributions of recompressed speech

On the other hand, the value of $P(d)$ is around 0 while d is in [-9, -5] and [6, 9], which means the difference value d is mainly concentrated in the range of [-4, 5]. As a result of that, [-4, 5] is chosen as the threshold for the difference value d. C-PDDPD feature can be calculated as follows:

$$\Delta P(d) = P_{\text{original speech}}(d) - P_{\text{recompressed speech}}(d) \quad d \in [-4,5] \tag{10}$$

where $\triangle P(d)$ is the difference value between $P(d)$ extracted from the original speech and its recompressed one. Because the difference value d is only taking value from -4 to 6, the dimension of C-PDDPD feature is 10.

3.2. Markov Transition Probability Matrix of The Difference of Pitch Delay in Same Group (MDPD)

The difference of pitch delay characterizes the variation of pitch period in short duration. According to the characteristics of speech signals, the variation of speech signals in short duration should be smooth and stable, from which we can learn that the adjacent difference values should have strong correlation with each other. For a speech segment contains N frames, we can consider the sequence of difference values $D = \{d_{0,0}, d_{1,0}, d_{0,1}, d_{1,1}, ..., d_{1,N-1}\}$ as a Markov chain. And the sequence of the difference values should satisfy this formula:

$$P(d_{cur} \mid d_{prev}) = \begin{cases} P(d_{0,i} \mid d_{1,i-1}) = P(d_{0,i} \mid d_{0,0}, d_{1,0}, d_{0,1}, \ldots, d_{1,i-1}) & , d_{cur} = d_{0,i} \\ P(d_{1,i} \mid d_{0,i}) = P(d_{1,i} \mid d_{0,0}, d_{1,0}, d_{0,1}, \ldots, d_{0,i}) & , d_{cur} = d_{1,i} \end{cases} \tag{11}$$

Then we can obtain the Markov transition matrix M_D as follow:

$$\mathbf{M}_D = \begin{bmatrix} P(d_{cur} = -9 \mid d_{prev} = -9) & \cdots & P(d_{cur} = -9 \mid d_{prev} = 9) \\ \vdots & \ddots & \vdots \\ P(d_{cur} = 9 \mid d_{prev} = -9) & \cdots & P(d_{cur} = 9 \mid d_{prev} = 9) \end{bmatrix} \tag{12}$$

where $P(d_{\text{cur}} = x \mid d_{\text{prev}} = y)$ is the probability that current difference value is x while the previous one is y. $P(d_{\text{cur}} = x \mid d_{\text{prev}} = y)$ can be obtained by:

$$P(d_{cur} = x \mid d_{prev} = y) = \frac{P(d_{cur} = x, d_{prev} = y)}{P(d_{prev} = y)} \tag{13}$$

Figure 4 illustrates the average Markov transition matrix M_D extracted from 1000 cover speech samples and corresponding stego speech samples. The stego speech samples are obtained by using Huang's (2012) and Yan's (2015) steganographic method at 100% embedding bit rate. We can draw two conclusions from the observation on the images.

First, steganography actually changes the distribution of the matrix, which means the variation of pitch delay has been obviously affected by bits embedding. The average cover matrix has particular fluctuations that the values are concentrated in the center near the point at $P(d_{\text{cur}} = 0 \mid d_{\text{prev}} = 0)$. But the distribution of the steganographic matrix is more evenly than the one calculated by the cover speech samples. According to the short-term invariance of pitch period, the variation of pitch delay should be smooth and stable in each group, and the adjacent difference values should be similar, which

means the difference value should be near 0 in most cases. Due to the randomness of secret message bits, steganography will make the matrix extracted from the stego samples more evenly distributed.

Secondly, the matrices of cover speech and stego speech are all mainly concentrated in the range of [-4, 5] on both vertical and horizontal ordinates. The reason is that the difference value d is mainly concentrated in the range of [-4, 5] as we discussed above. The probability values far away from the center of the matrix are invariant even when the embedding bit rate is 100%, which means that they are not suitable for distinguishing the stego speech from the cover ones. As a result of that, we only take the center of the matrix (the range on both vertical and horizontal ordinates are [-4, 5]) as the feature for steganalysis. The dimension of the MDPD feature is $10\times10=100$.

3.3. Steganalysis Scheme

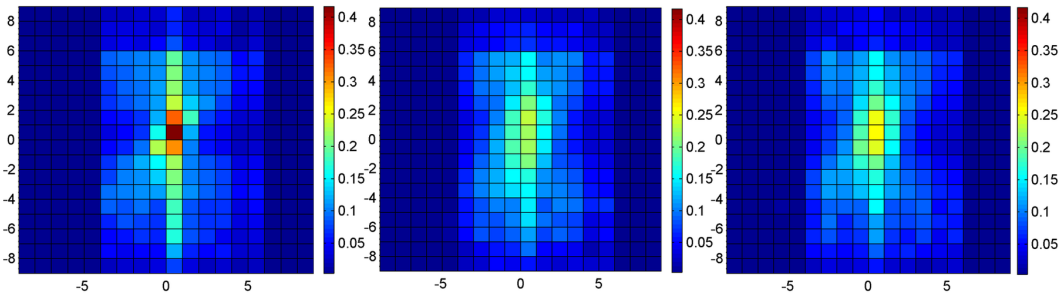
After the introduction of the features, we present a steganalysis scheme based on SVM. SVM is broadly employed in the field of steganalysis. As Figure 5 shows, the training procedure contains 4 steps:
Step 1: Collect large numbers of speech samples that half of the samples are produced by steganography method and the rest are compressed by the original AMR encoder. Then recompression is conducted to produce the recompressed cover and stego samples.
Step 2: Extract MSDPD, PDDPD from all speech samples including the original and recompressed ones, then calculate MDPD by only original samples.
Step 3: Obtain C-MSDPD and C-PDDPD by using MSDPD and PDDPD extracted from the original samples subject the ones extracted from the corresponding recompressed samples.
Step 4: Train the SVM-based classifier with C-MSDPD, C-PDDPD, MDPD and the cover or stego label of each sample.

Then the steps of test are described as follows:
Step 1: Recompress the tested speech segment with AMR encoder in the same mode.
Step 2: Obtain C-MSDPD, C-PDDPD, MDPD features from the original sample and the recompressed sample.
Step 3: Send the feature vector as input into the trained classifier to detect whether the speech is carrying with secret messages.

4. EXPERIMENTAL RESULT AND EVALUATION

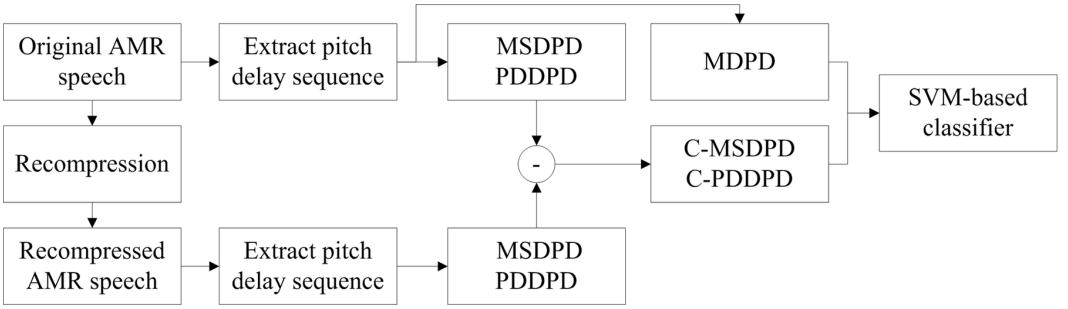
LibSVM is a famous open-source library for SVM. In this paper, LibSVM with default parameters ($c = 1$ and $g = 1 / 1064$) and RBF kernel is employed for training the SVM-based classifier. It is also the SVM used in Ren's (2017) method. To evaluate the performance of proposed approach, 2800 speech samples are collected for training and test in this paper. The sample set consists of four types

Figure 4. The average Markov Transition Matrix of the difference between pitch delay in the same group



(a) The average matrix of cover speech (b) The average matrix of stego speech produced by Yan's method (c) The average matrix of stego speech produced by Huang's method

Figure 5. The process of the proposed steganalysis method



of speech, which are English male speech, English female speech, Chinese male speech and Chinese female speech. The number of each type of speech is 700 and the length of each sample is 10s. Before starting the experiments, the whole speech samples are compressed by AMR encoder in 12.2kb/s mode to produce the cover speech samples. Then Huang's (2012) method and Yan's (2015) method are used for producing the stego speech samples at different embedding bit rates from 10% to 100%.

In each experiment, 1400 cover samples and their corresponding stego samples are randomly chosen as the training set. After the classifier is trained, the rest samples are employed for performance test. Table 1 and Table 2 respectively show the experimental results for detecting samples at different embedding bit rates and samples with various lengths. Table 1 records the detection accuracies of the proposed method and Ren's (2017) method when stego samples are produced by Huang's (2012) and Yan's (2015) steganographic method at different embedding bit rates from 10% to 100% with the length of 1s. From the table we can see that, the proposed method has obvious advantage than the state-of-the-art steganalysis method. For detecting Huang's (2012) steganographic method, the correct rate of the proposed method is up to 80.07% at the embedding bit rate of 70% while Ren's (2017) classifier can achieve only 78.46% correct rate even when the embedding bit rate is 100%. For detecting Yan's (2015) method, the advantage of the proposed method is more obvious. The proposed method can provide higher 7.5% accuracy than Ren's (2017) one at most. Table 2 shows the detection accuracies on detecting tested samples with different length from 1s to 10s, and the embedding bit rate of each stego sample is 50%. From the table it can be seen that, the proposed method is better than Ren's (2017) method under all speech lengths. The discrimination between the proposed method and Ren's (2017) method is obvious when the length of tested samples is only 1s. With the increase in speech length, the detection accuracies of both steganalysis methods are higher, but the proposed method still outperforms Ren's (2017) one in detecting all embedding methods.

To compare the performance of the two steganalysis methods more clearly, Figure 6 illustrates the receiver operating characteristic (ROC) curve when the two steganalysis schemes are employed for detecting speech samples with the length of 1s, 5s, and 10s at the embedding bit rate of 50%. ROC curve is often employed for illustrating the ability of binary classifier. If the area under the curve (AUC) is bigger, the performance of the classifier could be seen as better. From the figures it could be seen more clearly that the proposed method outperforms Ren's (2017) one with different speech lengths. When the tested sample is short, the discrimination between the two methods will be obvious.

The experimental result demonstrates that proposed method has better performance than the state-of-the-art one in all kinds of conditions and it is most effective on short-length speech detection, which means it is more suited to real-time speech steganalysis scenario because it can detect the covert communication only in 1s - 3s and the detection accuracy will increase during the process of communication. For the both two existing steganographic methods, the proposed method can all effectively detect the target from tested samples and provides high accuracy even when the tested sample is short or the embedding bit rate is low.

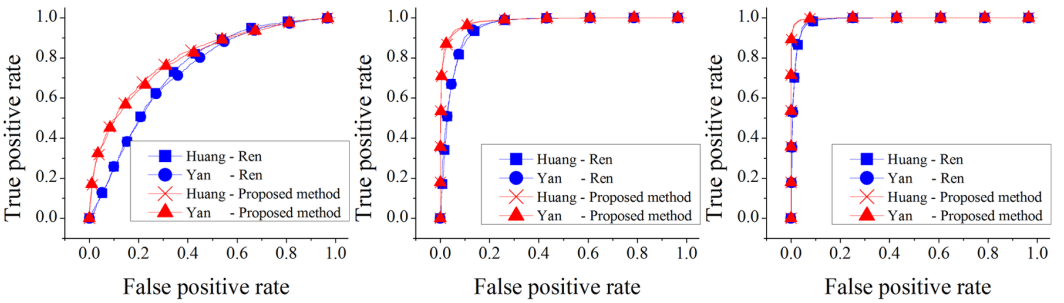
Table 1. The detection accuracies of two steganalysis methods at different embedding bit rate

Embedding method	Huang' steganography		Yan's steganography	
Embedding bit rate	Proposed method	Ren's method	Proposed method	Ren's method
10%	53.93%	52.93%	53.75%	50.82%
20%	59.54%	56.79%	58.14%	54.21%
30%	65.14%	62.21%	62.75%	59.89%
40%	68.21%	65.54%	68.04%	63.68%
50%	73.11%	69.07%	72.11%	66.14%
60%	76.68%	71.46%	75.64%	71.43%
70%	80.07%	74.79%	79.43%	74.14%
80%	81.39%	74.57%	83.39%	77.07%
90%	82.71%	77.21%	86.00%	78.50%
100%	84.86%	78.46%	87.89%	81.68%

Table 2. The detection accuracies of two steganalysis methods with different sample length

Embedding method	Huang' steganography		Yan's steganography	
Sample length	Proposed method	Ren's method	Proposed method	Ren's method
1s	73.11%	69.07%	72.11%	66.14%
2s	83.43%	80.29%	82.29%	78.04%
3s	88.50%	84.36%	87.68%	84.29%
4s	90.89%	86.54%	91.14%	87.75%
5s	93.18%	89.54%	93.79%	89.71%
6s	94.82%	91.25%	95.18%	91.61%
7s	96.11%	92.29%	96.46%	92.71%
8s	96.89%	93.00%	96.86%	93.50%
9s	97.32%	93.86%	97.54%	94.18%
10s	97.96%	94.32%	97.96%	94.75%

Figure 6. The ROC curve of two steganalysis methods



(a) The ROC of two steganalysis methods when speech length is 1s

(b) The ROC of two steganalysis methods when speech length is 5s

(c) The ROC of two steganalysis methods when speech length is 10s

5. CONCLUSION

In this article, we have proposed a novel steganalysis method for detecting steganography based on modification of pitch delay. By analyzing of the search strategy of pitch delay, we discovered that the correlation between adjacent pitch delay is different. Based on this, the pitch delay in the same frame is divided into some groups, and two new types of features (C-PDDPD and MDPD) based on the pitch delay in the same group and C-MSPDP feature are employed to train the classifier. Experimental result shows that the proposed method outperforms the state-of-the-art one especially in short length speech detection. Because of the high accuracy on detecting short length speech, the proposed method can meet the high requirement of real-time covert communication detection.

REFERENCES

Bas, P. (2010). Using high-dimensional image models to perform highly undetectable steganography. *Paper presented at International Conference on Information Hiding*, Calgary, Canada.

Denemark, T., Boroumand, M., & Fridrich, J. J. (2016). Steganalysis Features for Content-Adaptive JPEG Steganography. *IEEE Transactions on Information Forensics and Security*, *11*(8), 1736–1746. doi:10.1109/TIFS.2016.2555281

Ding, Q., & Ping, X. (2010, December). Steganalysis of Analysis-by-Synthesis Compressed Speech. *Paper presented at International Conference on Multimedia Information Networking and Security*, Nanjing, China. doi:10.1109/MINES.2010.148

Ekudden, E., Hagen, R., Johansson, I., & Svedberg, J. (1999, June). The adaptive multi-rate speech coder. *Paper presented at 1999 IEEE Workshop on Speech Coding Proceedings*, Porvoo, Finland.

Filler, T., Judas, J., & Fridrich, J. J. (2011). Minimizing Additive Distortion in Steganography Using Syndrome-Trellis Codes. *IEEE Transactions on Information Forensics and Security*, *6*(3), 920–935. doi:10.1109/TIFS.2011.2134094

Fridrich, J. J., & Kodovsky, J. (2012). Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security*, *7*(3), 868–882. doi:10.1109/TIFS.2012.2190402

Geiser, B., & Vary, P. (2008, March). High rate data hiding in ACELP speech codecs. *Paper presented at International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, NV. doi:10.1109/ICASSP.2008.4518532

Goljan, M., Fridrich, J. J., & Cogranne, R. (2014). Rich model for Steganalysis of color images. *Paper presented at International Workshop on Information Forensics and Security*, Atlanta, GA. doi:10.1109/WIFS.2014.7084325

3. GPP/ETSI. (2016). *Digital Cellular Telecommunications System (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Mandatory Speech Codec Speech Processing Functions; Adaptive Multi-Rate (AMR) Speech Codec; Transcoding Functions (3GPP TS 26.090 version 13.0.0 Release 13)", Technical Report TR 126 090*. France: Sophia Antipolis Cedex.

Hess, W., & Shaughnessy, D. O. (1984). Pitch Determination of Speech Signals: Algorithms and Devices by Wolfgang Hess. *The Journal of the Acoustical Society of America*, *76*(4), 1277–1278. doi:10.1121/1.391349

Holub, V., & Fridrich, J. (2013). Digital image steganography using universal distortion. *Paper presented at the 1st ACM workshop on information hiding and multimedia security*, Montpellier, France.

Huang, Y., Liu, C., Tang, S., & Bai, S. (2012). Steganography Integration Into a Low-Bit Rate Speech Codec. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1865–1875. doi:10.1109/TIFS.2012.2218599

Huang, Y., Yuan, J., Chen, M., & Xiao, B. (2011). Key distribution over the covert communication based on VoIP. *Chinese Journal of Electronics*, *20*(2), 357–360.

Jiang, Y., Tang, S., Zhang, L., Xiong, M., & Yip, Y. J. (2016). Covert voice over internet protocol communications with packet loss based on fractal interpolation. *ACM Transactions on Multimedia Computing Communications and Applications*, *12*(4), 54. doi:10.1145/2961053

Li, S., Jia, Y., & Kuo, C. C. J. (2017). Steganalysis of qim steganography in low-bit-rate speech signals. *IEEE/ACM Transactions on Audio Speech & Language Processing*, *25*(5), 1011-1022.

Li, S., Tao, H., & Huang, Y. (2012). Detection of quantization index modulation steganography in G.723.1 bit stream based on quantization index sequence analysis. *Journal of Zhejiang University Science C*, *13*(8), 624–634. doi:10.1631/jzus.C1100374

Li, S. B., Jia, Y. Z., Fu, J. Y., & Dai, Q. X. (2014). Detection of pitch modulation information hiding based on codebook correlation network. *Chinese Journal of Computers*, *37*(10), 2107–2117.

Liu, J., Tian, H., Lu, J., & Chen, Y. (2016). Neighbor-index-division steganography based on qim method for g.723.1 speech streams. *Journal of Ambient Intelligence and Humanized Computing*, *7*(1), 1–9. doi:10.1007/s12652-015-0315-6 PMID:27042240

Liu, K., Yang, J., & Kang, X. (2017). Ensemble of CNN and rich model for steganalysis. *Paper presented at International Conference on Systems, Signals and Image Processing*, Poznan, Poland. doi:10.1109/IWSSIP.2017.7965617

Luo, X., Song, X., Li, X., Zhang, W., Lu, J., Yang, C., & Liu, F. (2016). Steganalysis of hugo steganography based on parameter recognition of syndrome-trellis-codes. *Multimedia Tools and Applications*, *75*(21), 13557–13583. doi:10.1007/s11042-015-2759-2

Mazurczyk, W. (2013). VoIP steganography and its detection—a survey. *ACM Computing Surveys*, *46*(2), 1–21. doi:10.1145/2543581.2543587

Mazurczyk, W., & Lubacz, J. (2010). LACK--a VoIP steganographic method. *Telecommunication Systems*, *45*(2), 153–163. doi:10.1007/s11235-009-9245-y

Miao, H., Huang, L., Chen, Z., Yang, W., & Alhawbani, A. (2012). A new scheme for covert communication via 3G encoded speech. *Computers & Electrical Engineering*, *38*(6), 1490–1501. doi:10.1016/j.compeleceng.2012.05.003

Miao, H., Huang, L., Shen, Y., Lu, X., & Chen, Z. (2013, October). Steganalysis of Compressed Speech Based on Markov and Entropy. *Paper presented at International workshop on digital watermarking*, Auckland, New Zealand.

Provos, N., & Honeyman, P. (2003). Hide and seek: an introduction to steganography. *IEEE Symposium on Security and Privacy*, 99(3), 32-44. doi:10.1109/MSECP.2003.1203220

Ren, Y., Cai, T., Tang, M., & Wang, L. (2015). AMR Steganalysis Based on the Probability of Same Pulse Position. *IEEE Transactions on Information Forensics and Security*, *10*(9), 1801–1811. doi:10.1109/TIFS.2015.2421322

Ren, Y., Yang, J., Wang, J., & Wang, L. (2017). Amr steganalysis based on second-order difference of pitch delay. *IEEE Transactions on Information Forensics and Security*, *12*(6), 1345–1357. doi:10.1109/TIFS.2016.2636087

Sedighi, V., Cogranne, R., & Fridrich, J. J. (2016). Content-Adaptive Steganography by Minimizing Statistical Detectability. *IEEE Transactions on Information Forensics and Security*, *11*(2), 221–234. doi:10.1109/TIFS.2015.2486744

Tian, H., Liu, J., & Li, S. (2014). Improving security of quantization-index-modulation steganography in low bit-rate speech streams. *Multimedia Systems*, *20*(2), 143–154. doi:10.1007/s00530-013-0302-8

Tian, H., Wu, Y., Chang, C. C., Huang, Y., Chen, Y., Wang, T., & Liu, J. et al. (2017). Steganalysis of adaptive multi-rate speech using statistical characteristics of pulse pairs. *Signal Processing*, *134*(C), 9–22. doi:10.1016/j.sigpro.2016.11.013

Wu, Z., Cao, H., & Li, D. (2015). An approach of steganography in g.729 bitstream based on matrix coding and interleaving. *Chinese Journal of Electronics*, *24*(1), 157–165. doi:10.1049/cje.2015.01.026

Wu, Z., & Yang, W. (2006). Speech information hiding in g.729. *Chinese Journal of Electronics*, *15*(3), 545–549.

Xiao, B., Huang, Y., & Tang, S. (2008, December). An Approach to Information Hiding in Low Bit-Rate Speech Stream. *Paper presented at Global Telecommunications Conference*, New Orleans, LA. doi:10.1109/GLOCOM.2008.ECP.375

Yan, S., Tang, G., & Sun, Y. (2015). Steganography for low bit-rate speech based on pitch period prediction. *Jisuanji Yingyong Yanjiu*, *32*(6), 1774–1777.

Yanpeng Wu was born in 1989. He received the B.S. degree in Materials Forming and Control Engineering in 2012 and the M.S. degree in Computer Application Technology in 2016 from National Huaqiao University. He is currently working at Mobile Forensics R&D Center of Xiamen Meiya Pico Information Co., Ltd., China. His research interests include multimedia security, steganalysis and mobile forensics.

Huiji Zhang was born in 1980. He received the B.S. degree in Electronic and Information Engineering from Xiamen University in 2003. He is currently working at Mobile Forensics R&D Center of Xiamen Meiya Pico Information Co., Ltd., China. His research interests include digital forensics, data recovery and artificial intelligence in digital forensics.

Yi Sun was born in 1986. He received the B.S. degree in Business Administration from Jimei University in 2009 and the M.S. degree in Computer Science from Xiamen University in 2014. He is currently working at Mobile Forensics R&D Center of Xiamen Meiya Pico Information Co., Ltd., China. His research interests include IoT forensics and mobile forensics.

Minghui Chen was born in 1983. He received the M.S. degree in Computer Software and Theory from University of Science and Technology of China in 2009. He is currently working at Mobile Forensics R&D Center of Xiamen Meiya Pico Information Co., Ltd., China. His research interests include computer security, data recovery and mobile forensics.