J. Cent. South Univ. (2014) 21: 4544−4552
DOI: 10.1007/s11771-014-2459-5

Springer

Frame-bitrate-change based steganography for voice-over-IP

LIU Jin(刘进)[1, 2], TIAN Hui(田晖)[3], ZHOU Ke(周可)[1, 2]

1. School of Computer Science and Technology, Huazhong University of Science and Technology,
Wuhan 430074, China;
2. Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology,
Wuhan 430074, China;
3. College of Computer Science and Technology, National Huaqiao University, Xiamen 361021, China

Abstract: Steganography based on bits-modification of speech frames is a kind of commonly used method, which targets at RTP payloads and offers covert communications over voice-over-IP (VoIP). However, direct modification on frames is often independent of the inherent speech features, which may lead to great degradation of speech quality. A novel frame-bitrate-change based steganography is proposed in this work, which discovers a novel covert channel for VoIP and introduces less distortion. This method exploits the feature of multi-rate speech codecs that the practical bitrate of speech frame is identified only by speech decoder at receiving end. Based on this characteristic, two steganography strategies called bitrate downgrading (BD) and bitrate switching (BS) are provided. The first strategy substitutes high bit-rate speech frames with lower ones to embed secret message, which introduces very low distortion in practice, and much less than other bits-modification based methods with the same embedding capacity. The second one encodes secret message bits into different types of speech frames, which is an alternative choice for supplement. The two strategies are implemented and tested on our covert communication system StegVoIP. The experiment results show that our proposed method is effective and fulfills the real-time requirement of VoIP communication.

Key words: covert communication; steganography; multi-rate speech codec; voice-over-IP (VOIP)

1 Introduction

Since the advent of IP network application, traditional public switched telephone network (PSTN) service has been gradually invaded and occupied by packet-switched IP telephony, such as voice-over-IP (VoIP) services. People prefer to use cost-effective online calls. And some of those VoIP applications have already occupied an important position in our daily lives. However, the problems of privacy leak and insecure communication come along with them. Steganography, an emerging technology that embeds secret message in ordinary covering media (such as image, audio and other multimedia) without knowing the very existence of hiding procedure itself, brought challenges to information and communication security for IP telephony society. And also, individuals and organizations may employ these techniques for advantageous covert communication [1].

Differing from static covering media (image, audio or text), steganography in VoIP streams is a real-time application, which means restricted bandwidth and latency. That gives limited embedding capacity and processing time for steganography [2]. Fortunately, the existence of a range of communication protocols for VoIP service increases the extent of redundancy, which brings more opportunities for steganography. These protocol-based steganographic methods include hiding in the packet headers of IP, TCP/UDP, SIP, SDP, RTP/RTCP [3], and etc. Most of them are based on the modification of some unused bits in protocol headers, which may be vulnerable by the opening of steganographic algorithms according to Kerckhoff's principle in cryptography.

Other steganography methods are often based on VoIP payloads (speech frames), which are encoded and decoded by speech codecs such as ITU-T Recommendation G.711, ITU-T Recommendation G.723.1, ITU-T Recommendation G.729, IETF RFC 3951 (iLBC), Speex, GSM RTE-LTP, and etc. All these

Foundation item: Project(2011CB302305) supported by National Basic Research Program (973 Program) of China; Projects(61232004, 61302094) supported by National Natural Science Foundation of China; Project(ZQN-PY115) supported by Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University, China; Project(JA13012) supported by Education Science Research Program for Young and Middle-aged Teacher of Fujian Province of China; Project(2014J01238) supported by Natural Science Foundation of Fujian Province of China
Received date: 2013−08−01; **Accepted date:** 2013−11−19

Corresponding author: ZHOU Ke, Professor, PhD; Tel: +86−13971075916; E-mail: k.zhou@hust.edu.cn

payloads based methods that directly modify the speech frames [4] may result in obvious speech quality distortion for human perception. This characteristic limits their usage (less embedding capacity). In order to resist statistic steganalysis [5] or other potential attacks [6], the similarities between secret messages and cover speech [7] can also increase to reduce the distortion of speech quality or use pseudorandom sequences [8] to eliminate the statistic characteristics of secret messages and enhance the security of it. Furthermore, LSB classifying [9] based matrix coding and state-based least-significant-digits embedding methods [4] can be utilized to reduce introduced speech quality distortion, or in other words, to enhance embedding capacity.

However, most of the above steganographic methods are based on the bits-modification of speech frames and independent of the internal features of speech, which may lead to detectable degradations of speech quality. From the perspective of implementation, the way to obtain embedding capacity with minimum distortion is to replace cover speech frames by modified frames with the best speech quality under the condition of smaller frame size. To achieve this goal, replacing ordinary frames with speech frames encoded by existing lower bitrate speech codec is always the best choice. That's the main principle of our proposed steganographic method in this work.

2 Related works and motivations

To obtain the best speech quality through steganography, it is feasible to substitute current covering speech frames with frames encoded by a different encoding algorithm with smaller size. Consequently, if the smaller frames are correctly decoded, the steganography with less speech quality degradation is obtained. A silent frame steganography algorithm is proposed in Ref. [10]. It is based on the feature of ITU-T Recommendation G.723.1 speech codec, in which a piece of voice will be encoded into smaller silent frames once detecting that there is no active voice on proceeding call. Then, the embedding procedure is completed by substituting inactive frame (silent frame) with an active one (voice frame) and applies commonly used least-significant-bit (LSB) substitution method to it. It is proved that this method introduces less distortion than simple LSB substitution. However, primitive encoded frames are still explicitly modified to embed secret messages, which cause an extra degradation of speech quality. And the substituted silent frames will incur much more attacks for the potential possibility of steganography.

A TranSteg steganographic method is proposed in Ref. [11], which uses transcoding technique for covert communication in VoIP networks. The main idea is to substitute original speech signal with lower bitrate speech encoded by a different encoding algorithm with similar speech quality but smaller payload size. As a result, the remaining extra space can be filled with secret messages. The embedding operation makes no modification in RTP headers during the steganography procedure. So, any steganalytic algorithm performed on VoIP protocol headers is in vain. And the payload safety is ensured by SRTP protocol which encrypts the RTP payload for secure transmission. Experimental results showed that this method brought almost undetectable speech quality degradation and considerable embedding capacity. However, this method has mainly two restrictions. Firstly, the required speech codecs are determined by the client system. That is, multiple speech codecs are needed. Next, in the embedding procedure, one coverted conversation needs at least four times of coding operations (decoding or encoding). Considering the processing time and look-ahead delay of some codecs (e.g., 37.5 ms for G.723.1), that will result accumulated latency and extra speech quality degradation.

A new steganographic method using the variable bitrate (VBR) feature within one specific speech codec is proposed in this work, so that encoding and decoding procedures can be correctly performed without the need of multiple codecs support. Two embedding strategies are implemented based on this method, named bitrate downgrading (BD) and bitrate switching (BS). BD strategy is based on the speech frame size downgrading in RTP packets, where the downgraded frame is encoded by the same speech codec, the caused speech quality degradation is tiny and could not be aware by listeners or eavesdroppers in a limited time period. BS embedding strategy is based on bitrate change according to secret message bits, which can resist detections on the transmission path towards unencrypted RTP payloads.

The proposed frame-bitrate-change based steganography takes widely used multi-rate speech codec ITU-T Recommendation G.723.1 [12] (G.723.1 for short) as an example, while the principle is applicable to other multi-rate speech codecs or even other streaming media formats with multiple types. In BD strategy, the higher bitrate frame type (6.3 kb/s) is the nominal coding bitrate while actually lower bitrate frames (5.3 kb/s) are packaged and the remaining spaces are utilized for embedding. The resulted speech quality outperforms vast majority of payload based steganography methods. And it introduces less processing time and negligible transmission delay. BS strategy is based on the uncertainty of frame bitrate and encodes the bitrate change manner for secret message embedding. The transmitting speech frames can be decoded by normal speech codecs correctly. This strategy has a lower embedding capacity yet more suitable for an unencrypted network environment.

3 Covert communication over VoIP

Typical covert communication model is derived from classic prisoner-warden model in Ref. [13], where the covert channel is derived from any available “redundant information” selected from general information exchange procedure. This procedure can be simply formalized as Eqs. (1) and (2). s is the covering media with secret message inside (stego cover). s' denotes the received version of s through covert communication channel. If no error occurs during transmission, s is equal to s'. m denotes secret message and c indicates covering speech stream. E and R represent the embedding and retrieving functions, respectively. k_{ERS} represents the embedding and retrieving secret key for security enhancement. m' indicates the secret message retrieved at the receiving end. Similarly, m is equal to m' if no error occurs.

$$s = E\ (m,\ k_{\text{ERS}},\ c) \tag{1}$$

$$m' = R\ (s',\ k_{\text{ERS}}) \tag{2}$$

3.1 Covert communication channels in VoIP

When it comes to VoIP domain, a family of network protocols are available for steganography. Figure 1 describes a commonly used basic covert communication model for VoIP. Alice (A) wants to send a piece of secret message to Bob (B). However, the line is monitored by a malicious eavesdropper Clark (C). So, they negotiate beforehand to transfer message through covert channels in VoIP streams and keep the normal communication unaffected in order not to draw C’s attention. Many protocols are available for this purpose as shown in the figure, and also some out-of-band covert channels (such as related steg) exist. The speech covert channel represents steganography based on speech payload, which includes all payload alteration based steganography including proposed method. RTP/RTCP covert channel is used for reliable transmission for our covert communication, which will be mentioned later.

3.2 Speech payload negotiation

At the beginning of a VoIP conversation, participants negotiate about the proper speech codecs of the following communication. Generally, more and more VoIP applications adopt session initiation protocol (SIP) as their signaling protocol which makes use of session description protocol (SDP) to designate supported media codec and other useful information. Once the conversation is established, the participants declare the upcoming payload type (PT) in RTP packet’s PT field, which is a protocol in application layer built on connectionless protocol UDP for media transmission on Internet. In VoIP communication it is used as the underlying speech payload carrier.

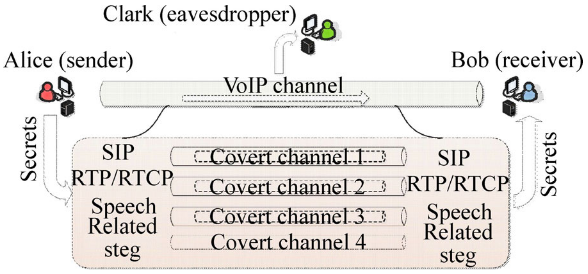

Fig. 1 Covert communication channels in VoIP

In proposed method, only the PT field in RTP is focused on, which indicates the carrying audio or video payload type in current RTP packet. PT has 7 bits and value 4 indicates G.723.1 speech codec [14]. If certain speech codec is not declared in the standard, the participants need to negotiate the specifications beforehand in the range of PT values assigned as dynamic in Table 1. Otherwise, the receiver will ignore RTP packets they do not recognize.

Table 1 Value assignment of RTP’s PT field

Value	Encoding	Media type
0−18	Static	Audio
25, 26, 28, 31−34	Static	Video
96−127	Dynamic	Audio/Video
Other	Unassigned/Reserved	Audio/Video

3.3 ITU-T Recommendation G.723.1 speech codec

In order to adjust to unstable network environment, many kinds of speech codecs contain multi-rate speech coding algorithms such as ITU-T Recommendation G.723.1 [12], IETF iLBC, GSM RTE-LTP and open-source Speex codec. In this work, G.723.1 speech codec is selected as an example while the principle can also be extended to other multi-rate codecs, even for images and videos. G.723.1 codec has two bitrates: 6.3 kb/s and 5.3 kb/s, which represent two kinds of frames. Both of them are mandatory parts of the encoder and decoder. The speech frames are divided into four types with the lowest two bits of the first byte for bitrate and silent frame indication.

Table 2 gives the specification of the two flag bits. SID (silence insertion descriptor) frame is used to

Table 2 ITU-T Recommendation G.723.1 flag bits

Value	Frame type	Frame size (Octets)
00	6.3 kb/s	24
01	5.3 kb/s	20
10	SID	4
11	Reserved	−

indicate comfort noise if silent compression algorithm is activated to reduce the transmission rate. The 5.3 kb/s frame has a smaller size than 6.3 kb/s frame. They are generated by different encoding and decoding algorithms in G.723.1 codec, and the decoding speech quality of the former is a little worse than the latter.

4 Proposed steganography method

A novel VoIP covert channel for G.723.1 speech codec is exhibited in the work, which is based on the codec's "dual rate" characteristic. As aforementioned, while G.723.1 speech frames are transmitted in RTP packets, the PT field in RTP header for each packet is exactly equivalent to four, indicating the transmitted media are G.723.1 speech frames. However, more specific speech frame bitrate (6.3 or 5.3 kb/s) is still unassigned (although one can still have a more specific assignment through SDP, which is not a typical practice). At the receiving end, speech decoder parses the two flag bits and passes them to corresponding decoding algorithms. Generally speaking, proper bitrate is selected in practice according to network condition. In this way, in a stable network environment, this characteristic can also be exploited to embed secret message as long as the decoder has the right decoding.

Two strategies are presented in our proposed method based on above features, bitrate downgrading (BD) strategy and bitrate switching (BS) strategy. Under the condition of right decoding, BD strategy replaces selected 6.3 kb/s frames (24 bytes) by 5.3 kb/s frames (20 bytes), leaving remaining 4 bytes space for steganography. BS strategy utilizes the fact that there is no restriction on how the two bitrates intermixed in speech streams [12]. So, the bitrate could be discretionarily changed, thus the change manner can be encoded to represent secret message bits for steganographic embedding. Both of these strategies keep PT value unchanged. As the two bitrates are mandatory requirement within G.723.1 codec, no extra speech codec support is needed either.

4.1 BD steganography strategy

In G.723.1 speech codec, active voice is mainly encoded into 6.3 kb/s and 5.3 kb/s frame types, and the two types have 24 and 20 bytes, respectively. When speech frames are transmitted in an encrypted network environment (via SRTP protocol which is an extension of RTP), which is a common practice, the specific frame types cannot be inferred from RTP header as aforementioned. Therefore, 5.3 kb/s frames can also be utilized to substitute the 6.3 kb/s frames and fill the remained 4 bytes extra space with secret message bits and keep the payload size unchanged. This strategy introduces a little speech quality distortion caused by frame bitrate downgrading, which will be evaluated later. What's more, eavesdroppers cannot discover any abnormity out of intermediate network nodes.

Figure 2 gives an example of secret message embedding using BD strategy for one speech frame. At most 4 bytes of secret message is allowed to interpolate into a 20 bytes 5.3 kb/s frame by a secret key indicating the embedding positions to form a 24 bytes frame. The secret key is shared by the sender and receiver to enhance the security of steganography. The key space size S_{key} is figured out according to Eq. (3) (retain the order of bits), which is equivalent to 118 bits of secret key space size. At the receiver end, this key is needed for secret message retrieving and the right decoding of speech frames by the 5.3 kb/s decoder of G.723.1 codec. Note that the two flag bits are located in the first byte, so only the rest 23 bytes are selected for secret bits insertion.

$$S_{\text{key}} = C_{23\times 8}^{4\times 8} = \frac{(184)!}{(184-32)!32!} \approx 6.46\times 10^{35} \approx 1.94\times 2^{118} \tag{3}$$

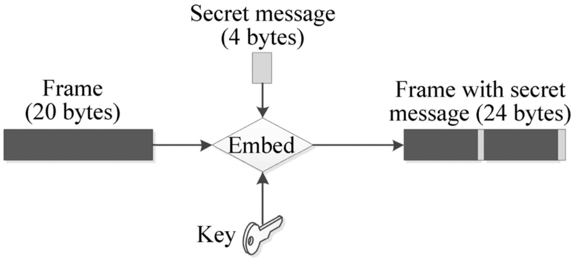

Fig. 2 BD embedding in one speech frame

As a result of steganography, there exist totally four different types of frames: 6.3 kb/s frames, 5.3 kb/s frames, SID frames and 5.3 kb/s frames with secret message, so the reserved state of flag bits in Table 2 can be employed to indicate the last frame type. Table 3 gives the specification of these types and their corresponding sizes, where the first three can be decoded by normal G.723.1 codec, but the last type is often neglected by most speech codec implementation. Therefore, the decoder needs to be modified to identify this characteristic.

Figure 3 gives the flow chart of the embedding and retrieving procedures of BD strategy, where the sender

Table 3 Flag bits for normal and covert communication

Flag bit	Frame type	Size/bytes
00	6.3 kb/s	24
01	5.3 kb/s	20
10	SID	4
11	5.3 kb/s with steganography	24

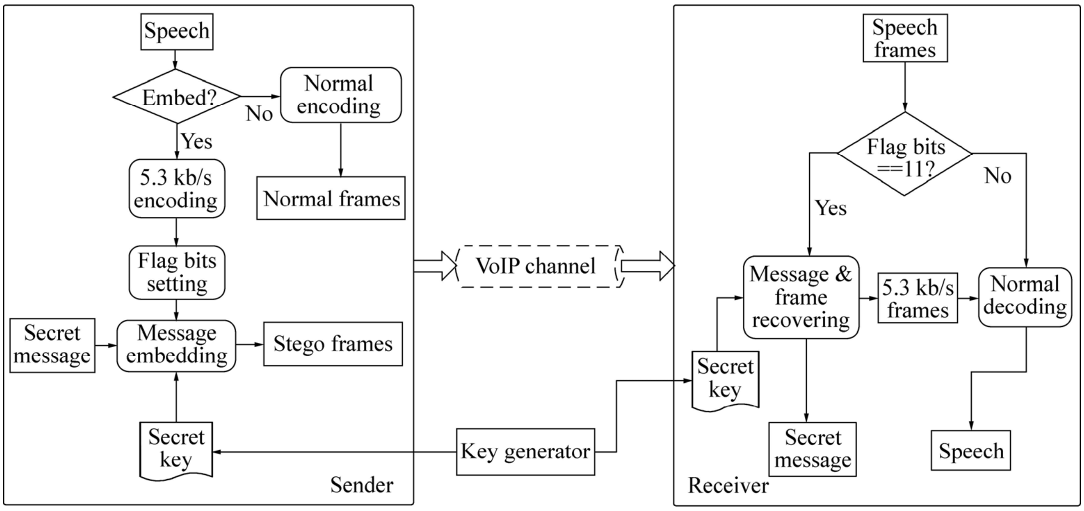

Fig. 3 Flow chart of steganography of BD strategy

and receiver share the same secret key generated from a key generator to indicate the embedding and retrieving of secret message as well as the assembling of received speech frames. The two flag bits need to be set before embedding and be parsed prior to the frame decoding and message retrieving procedure at the receiving end.

4.2 BS steganography strategy

In G.723.1 standard [12], it is said to be possible to switch between the two bitrates of it at any 30 ms frame (5.3 kb/s and 6.3 kb/s frames) boundary. This can be treated as redundant information, thus is available for steganography. The alteration manner of frame types (only consider 5.3 kb/s frame and 6.3 kb/s frame) is coded to represent secret message bits for BS strategy.

The transmitting information in VoIP channel can be regarded as speech frame stream. Let l_i be the frame state of the i-th frame defined in Eq. (4) where $s(i)$ is the size of it. Accordingly, secret message bit m_p is defined in Eq. (5) where $f_k \in F=\{f_1, f_2, \cdots, f_k, \cdots\}$. F is a set of encoding functions and f_k is a k-ary function with k independent variables l_o, $l_{\mathrm{o}+1}$, $\cdots$, $l_{\mathrm{o}+k-1}$. Here, k parameter indicates the states of k selected frames which come from one or more RTP payload, and l_o represents the state of the first of these k frames.

$$l_i=\begin{cases}0, \text{ if } s(i)=24\text{ bytes}\\ 1, \text{ if } s(i)=20\text{ bytes}\\ \text{Undefined, else}\end{cases} \tag{4}$$

$$m_\mathrm{p}=f_k(l_\mathrm{o}, \cdots, l_{\mathrm{o}+k-1}),\ m_\mathrm{p}\in\{0,1\} \tag{5}$$

To enhance the reliability of decoding at the receiving end, a state set $S_f=S_{f0}\cup S_{f1}$ ($S_{f0}\cap S_{f1}=\varnothing$) is defined for f_k. If current frame states sequence received $S_i=(l'_\mathrm{o}, l'_{\mathrm{o}+1}, \cdots, l'_{\mathrm{o}+k-1})$ belongs to S_{f0} or S_{f1}, the decoded message bit $m'_\mathrm{p}=f_k(S_i)$ may be 1 or 0. Otherwise, there is no BS steganography in these frames. Accordingly, when there is no information to hide, if sequence (l_o, $\cdots$, $l_{\mathrm{o}+k-1}$) happened to belong to S_f, the states of some frames are deliberately changed to avoid retrieving error. For function f_k, the selection of frames and value k are all not fixed, the way of bitrate change of normal frames to ensure reliable covert communication also depends.

Note that the parametric frames of f_k may not be consecutive. Practical embedding procedure depends on embedding function E_f (incl. embedding position) and embedding rate a, so Eq. (5) may be expanded to Eq. (6), which is a special case of Eq. (6) when a=100% and the frame set for f_k is consecutive. Parameter value k affects the encoding efficiency and the performance of steganography. And it also has something to do with local network condition. Assuming a=100% and the frames in one embedding unit (k frames) are consecutive, thus the lowest embedding capacity is $1/k$ bit per frame when S_{f0} and S_{f1} both have equally one element.

$$m_\mathrm{p}=E_f(f_k(l_\mathrm{o}, \cdots, l_{\mathrm{o}+k-1}), a) \tag{6}$$

The bit encoding of this strategy for message embedding is simply depicted in Fig. 4, where one exemplary encoding method is applied. In one bit per frame case 6.3 kb/s and 5.3 kb/s frame denotes bit 0 and 1, respectively. In $1/k$ bit per frame (k>1) case two groups of k frames are selected from S_{f0} and S_{f1} for one secret message bit, the remainders are normal frames (without embedding). The receiving end needs the information of encoding scheme to have the right retrieve of secret message bit. Here, only one of 2^k-1 usable encoding schemes is selected, that is $1/(2^k-1)$. Furthermore, the selection of encoding scheme could be dynamically changed to enhance the security.

In one bit coding, the embedding capacity is 1 bit/frame, thus extra state was supposed to be needed to indicate normal frame state. However, due to the high distortion it introduces when $k=1$, this case is never

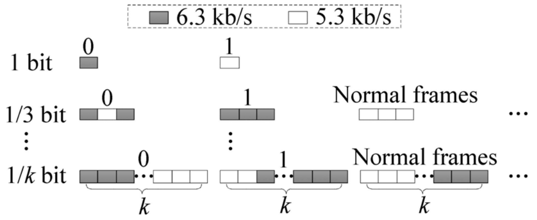

Fig. 4 Encoding frame states of BS strategy

used in practice. To enhance the reliability of BS strategy, more frames can be chosen and more states can be selected in k frames for steganography. In order to resist possible stream analysis, BS strategy is fit for covert communication with little information embedded, such as key exchange and session signal negotiation.

Figure 5 depicts the flow chart of BS strategy where the frame encoding scheme is shared by the sender and receiver. If embedding is not allowed, the normal frame bitrates need to be adjusted so as not to be treated as stego frames. The frame bitrates are determined before speech encoding in the embedding procedure so as to call corresponding speech encoder. The receiver needs to extract the frame states and judge if S_i belongs to S_f (S_{f1} or S_{f2}) or not to retrieve secret information and then decode received frames.

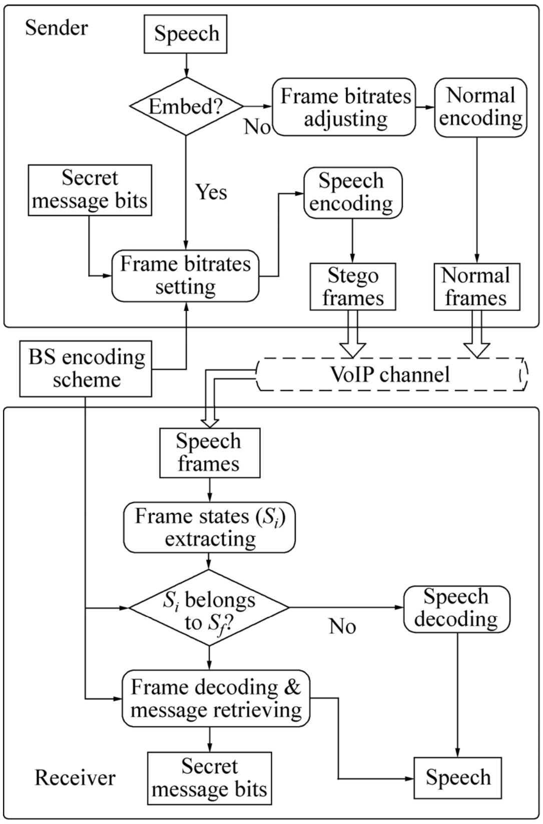

Fig. 5 Flow chart of steganography of BS strategy

5 Implementation and evaluation

In order to apply proposed method to practical VoIP system, G.723.1 codec must be supported and is the only necessary speech codec. The proposed method is implemented on our StegVoIP [8] system, which is a covert communication solution based on VoIP and customized for speech steganography. The experiments are performed in a LAN environment with two computers both of which are made up of Pentium Dual-core CPU E5200 2.50 GHz, 2.00 GB RAM and 32-bit operating system of Windows® 7.

5.1 Steganography model based on VoIP

A reliable covert transmission protocol is constructed and implemented in StegVoIP based on RTP covert channel [3, 15] in application layer to solve packet loss problem caused by the connectionless service of UDP protocol. This function is optional if packet-loss insensitive information is embedded and transmitted, such as very low bitrate speech [16]. For proposed method, it is necessary especially when selected frames for BS strategy are situated in different RTP packets.

Figure 6 depicts the simplified steganographic model

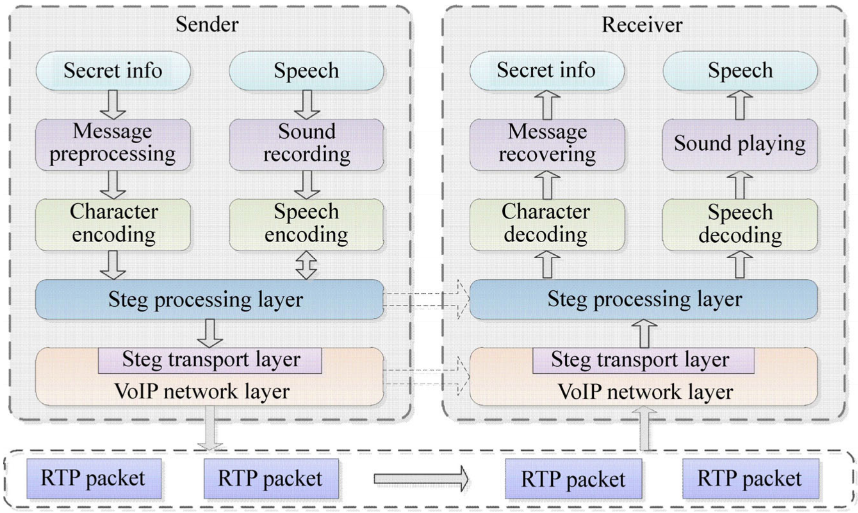

Fig. 6 Steganography model for StegVoIP

of StegVoIP, where a three-level protocol [15] is employed to ensure the completion of steganography algorithm, reliable covert communication and the independence of steganography from the underlying VoIP environment. The steganographic embedding and retrieving procedures are performed in the stego processing layer. The stego transport layer is used to detect and retransmit lost secret message in RTP packet and transmit extra information, e.g., the indication of the proposed two strategies.

5.2 Secret message encoding

To verify proposed method, instant messages are employed as the secret information to be embedded. Due to the restriction of speech quality distortion and inherent nature of compressed speech (low redundancy), embedding capacity in low bitrate speech is generally in a low level. Therefore, to decrease the secret message size beforehand is necessary. However, commonly used data compression algorithm is always complicated thus time-consuming, which is unsuitable for real-time VoIP communication. And for covert communication in VoIP, transmitting short pieces of information or brief messages is more acceptable and reliable than large-sized multimedia and packet-loss-sensitive data.

For this reason, the character encoding scheme of secret messages is redefined in Table 4 by reducing the number of encoding bits. Here, only some commonly used characters are defined, which are sufficient for transmission of understandable messages. In the table, a 4−8 bits isometric extension encoding method (14/32) is employed to encode the most commonly used characters according to the probability of occurrence of English letters tested by NORVIG [17].

In this coding method, some seldom used and unrecognized characters are neglected and will be transcoded into ‘?’. Equation (7) gives the expression to compute the average code length L for these letters, where p_i is the probability of occurrence of letter i and l_i is the coding length of it. n represents the number of letters calculated. It is deduced from Eq. (7) that L is approximately equal to 4.5 bits according to the letter statistics in Ref. [17], which is a big degradation compared with 8-bit length ASCII coding method and dramatically saves the steganography bandwidth for covert communication.

$$L=\sum\nolimits_{i=1}^{n} p_i l_i \qquad (7)$$

Table 4 Character encoding scheme for secret messages

Character	Code	Character	Code	Character	Code
E & e	0010	F & f	00000010	2	00010010
T & t	0011	P & p	00000011	3	00010011
A & a	0100	G & g	00000100	4	00010100
O & o	0101	W & w	00000101	5	00010101
I & i	0110	Y & y	00000110	6	00010110
N & n	0111	B & b	00000111	7	00010111
S & s	1000	V & v	00001000	8	00011000
R & r	1001	K & k	00001001	9	00011001
H & h	1010	X & x	00001010	Space	00011010
L & l	1011	J & j	00001011	,	00011011
D & d	1100	Q & q	00001100	.	00011100
C & c	1101	Z & z	00001101	‘	00011101
U & u	1110	0	00001110	?	00011110
M & m	1111	1	00001111	Padding	0000(0000)

Note: 0000 is used for padding, and 00000000 is the same meaning as NULL.

5.3 Speech quality evaluation

To evaluate the speech quality degradation caused by steganography, commonly used perceptual evaluation speech quality (PESQ) [18] criterion is selected, which is an ITU-T Recommendation to objectively evaluate speech quality with a value between −0.5 and 4.5. Higher PESQ value means better speech quality, that is, less speech quality distortion is introduced. PESQ needs two speech segments for the test, one is used as reference speech and the other is the degraded version. Here, the reference speech signal is the primitive sampled PCM signal and the degraded signal is the speech recovered from G.723.1 codec through steganography or not. Test speech samples are divided to four types: Chinese man (CM), Chinese woman (CW), English man (EM) and English woman (EW), which are all collected from daily conversations and newspaper reports.

In BD strategy, the resulted speech quality degradation is caused by the decrease of speech encoding output bits from 24 bytes (6.3 kb/s frame) to 20 bytes (5.3 kb/s frame), that is, the change of speech coding algorithm in G.723.1 codec. Figure 7 compares the PESQ values of BD strategy, LSB substitution, LSD steganography (its BS strategy) [4] and the decoding output of G.723.1 6.3 kb/s coding algorithm with approximately equal embedding bits, where a=100% and

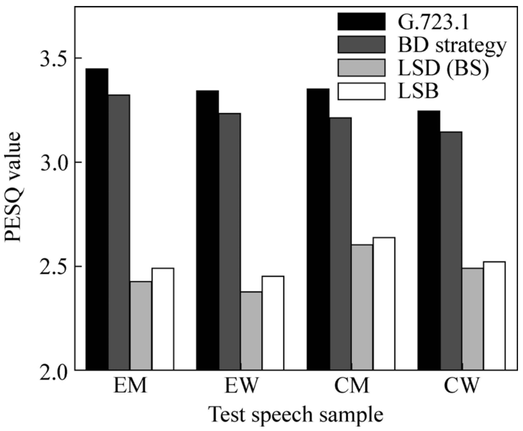

Fig. 7 PESQ comparison of three steganography methods and G723.1

some daily used dialogues are employed as embedded secret message.

The embedding space for BD strategy is derived from the redundant information caused by different output encoding bits of different speech encoding algorithms. Differing from other frame-modification based steganography methods in the figure, the generation of embedding space of BD strategy conforms to the essential features of speech. As a result, less distortion is introduced. It is seen from Fig. 7 that BD strategy has the highest PESQ values compared with other steganography methods, which corroborates its effectiveness. Note that the other two steganographic methods results in bad speech quality, that's due to much more embedded bits (about 32 bits to compare with BD strategy) than their usual application.

To evaluate how the change rate of frame bitrate affects speech quality, the pair of states $f_k(0, 0, \cdots, 0)$ and $f_k(1, 1, \cdots, 1)$ are selected for secret bits encoding and they are consecutive frames in VoIP stream. According to the principle of BS strategy, the embedding capacity varies from 1 bit/frame to 1/*k* bit/frame, which is a very low embedding level compared with other steganography methods aforementioned. However, due to the persistence of VoIP streams, this strategy is still useful and advantageous. Although it is possible to switch between the two bitrates at any frame boundary, the linear prediction coder (LPC) in the speech codec employs the information of previous frames. As a result, the frequent bitrate switching may affect decoded speech quality a lot.

Figure 8 depicts the PESQ values of BS based steganography from *k*=1 to *k*=33 of four speech types, which corresponds to embedding capacity from 1 bit/frame (33 b/s) to 1/33 b/frame (1 b/s). It is seen that when *k* value is less than 10 (high change rate of bitrate), the speech quality distortion is very high, and so it is not suitable for embedding in this situation. For BS strategy with *k* values greater than 10, the PESQ values exhibits less distortion, which can be utilized for steganography, especially when speech payload modification is restrained but frame bitrate adjusting is allowed.

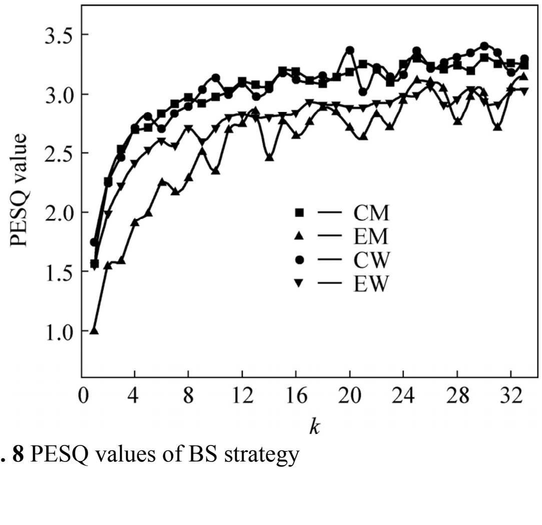

Fig. 8 PESQ values of BS strategy

5.4 Steganography delay

In StegVoIP environment, BD and BS strategies are disposed on users' end rather than intermediate network nodes. So, the introduced communication latency mainly depends on the processing delay of embedding and extracting operations as well as the waiting delay of RTP packaging. The waiting delay is caused by speech samples collection introduced by RTP packaging, which is about 37.5 ms per frame for G.723.1 speech codec [12] and is not needed to be considered for steganography. The other delays have nothing to do with the network condition, so the experimentation is only performed on LAN environment for simplicity. Compared with the method in Ref. [11], only a couple of encoding and decoding operations need to be performed, which decreases the processing delay a lot.

Table 5 shows the evaluation results, where the processing delay is broke up into secret message coding delay and steganographic operation delay. As the delays are affected by system hardware and implementation algorithm, the delay of G.723.1 5.3 kb/s codec is presented here for reference.

Table 5 Processing delay of proposed method

Operation	Delay/(μs·frame^{-1})
Message encoding (BD)	1−10
Message encoding (BS)	2−3
Message decoding (BD)	7−26
Message decoding (BS)	1−7
Message embedding (BD)	< 200
Message embedding (BS)	< 800
Message retrieving (BD)	< 200
Message retrieving (BS)	< 1
G723.1 5.3 kb/s encoding	< 800
G723.1 5.3 kb/s decoding	< 80

The experimental results show that secret message bits and the number of coding frames (*k* for BS strategy from 1 to 33) have little effect on the processing delay. In other words, compared with the 37.5 ms algorithmic delay of G.723.1 and the 150 ms end-to-end one-way latency limitation for speech communication [19], the processing delay for proposed method can be neglected.

5.5 Discussion on potential steganalysis

HUANG et al [5] proposed a steganalysis method based on second statistics to detect the presence of steganography and estimate the secret message length. LU et al [20] discovered the characteristic of the pulse position parameter of normal G.723.1 frames, while many steganography methods destroyed it. LI et al [21] constructed the statistical model of the speech features affected by steganography and successfully detect the quantization index sequence (QIM) based steganography. However, these steganalysis methods are all aimed at steganography based on bits modification of speech frames, which are ineffective for proposed method.

To the best of the authors' knowledge, there has been no effective steganalysis method to detect steganography based on bitrate-modification so far. Similar to that mentioned in Ref. [11], it is assumed that BD strategy is always performed in SRTP environment for privacy protection. In that condition, steganalyzer is unable to detect the existence of steganography at intermediate node of the network. Moreover, in an open network environment, BS strategy is able to be decoded correctly by normal speech codec. The frequent switching of bitrate may draw eavesdropper's attention and affect the predictive algorithm of speech codec. That can be alleviated or avoided by increasing the *k* value. Moreover, it is useful at the beginning or intervals of speech conversation for transmission of bits of synchronization information as less human's voice is generated therein [10]. For a network administrator, it is useful to avoid possible information leakage by strict active management when these two strategies are known.

6 Conclusions and future work

1) A steganography method based on multi-rate speech frames is proposed. Differing from previous algorithms, the proposed method does not modify the speech coding algorithm itself, which decreases introduced distortion largely. Two embedding strategies are presented with experimental results here, and in order to decrease the secret message size, a message encoding method is given to transmit more secret information. The evaluation results show this method (BD strategy) outperforms previous steganography methods on imperceptibility and introduces negligible delays.

2) As a hypothesis, BS strategy may be well suitable for silent frames in communication to increase its imperceptibility. To combine BS strategy with steganography in Ref. [10] or other possible steganography methods will be our future work. And the practical application of them for network information exchange seamlessly with RTP protocol will be investigated further. To increase the applicability, BD and BS strategies in commonly used VoIP systems using their build-in speech codecs will also be studied.

References

[1] KEROMYTIS A. Voice-over-IP security: Research and practice [J]. IEEE Security Privacy, 2010, 8(2): 76−78.

[2] TIAN Hui, ZHOU Ke, FENG Dan. Dynamic matrix encoding strategy for voice-over-IP steganography [J]. Journal of Central South University of Technology, 2010, 17(6): 1285−1292.

[3] YING Li-zhi, HUANG Yong-feng, YUAN Jian, LINDA B. A novel covert timing channel based on RTP/RTCP [J]. Chinese Journal of Electronics, 2012, 21(4): 711−714.

[4] ZHOU Ke, LIU Jin, TIAN Hui, LI Chun-hua. State-based steganography in low bit rate speech [C]// Proceedings of the 20th ACM International Conference on Multimedia. Nara: ACM, 2012:1109−1112.

[5] HUANG Y, TANG S, BAO C. Steganalysis of compressed speech to detect covert voice over Internet protocol channels [J]. IET Information Security, 2011, 5(1): 26−32.

[6] ZANDER S, ARMITAGE G, BRANCH P. Covert channels and countermeasures in computer network protocols [J]. IEEE Communications Magazine, 2007, 45(12): 136−142.

[7] TIAN Hui, JIANG Hong, ZHOU Ke, DAN Feng. Transparency-orientated encoding strategies for voice-over-IP steganography [J]. The Computer Journal, 2012, 55(6): 702−716.

[8] TIAN Hui, ZHOU Ke, LU Jing. A VoIP-based covert communication scheme using compounded pseudorandom m sequence [J]. International Journal of Advancements in Computing Technology, 2012, 4(1): 223−230.

[9] LIU Jin, ZHOU Ke, TIAN Hui, LI Chun-hua. Efficient least-significant-bits steganography for VoIP [J]. International Journal of Advancements in Computing Technology, 2012, 4(10): 297−305.

[10] HUANG Y, TANG S, YUAN J. Steganography in inactive frames of VoIP streams encoded by source codec [J]. IEEE Transactions on Information Forensics and Security, 2011, 6(2): 296−306.

[11] MAZURCZYK W, SZAGA P, SZCZYPIORSKI K. Using transcoding for hidden communication in IP telephony [J]. Multimedia Tools and Applications, 2011: 1−27.

[12] Telecommunication standardization sector of ITU. ITU-T G.723.1-Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s [R]. Geneva, Switzerland: International Telecommunication Union, 2006.

[13] SIMMONS G. The prisoner's problem and the subliminal channel [C]// Proceedings of Crypto: Plenum 1984: 51−67.

[14] SCHULZRINNE H, CASNER S. RTP profile for audio and video conferences with minimal control [R]. IETF Network Working Group, 2003.

[15] XIAO Bo, HUANG Yong-feng. Modeling and optimizing of the information hiding communication system over streaming media [J]. Journal of Xidian University, 2008, 35(3): 554−558. (in Chinese)

[16] WANG Chungyi, WU Quincy. Information hiding in real-time VoIP streams [C]// Proceedings of the 9th IEEE International Symposium on Multimedia. Taiwan: IEEE, 2007: 255−262.

[17] NORVIG P. English letter frequency counts: Mayzner revisited [EB/OL]. [2012−12−20]. http://norvig.com/mayzner.html.

[18] Telecommunication standardization sector of ITU. 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 [R]. Geneva, Switzerland: International Telecommunication Union, 2001.

[19] Telecommunication standardization sector of ITU. ITU-T G.114-One-way transmission time [R]. Geneva, Switzerland: International Telecommunication Union, 2003.

[20] LU Ji-cang, HUANG Yong-feng, LIU Fen-lin, LUO Xiang-yang. Pulse position checking-based steganalysis of G.723.1 compressed speech in VoIP [J]. International Journal Multimedia Intelligence and Security. 2011, 2(3/4): 225−237.

[21] LI Song-bin, TAO Huai-zhou, HUANG Yong-feng. Detection of QIM steganography in G.723.1 bit stream based on quantization index sequence analysis [J]. Journal of Zhejiang University Science C. 2012, 13(8): 624−634.

(Edited by DENG Lü-xiang)