This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2011 proceedings

An Approach of Covert Communication Based on the Adaptive Steganography Scheme on Voice over IP

Rui Miao
Department of Electronic Engineering,
Tsinghua University,
Beijing, China
rm870725@gmail.com

Yongfeng Huang
Department of Electronic Engineering,
Tsinghua University,
Beijing, China
yfhuang@tsinghua.edu.cn

Abstract—**An effective covert communication system can be achieved based on steganography in stream media such as VoIP (Voice over IP), and the LSB-based (Least-Significant-Bit) methods are the most popular strategies. However, LSB embedding in the flat regions of the speech could inevitably bring about the perceptible distortion leading to the degradation of speech quality. Besides, LSB methods are vulnerable for diverse corresponding steganalysis like RS detection. This paper presents an adaptive steganography scheme that selects lower embedding bit rate in the flat blocks, while chooses higher embedding bit rate in the sharp blocks. In addition, we design an overflow judgment to guarantee the synchronous secret data transmission. Further, the rational tradeoff between hiding requirements and speech quality is also taken into account to construct a practical covert communication system on G.711 VoIP. Extensive experiments showed that the proposed adaptive scheme outperformed the LSB and could evade RS steganalysis. This proposed scheme is efficient, simple and can be used in real-time VoIP network with high hiding capacity.**

***Index Terms*—Block steganography, adaptive steganography, steganalysis, covert communication**

I. INTRODUCTION

Information hiding refers to the secret data hidden in a cover object without causing appreciable distortion that aims at making them inconspicuous and unaware towards an external observer. The technology of information hiding usually is applied to implement a covert communication system. Differing from hiding secret data in storage media [1][2], information hiding based on network real-time communication and streaming media features such as VoIP (Voice over IP) has attracted wide research focus in recent years. This alternative scenario could potentially enhance the security and capacity for several inspirations. Firstly, streaming media communication occupies a large amount of network resources, and thus the secret data could more easily to camouflage among them. Secondly, the hidden message could be transmitted successively and dynamically due to its real-time scenario, and the detector has insufficient time and calculating capacity to trace and discern the secret data communication. Consequently, an effective and inconspicuous covert communication can be designed underlying the public network channels [3][4][5].

Nowadays, covert communication based on VoIP puts forward some more critical requirements in aspects of real-time, hiding capacity, speech quality and security. However, popular LSB-based (Least Significant Bit) methods fail to take into account the non-linear compression scheme of speech stream and the perceptible distortion for sensitive character of HAS (human auditory system)[6]. Besides, it is also vulnerable for various corresponding steganalysis like RS detection [7][8][9].

The algorithm to hide information in VoIP needs to closely scrutinize a variety of speech quality assessment criteria, thus determining the rational tradeoff between degradation of speech quality (additional delay, slight distortion) and covert communication requirements (imperative security, sufficient capacity). This paper proposes an adaptive steganography scheme in G.711 [10] speech stream. This method examines the smoothness of block in a speech stream to select corresponding hiding capacity with superior performance.

The rest of this paper contains four sections. Section 2 discussed recent advanced and problems in the related work and proposed our solutions. Section 3 described in detail the proposed adaptive steganography scheme. Section 4 analyzed the performance by experiment results. Section 5 made conclusions and scheduled future works.

II. RELATED WORKS

Recently, the LSB algorithm has been applied to a number of covert communication systems [3][4]. J. Dittmann [3] proposed distinguishing between active and silence intervals in speech, and Z. Wu [11] pointed out that large amplitude samples are also not suitable for hiding information, since it deduced that an embedding into these specific intervals is the most perceptible distortion in speech features. Nevertheless, the assessment of speech quality is inadequate and transmission errors are always existence because of its asynchronism. A. Ito [12] pointed out that reducing distortion does not necessarily improve subjective speech quality. A. Ito proposed an enhanced LSB substation algorithm based on estimation of tolerable distortion to achieve high speech quality. The implementation of the algorithm is that a low bit rate encoder, G.726 ADPCM, is employed as a reference for deciding the number of bits embedded into a sample.

Overall, the methods above can be concluded as improved LSB substitution algorithms that merely consider the number of bits hidden in the discriminative samples. Since the

This work was supported in part by grants from the National Foundation Theory Research of China (973 Program, No. 2007CB310806), and the National Natural Science Foundation of China (No. 60970148, and No. 60773140).

978-1-61284-233-2/11/$26.00 ©2011 IEEE

substitution operation could inevitably modify the spatial structure of the carry media and it is vulnerable for diverse corresponding steganalysis like RS detection [7][8][9]. Therefore, the security of the LSB-based algorithms above is being threatened. In addition, these algorithms also did not consider the feature of speech stream. Consequently, conventional substitution operation is not appropriate and spatial correlation of speech samples needs to be taken into account to achieve less degradation of speech quality.

One feasible strategy is to exploit multiple regions steganography [13][14]. On the other hand, there are increasing focuses on adaptive steganography on edge areas of image [1][2][15]. The adaptive steganography exploits the pixel-value differencing (PVD) of two consecutive pixels to determine the embedding capacity, and then secret information is embedded through modifying the difference value. However, the parameter configuration is too subjective and spatial correlation can be further explored to obtain higher stego quality. Y. Chen [16] examined spatial correlation of multiple pixel-values and presented a steganographic scheme for images by block smoothness ranging. However, there is no detailed discussion on being extended to compressed speeches.

In addition, these related works all focused on the steganography algorithms rather synchronous scheme for covert communication application.

Reviewing the inspiration of scrutinizing spatial correlation above, this paper proposed a block-based adaptive steganography method in speech streaming such as G.711. A rational tradeoff between speech quality and hiding capacity is taken into account to design an effective covert communication system based on VoIP.

III. ADAPTIVE STEGANOGRAPHY IN VOIP

A. The model of covert communication based on VoIP

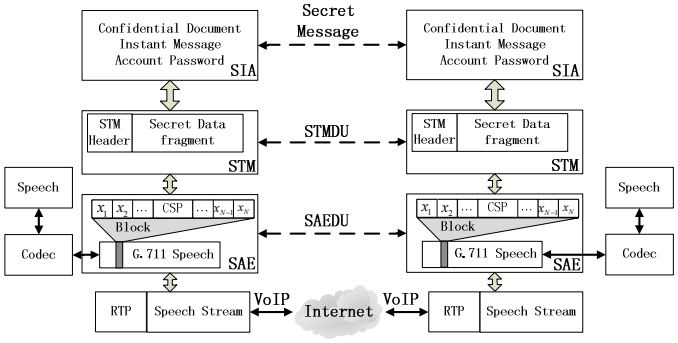

Figure 1. The model of covert communication on VoIP. The steganographic architecture contains three layers: SIA (Steganographic Information Application layer), STM (Steganographic Transmission Management layer) and SAE (Steganographic Adaptation & Execution layer). Specifically, SAE employs the proposed adaptive scheme to consecutively operate on every block along the G.711 speech stream. And then the stego-speech stream is delivered to the RTP protocol and transmitted on the public VoIP channel.

Our previous works [5][17][18] proposed a practical and effective model of covert communication on VoIP. In this paper, we utilize our previous architecture and employ our proposed adaptive scheme. An overall architecture of achieving a covert communication channel on VoIP is depicted in Fig.1. The proposed adaptive steganography scheme selects different embedding bit rate according to the smoothness of the sample block, bolstering rational balance between speech qualities and hiding requirements.

The original inspiration of this block-based method is that the flat blocks in the cover speech, especially the silence intervals, are more sensitive from our analysis towards HAV and extensive experiments, and therefore data embedding in these flat blocks will inevitably cause perceptible contamination, which will lead to poor acoustical quality and low security. Consequently, less secret bits are embedded in the flat block, while the sharp block can camouflage more secret bits.

B. Embedding algorithm

1) Block division rule.

In following statement, we put our focus on a-law G.711 PCM (PCMA), though the adaptive scheme is also suitable for μ-law G.711 PCM. For requirement of numerical computation in steganographic algorithm, the folded binary code [10] of PCMA is needed to be converted into complement binary code. The mapping of the two code words is shown in Table I. And then, cover voice samples are divided into consecutive blocks with the same size $N = 2k + 1$. i.e. $B = (x_1 x_2 \ \dots \ x_N)$.

TABLE I. BINARY CODE MAPPING

Decimal	Folded binary	Complement
+128	11111111	01111111
~	~	~
1	10000000	00000000
-1	00000000	111111111
~	~	~
-128	01111111	10000000

2) Adaptive embedding scheme based on the smoothness of the block.

The block mean value is calculated by (1) with downward rounding.

$$\mu = \left\lfloor \frac{\sum_{i=1}^{N} x_i}{N} \right\rfloor \quad (1)$$

And then the block smoothness is measured according to the difference value d_i between each sample and block mean value. Central sample point (CSP) is not embedded, only offsets for modifications. We obtain

$$d_i = \mu - x_i, i \neq k + 1 \quad (2)$$

As we discussed above, in the first place, the different value indicates the smoothness of the block. Further, the essential purpose of the adaptive scheme is to discriminate the gradation of the difference value, and embed corresponding size of secret data. In the second place, the difference value is modified to camouflage secret data. For the strict consistency in retrieving process, the modified difference value should accord with the same gradation.

Let $\mathcal{D}$ denotes the all gradations of the difference value, and $\mathcal{M}$ denotes the set of secret message. We have:

$$f: \mathcal{D} \times \mathcal{M} \rightarrow \mathcal{D} \quad (3)$$

$$g: \mathcal{D} \to \mathcal{M} \tag{4}$$

Where, f is the embedding mapping and g is the retrieving mapping. To substantiate the validity of the transmitting, the algorithm needs to satisfy the following condition:

$$\forall (D, M) \in \mathcal{D} \times \mathcal{M}, g\big(f(D, M)\big) = M \tag{5}$$

According to the Tab 1, decimal numerical range of PCMA is $[-128,128]$, and thus the range of difference value is $I = [-256,256]$. A feasible gradation scenario is that

$$\begin{aligned} I &= \bigcup_{r \in R} D_r = D_{-7} \bigcup D_{-6} \ldots \bigcup D_{-1} \bigcup D_0 \bigcup D_1 \bigcup D_2 \ldots \bigcup D_7 \\ &= [-256,-128] \bigcup [-127,-64] \ldots \bigcup [-3,-2] \bigcup [-1,1] \bigcup [2,3] \bigcup [4,7] \ldots \bigcup [128,256] \end{aligned} \tag{6}$$

Let us assume $\exists r \in R$, that $d_i \in D_r = [l_i, u_i]$, where l_i, u_i denote the lower bound and upper bound of D_r respectively. Here, we can use $\forall d_i^{'} \in D_r$ to substitute the original d_i . Specifically, this selection among the difference set can manifest the secret message of $\lfloor log_2 \| D_r \| \rfloor$ bits. However, multiple bits modification should be deliberated and D_0 generally denotes the silence intervals, and therefore a limitation in embedding bits is needed as shown in Table II.

TABLE II. DIFFERENCE VALUE GRADATION

Gradation	D_0	D_1 D_{-1}	D_2 D_{-2}	D_3 D_{-3}	others
Embedding bits(n)	0	1	2	3	4

Further, when d_i is small, the number of elements in its gradation set is small too, which delicately corresponds with the adaptive principle that the flat region is more sensitive and should embedded less secret data bits. On the contrary, sharp region can camouflage more secret data bits.

m_i is the n bits secret message searched from TABLE II. The original difference value d_i is substituted by $d_i^{'}$ with the secret data. We have:

$$d_i^{'} = \begin{cases} l_i + m_i, & if\ d_i \geq 0 \\ u_i - m_i, & otherwise \end{cases} \tag{7}$$

Then, the steganographic voice sample values are given by (8). Particularly, CSP offsets these modifications for maintaining consistent block mean value.

$$\begin{cases} x_i^{'} = \mu - d_i^{'}, i \neq k+1 \\ x_{k+1}^{'} = \mu + \sum_{i=1, i \neq k+1}^{N} d_i^{'} \end{cases} \tag{8}$$

3) Overflow judgment.

From our analysis of G.711 non-linear compression scheme as well as extensive experiments, we deduce that an embedding into samples of large amplitude will bring about strong noise. Besides, the offset operation on CSP could probably overflow the value range (beyond 128 or under -128). As a result, restricting the steganographic modification under an amplitude threshold becomes necessary as a security and transparency enhancing measure.

Meanwhile, algorithm should guarantee that the overflow judgment carried out at the embedding process is entirely consistent with the retrieving process, so that the secret data can be embedded and retrieved synchronously and effectively without additional synchronous flag bits. Accordingly, the overflow judgment needs to select those identical parameters at both sides of the communication. That is: block mean value (μ), gradation ($\bigcup_{r \in R} D_r$), as well as gradation bound (l_i, u_i). Then the value ranges of the steganographic samples are shown in (9).

$$x_i^{'} \in \begin{cases} [\mu - l_i\ ,\ \mu - u_i]\ , 1 \leq i \leq N, i \neq k+1 \\ [\mu + \sum_{i=1, i \neq k+1}^{N} l_i\ ,\ \mu + \sum_{i=1, i \neq k+1}^{N} u_i], i = k+1 \end{cases} \tag{9}$$

The maximum possible amplitude of steganographic sample is estimated by (10).

$$\lambda_i = \begin{cases} max_{1 \leq i \leq N, i \neq k+1} \{ |\mu - l_i|, |\mu - u_i| \}, 1 \leq i \leq N, i \neq k+1 \\ max_{i=k+1} \{ \left|\mu + \sum_{i=1, i \neq k+1}^{N} l_i\right|, \left|\mu + \sum_{i=1, i \neq k+1}^{N} u_i\right| \}, i = k+1 \end{cases} \tag{10}$$

Hence, the algorithm defines a threshold λ_x. For a certain voice sample, the judgment $\lambda_i > \lambda_x$ means the steganographic operation induces a data overflow. If this sample is a non-central point of the block, then the steganographic operation will be cancelled only on this particular sample. However, if the data overflow occurs on the CSP, in this case, steganographic camouflages on all samples of the block should be cancelled.

C. Retrieving algorithm

According to the same principles and the identical parameters as the embedding algorithm, when the sample value is no overflow, the embedded secret message can be retrieved synchronously by

$$m_i^* = \begin{cases} -l_i + d_i^{'}, & if\ \ d_i^{'} \geq 0 \\ u_i - d_i^{'}, & otherwise \end{cases} \tag{11}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

We conducted the experiments to estimate the embedding capacity and the speech quality with the proposed adaptive scheme. We adopted the steganographic architecture according to the Fig. 1 to implement a covert communication system on VoIP. The performance of covert communication system is evaluated by three aspects: objective testing for speech quality, subjective testing based ABX method and the performance against RS steganalysis. Test environment is:

TABLE III. TEST SPEECH SAMPLES

Grouping	Group I	Group II	Group III	Group IV
Style	reading, Male	reading, Female	dialogue, Males	dialogue, Females
Num.&length	10,60s/per	10,60s/per	10,60s/per	10,60s/per
Noise	Small	Small	Strong	Strong
Silence intervals	More	More	Few	Few

A. Objective Testing for the Speech Quality

Objective testing for the speech quality adopted the evaluation criterions of SNR (signal to noise ratio) and MOS-LQO (Mean Opinion Score-Listening Quality Objective). MOS-LQO [19], ITU P.862.1 objective standard for speech quality, is compatible with the universal MOS value of the subjective speech quality evaluation [20]. MOS-LQO value is

converted from ITU-T P.862 PESQ (Perceptual Evaluation of Speech Quality) [21].

1) Overall Test

We took the overall tests on the 40 speech samples in Table II with $N = 13$ and $\lambda_x = 49$. Average values of objective parameters for each Group were calculated and listed in Table IV.

TABLE IV. OBJECTIVE TESTING RESULTS ($(N = 13, \lambda_x = 49)$)

Group	SNR	Original SNR	MOS-LQO	Original MOS-LQO	Hiding Bit rate
I	36.68	37.43	4.23	4.42	2276
II	36.57	37.41	4.07	4.31	2476
III	36.56	37.53	3.59	3.93	7663
IV	35.98	37.51	3.50	4.07	7435

Original SNR and MOS-LQO denote the referential quality of carry-speech processed by G.711 non-linear compression without steganography.

More silence intervals and weaker background noise indicate more flat speech blocks, implying lower hiding capacity according to our proposed adaptive scheme, as listed in Table IV (Group I & II). On the contrary, few silence intervals and stronger background noise indicate more sharp blocks, resulting in higher hiding capacity (Group III & IV).

2) Characteristic curve of λ_x

In order to analyze the impact of λ_x, we arbitrarily selected one test sample from Group III in Table II, and fixed the block length $N = 13$, and successively increase the overflow threshold (λ_x). We recorded and analyzed the relationship between embedding capacity and speech quality. The SNR and MOS-LQO values are plotted against the bit rate of embedding bits in Fig. 2.

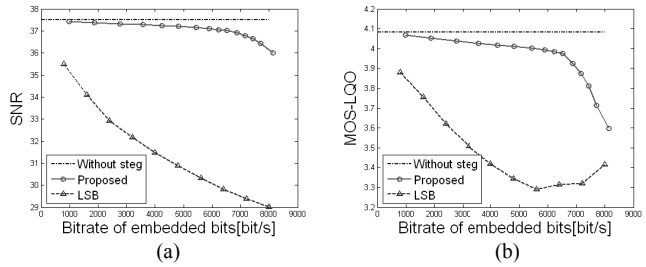

Figure 2. λ_x curve(a) SNR/Bitrate;(b) MOS-LQO/Bitrate. LSB is employed as a reference with the same embedding strength. As a reference, the carry-speech quality (Dash-and-dot line) also suffered the degradation from the G.711 non-linear compression scheme. By contrast, the degradation of stego-speech quality is caused by both the G.711 non-linear compression and steganography.

In Fig. 2, curves reveal the relationship between the hiding capacity (embedding bit rate) and the speech quality (SNR or MOS-LQO), with the different value of overflow threshold (λ_x). Specifically, when λ_x value is small, stego-speech can obtain good SNR and MOS-LQO values. With the increase in λ_x, embedding capacity increases but speech quality degrades, and this degradation approximates linear. However, when λ_x increases beyond a certain value, it appears a sharp decline in SNR and MOS-LQO, reflecting the drastic degradation of speech quality. At any rate, compared with the LSB substitution, the algorithm proposed in this paper achieves higher SNR and MOS-LQO, proving its superior effectiveness.

3) Characteristic curve of N

For λ_x taking several specific values, we examined the performance of steganography along with the changes in block length (N). As shown in Fig. 3, the modification of N leaded to a small fluctuation on the referential λ_x curve (dash-and-dot line). That is, according to a larger value of N, the stego-speech would acquire a better speech quality and sacrifice partial capacity. The variation is basically consistent with the reference dash-and-dot line.

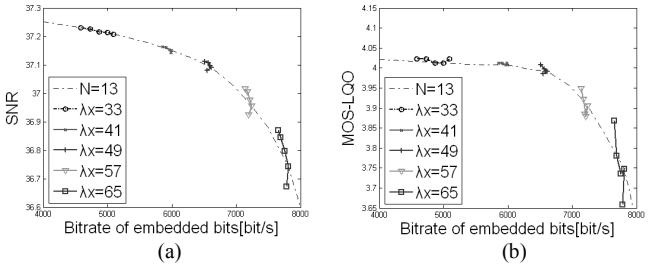

Figure 3. N curve(a) SNR/Bitrate;(b) MOS-LQO/Bitrate. The referential dash-and-dot line was the λ_x characteristic curve from Fig. 2 with N=13

To sum up, the regulation of tradeoff between the hiding capacity and speech quality mainly depend on the configuration of λ_x. In addition, the parameter N can partially and delicately adjust the performance. These two parameters could be cooperatively configured for a particular practical application.

B. Subjective Testing based ABX method

The so-called "A/B/X" test method [22] was employed to estimate the subjective quality of the stego-speech with the proposed adaptive steganography. We invited 40 testers and each one respectively listened 6 groups of A/B/X files. We recorded the frequency of correct judgments and plotted its statistic histogram, as shown in Fig. 4.

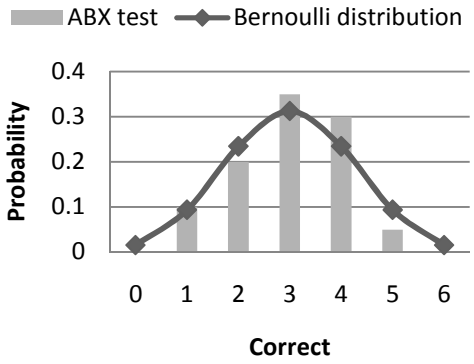

Figure 4. ABX test results distribution. As a reference, we also depicted the Bernoulli distribution on it, i.e. ξ~B(6,0.5).

From Fig. 5 we can see that the statistic histogram basically coincides with the Bernoulli distribution, implying that it is

more likely that the tester randomly decided the judgment rather than genuinely distinguished them depending on human acoustic systems. In other words, the indistinguishability of stego-speech and original speech reveals the good security of our proposed algorithm.

C. Performance against the RS Steganalysis

We adopted an effective steganalysis method to detect covert communication, which proposed by our previous works [8][9] with a unique sliding window mechanism and an improved RS algorithm. RS algorithm [7] is a very accurate and reliable detection measure in both digital images and speeches. Configure the flipping mask M be $[0\ 1\ 1\ 0]$. We randomly chose 20 test samples from Table II for RS steganalysis, and we compared these results with no steganography, LSB method and adaptive steganography in same cover speeches. It is shown in Fig. 5.

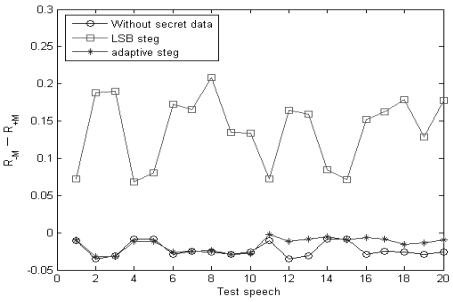

Figure 5. Performance against the RS steganalysis

From the difference diagram of R_{-M} and R_{+M}, the RS steganalysis gave us accurate and reliable detections for stego-speeches with LSB algorithm. However, stego-speeches with adaptive steganography were hardly distinguished from speeches without embedding. Namely, the RS steganalysis cannot detect the hidden information in stego-speeches with adaptive embedding method.

V. CONCLUSION

In this paper, we have presented an effective covert communication system based on VoIP, adopting an adaptive steganography scheme to embed different size of secret data in speech stream according to the smoothness of speech block. Compared with LSB method at the same hiding capacity, our proposed scheme could achieve much better speech quality, revealing the sufficient capacity and superior security. Specifically, around 7.5 kbps of secret information could be embedded while the degradation of the MOS-LQO value is less than 0.5.

Besides, compared with the vulnerable security of LSB method, our proposed method could substantially evade the current steganalysis like RS detection.

Depending on required application background, the proposed algorithm determines the tradeoff between the embedding capacity and the degradation of speech quality by selecting the appropriate parameters of block length (N), and the overflow threshold (λ_x).

Since there are increasing demanding on speech quality and hiding capacity in covert communication system, further exploring the character of human auditory system to optimize adaptive steganography scheme would be one of our future works.

REFERENCES

[1] W. Luo, F. Huang, J. Huang, "Edge adaptive image steganography based on LSB Matching Revisited." Information Forensics and Security, IEEE Transactions on, 2010, 5(2): 201-214.

[2] C. Yang, C. Weng, S. Wang, H. Sun, "Adaptive data hiding in edge areas of images with spatial LSB domain systems." Information Forensics and Security, IEEE Transactions on, 2008, 3(3): 488-497.

[3] J. Dittmann, T. Vogel, R. Hillert, "Design and evaluation of steganography for voice-over-IP". ISCAS 2006

[4] C. Wang and Q. Wu, "Information hiding in real-time VoIP streams. Ninth IEEE International Symposium on Multimedia", 2007, pp.255-262

[5] Y. Huang, J. Yuan, S. Tang, C. Wang, "Steganography in inactive frames of the source codec", IEEE Transactions on Information Forensics and Security, in press.

[6] W. Bender, D. Gruhl, N. Morimoto, "Techniques for data hiding". IBM System Journal, 1996, 35(3, 4):pp.313~336

[7] J. Fridrich, M. Goljan, R. Du, "Detecting LSB steganography in color, and gray-scale images." Multimedia, IEEE, 2001, 8(4): 22-28.

[8] Y. Huang, Y. Zhang, S. Tang, "Detection of covert voice-over Internet protocol communications using sliding window-based steganalysis",IET communications, in press.

[9] Y. Huang, S. Tang, C. Bao. Y.J. Yip, "Steganalysis of compressed speech to detect covert voice over Internet protocol channels", IET Information Security, in press.

[10] ITU-T G.711, Pulse code modulation (PCM) of voice frequencies, 1988.

[11] Z. Wu and W. Yang, "G.711-Based adaptive speech information hiding approach". ICIC 2006, pp.1139~1144

[12] A. Ito, S. Abe, Y. Suzuki, "Information hiding for G.711 speech based On substitution of Least Significant Bits and estimation of tolerable distortion", ICASSP 2009, pp.1409-1412.

[13] W. Zhang, X. Zhang, S. Wang, "A double layered "Plus-Minus One" data embedding scheme." Signal Processing Letters, IEEE, 2007, 14(11): 848-851.

[14] J. Guo and T. Le, "Secret communication using JPEG double compression." Signal Processing Letters, IEEE, 2010, 17(10): 879-882.

[15] D. Wu, W. Tsai, "A steganographic method for images by pixel-value differencing," Patten Recognition Letters, 2003, 24(9/10): 1613-1626.

[16] Y. Chen, A steganographic scheme for images by block smoothness ranging, Department of Communications Engineering, Feng Chia University, 2005

[17] H. Tian, et al. "An M-Sequence based steganography model for Voice over IP," Communications, IEEE International Conference on (ICC '09), 2009, pp. 1-5.

[18] B. Xiao, Y. Huang. "Modeling and optimizing of the information hiding communication system over streaming media," Journal of Xidian University, Chinese, 2008, Vol35, pp:554-558

[19] ITU-T P.862.1, Mapping function for transforming of P.862 to MOS-LQO, 2003

[20] ITU-T P.800.1, Mean Opinion Score (MOS) terminology, 2006

[21] ITU-T P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001

[22] Z.M. Lu, B. Yan, and S.H. Sun, "Watermarking Combined with CELP Speech Coding for Authentication," IEICE Transactions on Information and System, 2005, E88-D(2), pp. 330-334.