An Adaptive Steganography Scheme for Voice over IP

[*]Hui Tian, [*]Ke Zhou, [†]Hong Jiang, [‡]Yongfeng Huang, [*]Jin Liu, [*]Dan Feng
[*]Wuhan National Lab for Optoelectronics, School of Computer, Huazhong University of Science and Technology, Wuhan, China
[†]Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0150, USA
[‡]Department of Electronic Engineering, Tsinghua University, Beijing, China
{huitian, jinliu}@smail.hust.edu.cn, {k.zhou, dfeng}@hust.edu.cn, jiang@cse.unl.edu, yfhuang@tsinghua.edu.cn

***Abstract*—This paper presents an adaptive steganography scheme for Voice over IP (VoIP). Differing from existing steganography techniques for VoIP, this scheme enhances the embedding transparency by taking into account the similarity between Least Significant Bits (LSBs) and embedded messages. Moreover, we introduce the notion of Partial Similarity Value (PSV). By properly setting the threshold PSV, we can adaptively balance the embedding transparency and capacity. We evaluate the effectiveness of this scheme with G.729a as the codec of the cover speech in StegTalk, a covert communication system based on VoIP. The experimental results demonstrate that our technique provides better performance than the traditional method.**

I. INTRODUCTION

Recently, steganography has drawn increasing attention. Many different steganographic methods for storage media (e.g., image [1], audio [2], text [3], etc.) have been proposed in the last few years. However, the area of steganography for real-time systems is largely unexplored. In part due to the fact that the real-time characteristic of real-time systems is a double-edged sword. While the real-time nature potentially offers better security for secret messages by virtue of its instantaneity, it does not allow many complex operations, which increases the difficulty in assuring security. Nevertheless, given its potential advantages, steganography for real-time systems may soon become a worthy subject of further studies. In this paper, we focus on one of the typical real-time communication systems, Voice over IP (VoIP), as a possible carrier to apply steganography to provide security for secret messages.

VoIP is a promising technique to enable telephone calls via a broadband Internet connection. Owing to its advantages of low cost and advanced flexible digital features, VoIP has become a popular alternative to the public-switched telephone network (PSTN), and extensive research on it has been conducted [4]. The main motivations for our VoIP-based steganography study are twofold. First, the ongoing conversation of VoIP can offer an ideal camouflage for secret messages, because the voice data is naturally assumed to be the only data carried in a given VoIP channel. Second, a typically short VoIP connection does not give eavesdroppers a sufficient amount of time to detect possible abnormity due to hidden messages. Recently, some researchers have noticed the advantages of and carried out useful studies on steganography over VoIP [5-9]. However, all of these studies adopt the same simple embedding strategy, replacing the least significant bits (LSBs) of the cover speech with the binary bits of secret messages or their encrypted form without leveraging the characteristics of LSBs. Although LSBs modifications usually have little impact on the quality of cover media, blind substitution potentially risks being detected by statistic methods. This view has been widely accepted by researchers working on steganography on storage media [10-12]. In fact, the key criteria for steganography are perfect transparency to non-authenticated entities and high capacity for carrying secret messages. The first criterion, a measure of embedding distortion, is often more important. An acknowledged belief is that the smaller the embedding distortion, the harder the detection of the embedding changes. Therefore, in this study, we focus on an adaptive steganography scheme for VoIP, which aims at minimizing the distortion of speech quality to enhance the imperceptibility of the steganography system. While the pursuit of transparency is bound to decrease the capacity, as is well known, we are interested in proving an acceptable balance between these two criteria.

To make this paper self-contained, in Section II, we briefly summarize the necessary background of VoIP steganography and formally define some of the concepts discussed in this paper. The proposed adaptive steganography scheme for VoIP is described in Section III, which is followed by the evaluation of this scheme and its test results that are presented in Section IV. The paper is concluded with remarks on the paper's main contribution and future work in Section V.

II. BACKGROUND AND DEFINITION

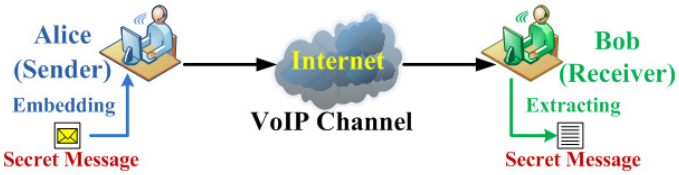

Fig.1. The framework of VoIP steganography

Fig. 1 depicts a general framework of VoIP steganography. Let us assume that Alice (the sender) wants to transmit some secret messages to Bob (the receiver), while they are talking about some inconspicuous topics via the VoIP system. For that,

This work is supported in part by National High Technology Research and Development Program of China (863 Program) under Grant No. 2006AA01Z444, National Basic Research Program of China (973 Program) under Grant No. 2004CB318201, Program for New Century Excellent Talents in University (No. NCET-06-0650), and National Natural Science Funds of China under Grant No. 60773140.

978-1-4244-3828-0/09/$25.00 ©2009 IEEE

Alice hides some secret messages into the VoIP stream with an embedding algorithm; after the voice is sent through the VoIP channel, Bob retrieves the secret messages with the corresponding restituting algorithm.

Based on the traditional definition of steganography [13], we formally describe VoIP steganography as follows:

Definition 1 (VoIP Steganography): Given the quadruple $\emptyset = <C, M, D, E>$, where C is the set of VoIP packet (VoIP stream), M is the set of secret messages with $|C| \geq |M|$, E: $C \times M \rightarrow C$ represents the embedding function and D: $C \rightarrow M$ represents the restituting function. A system with the property that $D(E(c, m)) = m$ for $\forall\, c \in C$ and $m \in M$ is called a **VoIP steganography system** .

As in most practical steganography systems, the two communication parties in the VoIP steganography system also must exchange secret messages without raising any suspicions. Therefore, we should guarantee the perceptual similarity between the cover and the corresponding stego-object (the cover embedded with secret message). The authors in [13] defined perceptual similarity via a similarity function as follows:

Definition 2 (Similarity Function): Let C be a non-empty set. A function sim: $C^2 \rightarrow (-\infty, 1]$ is called the **similarity function** on C, for $x, y \in C$, if $x = y$, $sim(x, y) = 1$; otherwise, $sim(x, y) < 1$.

Further, the transparency criterion for steganography can be formally described that for $\forall\, c \in C$ and $m \in M$ maximizing the value of the following function:

$$f(c,m) = sim(c, E(c,m)) - 1 \tag{1}$$

The LSBs substitution has a premise that the modification of LSBs is not enough to induce perceptual distortion. However, the selection of LSBs often largely depends on subjective opinions. Thus, the potential ineffectiveness of LSBs substitution on covers cannot be ignored, especially when employed in applications with high security requirements. In the following text, we will present a novel adaptive steganography approach for VoIP, which enhances the embedding transparency by taking into account the similarity between LSBs and embedded messages.

III. The Adaptive Steganography Scheme for VoIP

Let us consider the following hypothetical case: if the LSB stream of the cover speech is exactly the same as the bit stream of secret messages to be embedded, the steganography has the best transparency without any induced distortion of speech quality, i.e. $f(c, m)$ reaches its maximum value 0. This observation suggests that an ideal cover may be obtained if its LSBs match perfectly with the binary bits of the secret messages. However, it is impossible to find such an ideal match for most given secret messages, making this approach impractical. Another possible approach is to only replace the portion of LSBs that are equal to the binary bit of secret messages. This, however, will inevitably sacrifice the embedding capacity of the cover speech in favor of good transparency, which may not be a significant concern if the cover speech sufficiently long. Unfortunately, the length of the cover speech depends on the conversation, which is often short and unpredictable. Thus we must strike an acceptable balance between transparency and capacity. On the other hand, the selection of LSBs in this approach cannot be fixed beforehand and may even be uncertain, making it very difficult, if not impossible, for the receiver to determine which LSBs conceal the secret messages. The main idea behind our proposed new steganography approach, which attempts to strike a good balance between embedding transparency and capacity as detailed next, stems from these observations.

Let us assume that, $M = \{m_1, m_2, \ldots, m_L\}$ is the bit set of a given secret message (it may be encrypted beforehand, which is irrelevant in this scheme), where L is the length of the secret message; $B = \{b_1, b_2, \ldots, b_N\}$ is the LSB set of each VoIP packet, where, N is the total number of LSBs, and the selection of LSBs can refer to previous methods [9]. We divide B into S parts, namely, $B = \{B'_1, B'_2, \ldots, B'_S\}$, where $B'_i = \{b'_{i1}, b'_{i2}, \ldots, b'_{in}\}$, $i = 1, 2, \ldots, S$; $n = N/S$; $b'_{ij} = b_{(i-1)\cdot n + j}$, $j = 1, 2, \ldots, n$. Accordingly, we divide M into Q parts, i.e. $M = \{M'_1, M'_2, \ldots, M'_Q\}$, where $M'_i = \{m'_{i1}, m'_{i2}, \ldots, m'_{in}\}$, $i = 1, 2, \ldots, Q$; $Q = L/n$; $m'_{ij} = m_{(i-1)\cdot n + j}$, $j = 1, 2, \ldots, n$. Further, we define the following partial similarity function:

$$\varepsilon(X,Y) = count(x_i = y_i) \tag{2}$$

where, $X = \{x_1, x_2, \ldots, x_T\}$, $Y = \{y_1, y_2, \ldots, y_T\}$, T is the length of X and Y, x_i, $y_i = 0$ or 1, $i = 1, 2, \ldots, T$. It means that the value of $\varepsilon(X, Y)$ is the number of identical bits between X and Y. If X and Y are one part of B and M respectively, the value of $\varepsilon(X, Y)$ is called the **Partial Similarity Value (PSV)**. If the threshold PSV is set as η, for $\varepsilon(X, Y) \geq \eta$, X can be replaced with Y, and for $\varepsilon(X, Y) < \eta$, X is not changed. This process can be formalized as follows:

$$X_y = \varphi(X,Y) = \begin{cases} X, & \text{If } \varepsilon(X,Y) < \eta \\ Y, & \text{If } \varepsilon(X,Y) \geq \eta \end{cases} \tag{3}$$

Accordingly, function (1) can be converted into:

$$f(X,Y) = \varepsilon(X,Y)/T - 1 \tag{4}$$

For the given threshold PSV η, the condition of replacing X with Y is $f(X, Y) \geq \eta/T - 1$.Obviously, if $\eta = T$ (in this case, $T = n$), the embedding process has the best transparency but the smallest capacity; if $\eta = 0$, the embedding process can achieve the maximum capacity but the minimum transparency. However, we can adaptively balance the transparency and capacity via properly setting the threshold PSV for the given n. Fig. 2 illustrates this adaptive embedding process.

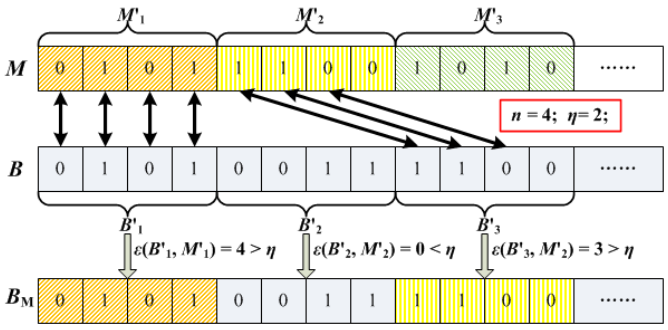

Fig.2. The adpative embedding process

In our scheme, the problem of how to tell the receiver the adopted LSBs is tantamount to how to indicate the embedded parts in each VoIP packet. Our solution is to set a flag bit for each part in every VoIP packet. According to the aforementioned definition, we need S flag bits, denoted by $FB = \{fb_1, fb_2, \ldots, fb_S\}$. fb_i ($i = 1, 2, \ldots, S$) can be determined as follows:

$$fb_i = \begin{cases} 1, & \text{If } \varepsilon(X_y, Y) = n \\ 0, & \text{If } \varepsilon(X_y, X) = n \end{cases} \quad (5)$$

In the above formula, $fb_i = 1$ indicates that the ith part has been replaced, otherwise, the ith part has not been changed. Obviously, if we divide B into more parts, we will need more flag bits. As far as the number of flag bits is concerned, a smaller S is preferable. However, the number of parts can also impact the performance of embedding, because the embedding transparency and capacity depend on the values of parameters n and η. Therefore, choosing a proper S involves a difficult tradeoff. In this prototypical scheme, we determine parameter S empirically. Another problem is where to place these flag bits. In our previous study [13], we present a synchronization mechanism using techniques of the protocol steganography, in which synchronization patterns are hidden in the unused and/or optional fields of the header of a certain packet. In the IP header, there are a total of 64 bits that can be used to embed messages. Moreover, the headers of upper-level protocols, such as UDP, RTP, etc, also have many unused or optional fields. Therefore, the sender can distribute the flag bits among those fields in a predetermined manner. Because the flag bits are often few and altered continually, such a transmission of flag bits is potentially hard to discover. The receiver knows exactly where the flag bits are embedded. When the receiver receives an IP packet, he (she) first checks for every flag bit, and then extracts the corresponding hidden parts. After collecting all these parts, the receiver can successfully restitute the whole secret message.

IV. Evaluation and Test

TABLE I. Steganography Modes

Mode	Description
1	Choosing 8 bits of LSBs in each frame and direct replacing them with secret message bits
2	Choosing 16 bits of LSBs in each frame and direct replacing them with secret message bits
3	Choosing 16 bits of LSBs in each frame and adaptive embedding secret message bits with $n = 4$, $\eta = 2$.
4	Choosing 16 bits of LSBs in each frame and adaptive embedding secret message bits with $n = 4$, $\eta = 3$.
5	Choosing 16 bits of LSBs in each frame and adaptive embedding secret message bits with $n = 4$, $\eta = 4$.

We evaluate the proposed scheme in StegTalk [9], a prototypical covert communication system based on VoIP. StegTalk supports typical coders, such as ITU-T G.711, G.723.1, G.729a, etc. In the tests, we typically choose G. 729a [14] as the codec of the cover speech. Similarly, we define the LSBs based on the observation that the parameters of fixed codebook in G.729a have the best transparency for information hiding [9, 13]. Our tests focus on performance comparison between the traditional LSBs substitution method and the proposed adaptive steganography method. Thus, we design five steganography modes as shown in TABLE I. The first two modes are based on the traditional LSBs substitution method, and the rest are based on the proposed adaptive steganography method. In what follows, we describe the results of the tests on a single phrase of speech and some audio samples.

A. Test One

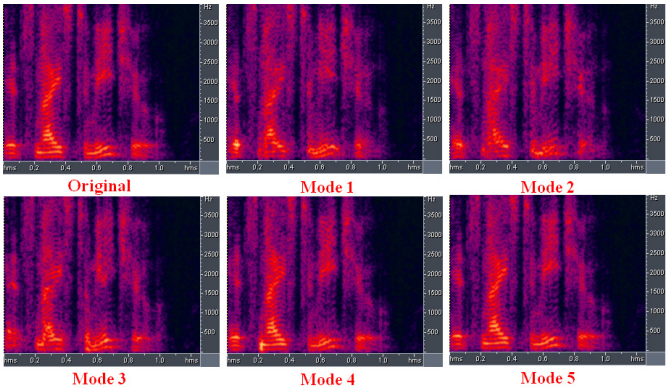

Fig.3. Spectrograms contrast

In the test, we compare the spectrograms of the cover speech without steganography and with steganography for each mode. We choose the phrase "It is nice to meet you" as the cover speech (denoted by P) and ISCAS 2009 CALL FOR PAPERS file (CFP15v.pdf) [15] as the secret message. After being encoded by G. 729a, P has 126 frames. TABLE II shows the lengths of the hidden secret message in each mode. Fig. 3 shows the spectrograms of the original speech and its steganographical versions with the secret message embedded in five different steganography modes respectively. From them, we can observe the following facts: (1) the spectrogram for Mode 2 shows the most visible differences from the original spectrogram, indicating that Mode 2 has the worst transparency. However, it provides the largest embedding capacity; (2) the spectrogram for Mode 5 is exactly the same as the original spectrogram, because the embedding process does not change any bits. However, it provides the smallest embedding capacity; (3) the spectrograms for Mode 1, Mode 3 and Mode 4 show relatively minor differences from the original spectrogram, which indicates that the speech quality degradations in these modes are negligible. Among them, Mode 4 offers the best transparency as it has fewer differences than the others but at the cost of embedding capacity. Mode 1 and Mode 3 can provide nearly the equal transparency, but Mode 3 offers larger embedding capacity than mode 1 (about 1.43 times). Therefore, Mode 3 keeps the best balance between embedding transparency and capacity; and (4) based on an overall consideration of transparency and capacity, the proposed adaptive steganography method provides better performance than the traditional LSB substitution method.

TABLE II. Steganography Capacity at Each Mode

Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
1008 bits	2016 bits	1444 bits	588 bits	128 bits

B. *Test Two*

To further compare the performances of the traditional LSBs substitution method and the proposed adaptive steganography method, we collect 150 one-minute audio samples. These samples consist of three categories: male speech, female speech and piano music. All the samples are recorded with 8000 HZ sampling rate, 16 bits quantization and mono, and encoded by G. 729a before the embedding process. For each sample, we perform the steganography experiment in the above five modes respectively. The secret message (also choosing CFP15v.pdf, actually only some forward parts) can be successfully embedded and retrieved in any case. Furthermore, for the evaluation of the embedding transparency, we also carry out Mean Opinion Score (MOS) [16] experiments on all steganographical samples. MOS is a mapping of perceived levels of distortion into the descriptive terms "excellent, good, fair, poor, unsatisfactory". Accordingly, MOS value is defined within the range of [1, 5]. Generally speaking, MOS value of speeches encoded by G. 729a is 4.0 [14].

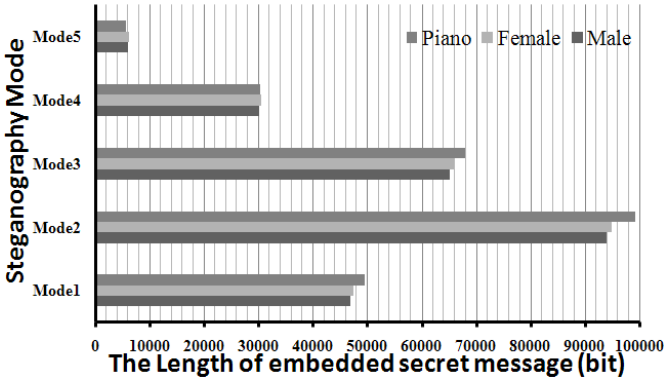

Fig.4. Test results of embedding capacity

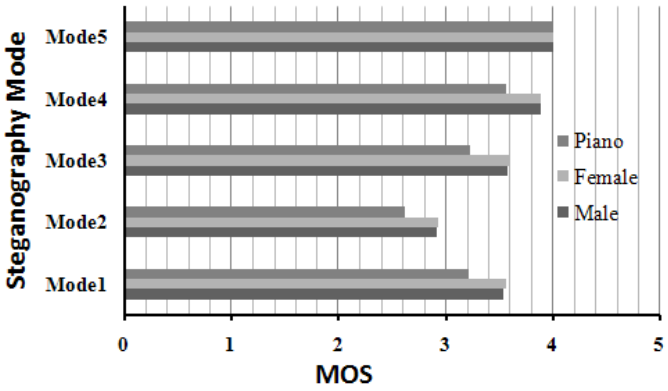

Fig.5. Test results of embedding transparency

Fig. 4 and Fig.5 show the statistical results of the mean length of the embedded secret messages and the mean MOS value for each category for all five modes. The data in the two charts further verify the conclusions drawn in Test one. In addition, we can observe that the quality degradations of the piano samples are more easily perceivable than the speech samples in the same steganography mode. This may attribute to the fact that piano music often has strong rhythms and accordingly its signal correlation is high, which makes changes more sensitive for the human audience. However, the sound quality of steganographical piano samples is still acceptable**.** Therefore, the adaptive steganography can be employed reliably and effectively in the VoIP steganography system.

V. Conclusion and Future Work

In this paper, we proposed an adaptive steganography scheme for VoIP. The salient advantage of this scheme is that it can adaptively balance the embedding transparency and capacity. Moreover, experimental results demonstrate that our method outperforms the traditional LSBs substitution method. For further work, we will study the effect of both parameter n and parameter η. Furthermore, the impact of latency induced by the steganography scheme on real VoIP conversations will be investigated.

References

[1] K. Bailey, K. Curran. “An evaluation of image based steganography methods” , Multimedia Tools and Applications, vol. 30, Issue 1, pp. 55-88, July 2006.

[2] M. Pooyan, A. Delforouzi. “LSB-based Audio Steganography Method Based on Lifting Wavelet Transform”, in Proc. of the 2007 IEEE International Symposium on Signal Processing and Information Technology, pp. 600 – 603, 15-18 Dec. 2007.

[3] Y. Liu, X. Sun, C. Gan,and H. Wang. “An Efficient Linguistic Steganography for Chinese Text”, in Proc. of 2007 IEEE International Conference on Multimedia and Expo, pp. 2094 – 2097, 2-5 July 2007.

[4] B. Goode. “Voice over Internet protocol (VoIP)”, Proceedings of the IEEE, vol. 90, Issue 9, pp. 1495-1517, Sept. 2002,.

[5] C. Wang, Q. Wu. “Information hiding in real-time VoIP streams”, in Proc. 9th IEEE International Symposium on Multimedia, pp. 255-262, 10-12 Dec. 2007.

[6] J. Dittmann, D. Hesse and R. Hillert. “Steganography and steganalysis in voice over IP scenarios: operational aspects and first experiences with a new steganalysis tool set”, in Proceedings of SPIE, vol. 5681, Security, Steganography, and Watermarking of Multimedia Contents VII, March 2005, pp. 607-618.

[7] C. Kratzer, J. Dittmann, T. Vogel and R. Hillert. “Design and evaluation of steganography for voice-over-IP”, in Proc. of 2006 IEEE International Symposium on Circuits and Systems, pp. 2397-2340, 21-24 May 2006.

[8] W. Mazurczyk, Z. Kotulski. “Covert Channel for Improving VoIP Security”, in Proc. of Multiconference on Advanced Computer Systems (ACS), pp. 311-320, Oct. 2006.

[9] H.Tian, K. Zhou, Y. Huang, J. Liu and D. Feng. “A Covert Communication Model Based on Least Significant Bits Steganography in Voice over IP”, in Proc. of the 9th International Conference for Young Computer Scientists, Nov. 2008, in press.

[10] H. Sun, K. Wang, C. Liang and Y. Kao. “A LSB substitution compatible steganography”, in Proc. of 2007 IEEE Region 10 Conference (TENCON 2007), pp. 1-3, Oct. 30-Nov. 2 2007.

[11] J. Fridrich. “Minimizing the embedding impact in steganography”, in Proc. of the 8th ACM workshop on Multimedia & security, pp. 2-10, Sep. 2006.

[12] J. Fridrich, T. Pevný and J. Kodovský. “Statistically undetectable jpeg steganography : dead ends challenges, and opportunities”, in Proc. of the 9th ACM workshop on Multimedia & security, pp. 3-14, Sep. 2007.

[13] H.Tian, K. Zhou, J. Hong,etc. “An M-Sequence Based Steganography Model for Voice over IP”, Technical Report, Department of Computer Science and Engineering, University of Nebraska-Lincoln, TR-UNL-CSE-2008-0007, September 2008.

[14] ITU-T, Recommendation G.729. “Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)”, Jan. 2007.

[15] Available at: http://conf.ncku.edu.tw/iscas2009/CFP15v.pdf.

[16] ITU-T Recommendation P.800. "Methods for subjective determination of transmission quality", Aug. 1996.