Audio steganography with AES for real-time covert Voice over Internet Protocol communications

Audio steganography with AES for real-time covert voice over internet protocol communications

TANG Shanyu[1*], JIANG Yijing[1], ZHANG Liping[1] & ZHOU Zhangbing[2]

[1] *School of Computer Science, China University of Geosciences, Wuhan 430074, China*

[2] *School of Information Engineering, China University of Geosciences, Beijing 100083, China*

Abstract As a popular real-time service on the Internet, Voice over Internet Protocol (VoIP) communication attracts more and more attention from the researchers in the information security field. In this study, we proposed a VoIP steganographic algorithm with variable embedding capacities, incorporating AES and key distribution, to realize a real-time covert VoIP communication. The covert communication system was implemented by embedding a secret message encrypted with symmetric cryptography AES-128 into audio signals encoded by PCM codec. At the beginning of each VoIP call, a symmetric session key (SK) was assigned to the receiver with a Session Initiation Protocol-based authentication method. The secret message was encrypted and then embedded into audio packets with different embedding algorithms before sending them, so as to meet the real-time requirements of VoIP communications. For each audio packet, the embedding capacity was calculated according to the specific embedding algorithm used. The encryption and embedding processes were almost synchronized. The time cost of encryption was so short that it could be ignored. As a result of AES-based steganography, observers could not detect the hidden message using simple statistical analysis. At the receiving end, the corresponding algorithm along with the SK was employed to retrieve the original secret message from the audio signals. Performance evaluation with state-of-the-art network equipment and security tests conducted using the Mann-Whitney-Wilcoxon method indicated that the proposed steganographic algorithm is secure, effective, and robust.

Keywords VoIP, steganography, AES, covert communication, Mann-Whitney-Wilcoxon

*Corresponding author (email: shanyu.tang@gmail.com)

1 Introduction

Voice over Internet Protocol (VoIP) communication is one of the most popular real-time services on the Internet. VoIP has more advantages than traditional telephony, since the Internet allows VoIP to provide low-cost, high-reliability, and global services. VoIP streams often have a highly redundant representation, which usually permits the addition of significantly large amount of secret data by means of simple and subtle modifications that preserve the perceptual content of the underlying cover object. With the increasing percentage of VoIP streams in all of the Internet traffic, VoIP is considered to be a better cover object for information hiding compared with “static” cover objects such as text files, image files, and audio files. Besides, VoIP connection is usually very short, and so it is unlikely for attackers to detect the hidden data within VoIP streams. Their real-time characteristics may be used to improve the security of the hidden data embedded in VoIP “dynamic” streams.

Steganography is a method of embedding secret data into a cover object, which should not cause unacceptable distortion and arouse observers’ attention. Both steganography and encryption technology keep the confidentiality of the secret data, but there are significant differences in many aspects between them. Encryption technology only protects the content of the secret data, making them unreadable. Thus, unauthorized users can know the existence, except the specific details about the secret data. Steganography hides the existence of the secret data, such that unauthorized users know neither the existence of the secret data nor the details of it.

Steganography is one of the most important areas of information security, becoming more and more flourishing to apply in various fields. For example, military communication systems usually need to be at higher security levels. They require not only encrypting the messages exchanged, but also hiding the existence of the messages, which means attackers even cannot perceive the existence of the messages. For protecting the intellectual property of digital products, merchants often embed their trade mark or unique logo into digital products with steganography. There are also some other applications of steganography.

In early steganography literature, steganography was widely used in image [1-3], audio [4-5], and video [6] files. For image steganography, the common method was to modify the least significant bit (LSB) of pixels in an image using LSB-based methods. Since Human Visual System (HVS) is not so sensitive, the differences between the original image and the image with a secret message are imperceptible through human eyes. Audio steganography was used to embed a secret message into an AU, WAV, or MP3 audio file. It is generally recognized that audio steganography is more challenging than image steganography for the wider dynamic range of Human Auditory System (HAS) in comparison with HVS. The basic LSB steganography was also widely used in audio and video files. With the rapid development of the Internet applications, steganography has been gradually applied to the Internet real-time data. The real-time requirements of VoIP communication provide basic security for the system, but it does not allow too many operations, making it difficult to add more operations (e.g., security measures) to improve the security level.

There have been some efforts to study VoIP steganography, which has potential practical applications in industries and the military. VoIP steganography can be divided into two categories. One is embedding secret data into some free or unused fields of TCP/IP protocol headers [7]. However, there is an underlying problem. When passing an intelligent router, VoIP packets with modifications in their protocol headers would be eliminated or discarded. The other one is hiding secret data into the payload of VoIP packets, which is a method to modify audio packets to embed secret data in different processing stages. For instance, Huang et al. [8] embedded secret data according to the characteristics of the coding algorithms at the coding phase. Codecs used were G.729, G.726, and G.723. Miao et al. [9] analyzed the character of audio, and implemented different embedding approaches in active frames and inactive frames.

Dittmann et al. [10] suggested a design of steganography for VoIP. Since the design substituted the bits of secret data for the least significant bits of the cover audio, the design was easy to be detected by simple steganalysis. They also suggested that secret data can be encrypted before embedding.

Tian et al. [11] proposed a covert communication model based on LSB steganography in VoIP, which simply encrypted secret data before embedding them into voice samples coded by G.729a codec. They claimed that the model could protect the security of the secret data in a short term and produce short latency. However, the encryption algorithm is not a standard algorithm, and the security of the encryption algorithm remains to be proved.

Wang and Wu [12] described a design of real-time speech hiding for G.711 codec, which was implemented in Linphone. They compressed the secret speech with Speex, before embedding it into the least significant bit of every two samples. But they did not consider the quality of stego-speech and failed to address the security problems such as encryption, authentication, key distribution, etc.

The encryption of secret data and the authentication between the communicating parties cost extra time before embedding the secret data, but if the costs were relatively lower, they would not affect the voice quality of VoIP. On the other hand, session key (SK) distribution could be performed ahead of each VoIP call, so that it would not have an impact on normal VoIP communications.

Based on the above analysis, in this study, we proposed a VoIP steganographic algorithm with variable embedding intervals incorporating Advanced Encryption Standard (AES) and key distribution to realize a real-time covert communication. The covert communication system was implemented by embedding a secret message encrypted with symmetric cryptography AES-128 into audio signals encoded by PCM codec. The secret message was encrypted and embedded with different algorithms before sending audio packets to meet the real-time requirements of VoIP communications. At the receiver end, the corresponding algorithm was employed to retrieve the secret message, and then to decrypt it to get the original message using the same SK. A large group of female and male speech samples were employed to evaluate the performance and security of the proposed steganographic algorithm and its impact on VoIP communications.

This article is organized as follows. Section 2 introduces our proposed real-time covert VoIP communication system with an underlying steganographic algorithm consisting of key distribution, encryption, embedding, and extraction. In Section 3, the experimental setup and test device are described in detail. The experimental results, performance evaluation, and security analysis of the proposed covert VoIP communication system using state-of-the-art network equipment Digital Speech Level Analyzer (DSLA) made in England are presented in Section 4. Finally, the conclusions of the article are given in Section 5.

2 Proposed Real-time Covert VoIP Communication System

2.1 Covert VoIP communication framework

The proposed covert VoIP communication system was based on the UDP protocol, which is used for connectionless and unreliable delivery. The system was implemented by embedding a secret message encrypted with symmetric cryptography AES-128 into audio signals encoded by PCM. At the receiving end, the secret message was retrieved and decrypted with the same SK, which was distributed before the communication session began.

Figure 1 depicts the covert VoIP communication model based on steganography and cryptography. At the beginning of each VoIP call, a symmetric SK was randomly generated by user A and assigned to the receiver (user B) with a Session Initiation Protocol-based authentication method. User A wants to send a secret message to user B through the VoIP channel. M stands for the secret message, M’ is the ciphertext, and E and D denote the encryption and decryption processes, respectively. User A encrypts the secret message M to M’, embeds M’ into the original cover-speech, and then sends the stego-speech to user B. Having received the stego-speech, user B retrieves the encrypted secret message, and decrypts M’ to get the original secret message with the same SK.

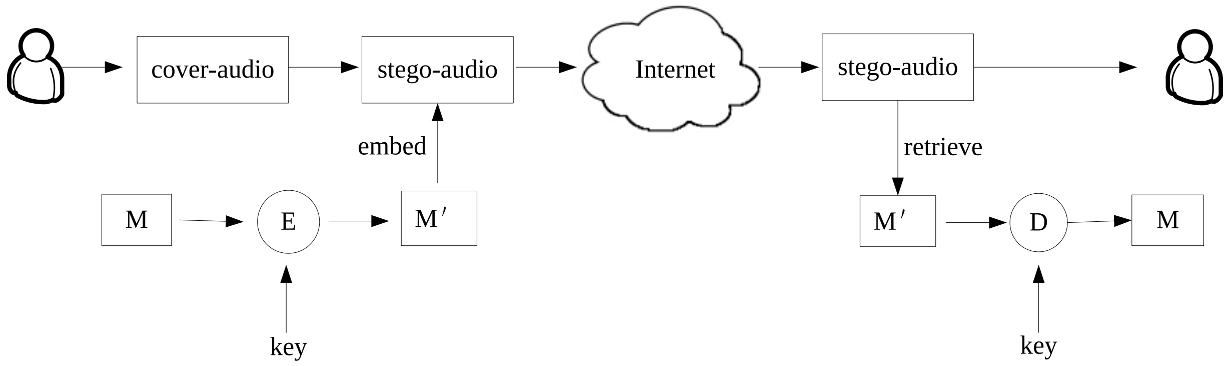


Figure 1 Covert VoIP communication model.

Since the VoIP system is a connectionless and unreliable delivery system, it leads to packet loss inevitably. To measure the packet loss ratio, we as user A sent a string “abcdefghijkl......” to the other communicating party (user B), by embedding one character in the least significant bits of audio packets. At the receiver end, we retrieved the string to determine the packet loss ratio. The test was implemented in a local area network, and the packet loss ratio was determined to be less than 1%, which is acceptable for VoIP communications of high quality.

2.2 Key distribution

The Session Initiation Protocol (SIP) is commonly used for establishing VoIP calls. SIP is an application-layer protocol which handles all the signaling requirements of the VoIP session, including initiating, managing, and terminating multimedia sessions across networks. Some SIP authentication key agreement protocols have been proposed to improve the security of SIP-based communications. Wang and Liu [13] proposed a key agreement scheme for VoIP communications. The scheme could achieve dynamic key changes during the session initiation. Yu et al. [14] proposed an identity-based mechanism capable of providing end-to-end identity authentication and key agreement for enhancing SIP security.

There are three communication ways in SIP: the direct call, the proxy server, and the redirect server. In the proposed covert VoIP communication system, the key delivery was implemented over a secret channel. The communication mode in the SIP protocol was the direct call, which was designed for initializing the interactive communication session between two communicating parties. The key distribution was achieved using an effective authenticated key agreement scheme for SIP based on the concept of identity-based encryption. The public key could be the users’ some public information. User A and User B had their own private key (*PR*) and public key (*PU*). User A could get user B’s public key through a certain way. Then user A generated a symmetric SK randomly, encrypted the SK with A’s private key and B’s public key in succession, and sent the encryption result to user B. Having received the result, user B decrypted the result with B’s private key and A’s public key successively to get the SK.

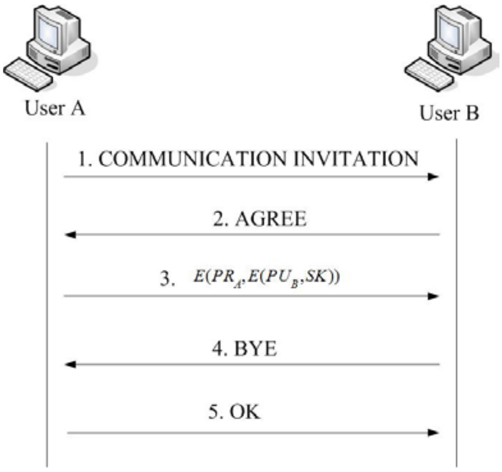


Figure 2 Initialization process of VoIP communication.

Figure 2 shows the process of initializing the communication between user A and user B. Firstly, user A sends "COMMUNICATION INVITATION" to invite B to join a session. If user B accepts the invitation, she or he returns "AGREE". Upon receipt of confirmation, user A generates a symmetric SK (*SK*) randomly, and then uses an encryption method $E(PR_A, E(PU_B, SK))$ to send the *SK* to B. User B decrypts SK to get the SK. If one communicating party wants to end the call, she or he could send "BYE", and the other party would reply "OK" to end the communication.

In each call, the SK was updated to avoid the key leakage problem, so as to improve the security of the covert VoIP communication between the communicating parties.

2.3 Encryption

Advanced Encryption Standard is a specification of standard encryption established by the U.S. NIST in 2001. It is based on Rijndael cipher algorithm [15], which is used widely. For AES-128, the key attackers need to break AES-128 with the computational complexity of $2^{126.1}$, indicating that the attacks are computationally infeasible. Considering the real-time requirements of VoIP communications, AES-128 was used in the proposed covert VoIP communication system, because AES-128 is one of the fastest encryption algorithms so far.

Table 1 The parameters of AES-128 used

Key Length (words/bytes)	Block Size (words/bytes)	Number of Rounds	Key Length in each Round (words/bytes)
4/16	4/16	10	4/16

In the AES algorithm, the lengths of the input block and the output block are 128 bits, which can be expressed as 16 bytes. Table 1 lists the parameters in the implementation of AES-128.

For its cipher and inverse cipher, the AES algorithm uses a round function that is composed of four different byte-oriented transformations: byte substitution (expressed as SubBytes()), shifting rows (expressed as ShiftRows()), mixing data in each column (expressed as MixColumns()), and adding Round Key (expressed as AddRoundKey()). Figure 3 describes the AES encryption pseudo code [16] used in the covert VoIP communication system. Nb is the block size in words and Nr is the number of rounds in words.

```
chiper(byte in [4*Nb],byte out [4*Nb],word w[Nb*(Nr+1)])
begin
    byte state[4, Nb]
    state = in
    AddRoundKey(state, w[0,Nb-1])
    for round =1 step 1 to Nr-1
        SubBytes(state)
        ShiftRows(state)
        MixCloumns(state)
        AddRoundKey(state, w[Nr*Nb, (Nr+1)*Nb-1])
    end for
    SubBytes(state)
    ShiftRows(state)
    AddRoundKey(state, w[Nr*Nb, (Nr+1)*Nb-1])
    out = state
end
```

Figure 3 Pseudo code of the cipher.

In our covert VoIP communication system, the AES-128 encryption algorithm was used to encrypt a secret message prior to embedding taking place. Before embedding the secret message, the whole length of the secret message was calculated and expressed as LoM. The LoM value was embedded into 16-bytes cover audio packets, and the range of LoM was between 0 and 2^{16}-1. Since it was necessary to minimize the time that encryption of the secret message costs to meet the real-time communication requirements, the size of the secret message embedded into each audio packet should be an integral multiple of 16 bytes and smaller than the size of the audio packet. As a result of that, AES-128 algorithm divided the plaintext into groups of 16 bytes, so each encryption time was less than 0.1 ms, which is acceptable for the real-time VoIP communication.

2.4 Embedding

The different embedding methods employed in the covert VoIP communication system were based on our newly designed data embedding algorithms with variable embedding capacities, which were achieved by varying embedding location intervals. Let us set a parameter R as the interval for embedding the bit stream of a secret message into each byte of audio packets. LLoM is defined as the length of the rest of the secret message that are embedded into the last packet. The various sizes of the secret message embedded into different audio packets are defined by

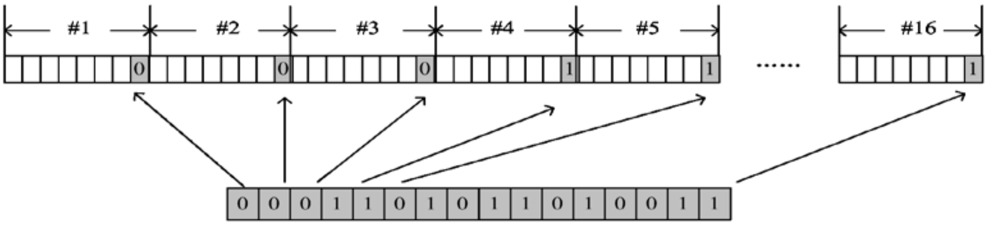
$$m_0 = \left\lceil \frac{LoM}{16} \right\rceil * 16 , \tag{1}$$

$$m_1 = \left\lfloor \frac{P-16}{R*8*16} \right\rfloor * 16 , \tag{2}$$

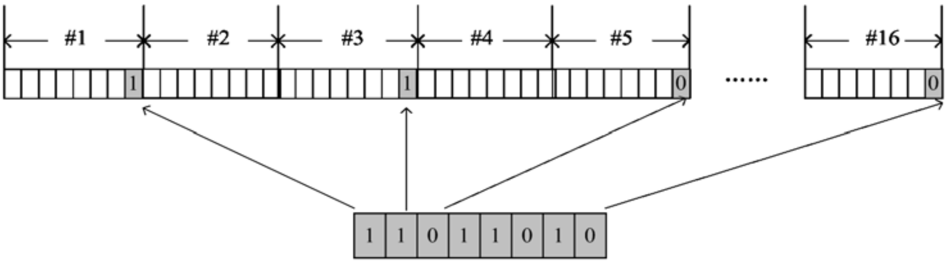
$$m_k = \left\lfloor \frac{P}{R*8*16} \right\rfloor * 16 , \tag{3}$$

$$m_n = \left\lceil \frac{LLoM}{16} \right\rceil * 16 , \tag{4}$$

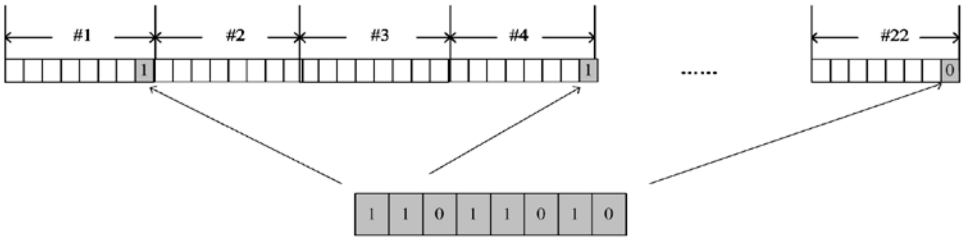
where m_0 is the whole size of the secret message to be embedded, m_1 denotes the length of the secret message embedded into the first packet, m_k represents the length of the secret message embedded into other packets, m_n denotes the length of the secret message in the last packet, *P* stands for the size of audio in each packet, and the value of *P* is 4096 bytes in our covert communication system. The larger the value of *R*, the less the secret message can be embedded into each audio packet. When the secret message is very short, LoM is smaller than m_1.



(a) the traditional LSB method



(b) embedding secret message when R=1



(c) embedding secret message when R=2

Figure 4 Embedding secret message with different values of R.

Figure 4(a) shows the traditional LSB method of embedding the bits of a secret message into the least significant bits of each byte of audio streams. For instance, 16 bytes of audio streams have an embedding capacity of 2 bytes message. Figure 4(b) illustrates the method for embedding a secret message into every 2 bytes of audio streams when the value of R is set to 1. One byte of the secret message is embedded into 16 bytes of audio streams. Figure 4(c) illustrates the method to embed a secret message into every 3 bytes of audio streams when the value of R is 2. So one byte of a secret message needs to be embedded into 22 bytes of audio streams.

The embedding process was designed as follows:

```
Step 1:  embed the length of the secret message expressed as LoM, and let length = LoM.

Step 2:  calculate $m_1$ and $m_k$,
          if LoM<$m_1$,
          encrypt M(0,$m_0$-1) as S(0,$m_0$-1), covert S(0,$m_0$-1) into bit stream B={b(0),b(1),...b(($m_0$-1)*8)},
              if b(i)=0,V(k)=V(k)&0xfe,
              if b(i)=1,V(k)=V(k)|0x01,
              k=k+R,
              length=0,
              end.

          if LoM>$m_1$,
          encrypt M(0,$m_1$-1) as S(0,$m_1$-1),covert S(0,$m_1$-1) into bit stream B={b(0),b(1),...b(($m_1$-1)*8)},
              if b(i)=0,V(k)=V(k)&0xfe,
            if b(i)=1,V(k)=V(k)|0x01,
            k=k+R,
            length=length-$m_1$.

Step 3:  if  length>$m_k$,
         encrypt M($m_1$,$m_1$+$m_k$-1) as S($m_1$,$m_1$+$m_k$-1),
         covert S($m_1$,$m_1$+$m_k$-1) into bit stream B={b(0),b(1),...b(($m_k$-1)*8)},
            if b(i)=0,V(k)=V(k)&0xfe,
            if b(i)=1,V(k)=V(k)|0x01,
            k=k+R,
            length=length-$m_{k,}$
         repeat Step 3 until length <$m_{k.}$

Step 4:  calculate $m_n$,
         encrypt M(LoM-length,LoM-1) as S(LoM-length, LoM-length+$m_n$-1),
         covert S(LoM-length, LoM-length+$m_n$-1) into bit stream B={b(0),b(1),...b(($m_n$-1)*8)},
            if b(i)=0,V(k)=V(k)&0xfe,
            if b(i)=1,V(k)=V(k)|0x01,
            k=k+R,
            length=0,
            end.
```

For instance, when the traditional LSB algorithm was used to embed a secret message, the LoM value was embedded into the first 16 bytes of the first audio packet, and the secret message was embedded into the rest of the first audio packet and the other packets. The rest of the first audio packet was about 4080 bytes, which was capable of embedding 510 bytes of the secret message according to Equation (2). To shorten the encryption time, we embedded 496 bytes of secret message into the first audio packet. The first 496 bytes of the secret message were encrypted with AES, and the resulting ciphertext was then embedded into audio packets using the embedding algorithm above. As for the other audio packets, we embedded 512 bytes of the secret message that were encrypted by AES into each audio packet. The rest of the secret message was embedded into the last audio packet, the size of which is expressed as LLoM. The number of LLoM might not be an integral multiple of 16 bytes, so it was possibly necessary to change it to an integral multiple of 16 bytes. It was larger than LLoM after encryption.

2.5 Extraction

The extraction is a reverse process of the embedding process. The corresponding retrieving algorithm was employed to get the secret message encrypted by AES, and decrypt it with the same SK to obtain the original message in each audio packet.

3 Experimental Setup

3.1 Experiment settings

To evaluate the performance and security of the proposed steganographic algorithm, we employed the VoIP speech samples coded by PCM as the cover-speech. The speech samples employed were classified into two groups, female speech samples and male speech samples. We implemented two sets of tests upon the female speech samples and the male speech samples, respectively.

The Mean Opinion Score (MOS) is the mean of the values on a predefined scale that subjects assign to their opinion of the performance of the telephone transmission system used either for conversation or for listening to spoken material [17]. But the MOS assessment is complex and expensive to perform. To evaluate the impact of steganography upon speech quality, we contrasted Perceptual Evaluation of Speech Quality (PESQ) [18] scores of the cover-speech without steganography and the stego-speech with steganography, respectively. PESQ provides an objective measure that predicts the results of subjective listening tests on telephony systems. The PESQ score is analogous to the subjective Mean Opinion Score measured using panel tests according to ITU-T P.800.

We collected a large number of stego-speech samples at the VoIP communicating receiver for testing. The audio samples were obtained using 2-channel and sampling at 11,025 Hz. There were 17 types of stego-speech files where a message in English was embedded with the different algorithms described in Section 2. The cover-speech without steganography was expressed as “no embedding”. The stego-speech with the English message embedded using the traditional LSB algorithm was expressed as “T_LSB”. The other stego-speech samples using the proposed algorithms with different R values were expressed as “R=value”, and the R value was between 1 and 40.

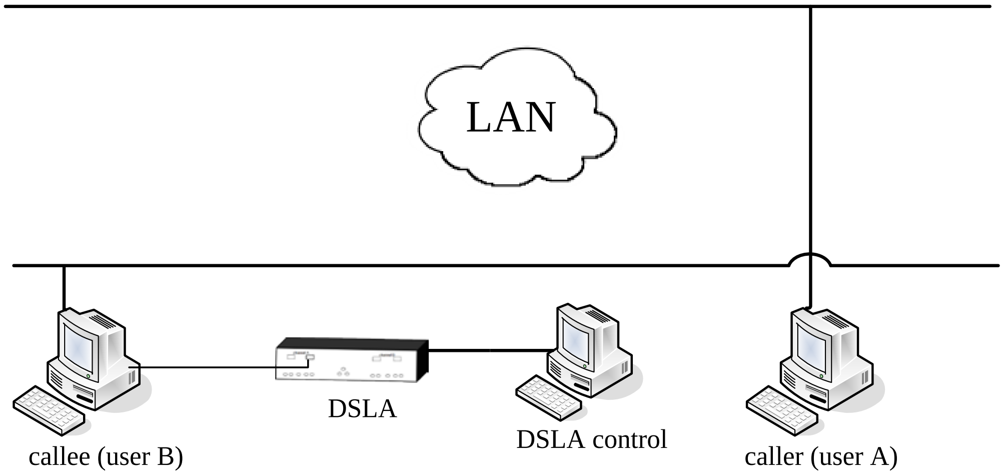


Figure 5 Speech quality testing with DSLA.

Figure 5 shows the experiments of testing the speech quality of the cover-speech and stego-speech streams with DSLA. In the experiments, we used a player to playback records of English audio as the cover-speech to microphone. The audio samples were standard English records from DSLA. Comparisons between cover-speech files and stego-speech files were carried out at the end of the VoIP call. At the receiving end (callee), we measured the PESQ scores of cover-speech samples and stego-speech samples using the DSLA, which is a high-accuracy tool to measure the strength and quality of speech signals.

3.2 DSLA principle

In our experiments, we used DSLA II made by Malden Electronics Ltd in the UK to measure the quality and strength of speech samples. The DSLA and user interface have been designed to provide a simple access to computing the PESQ score. PESQ takes into account the following sources of signal degradation: coding distortions, transmission errors, packet loss, delay and variable delay, and filtering in analogue network components. Thus, PESQ can provide an objective measure of speech quality.

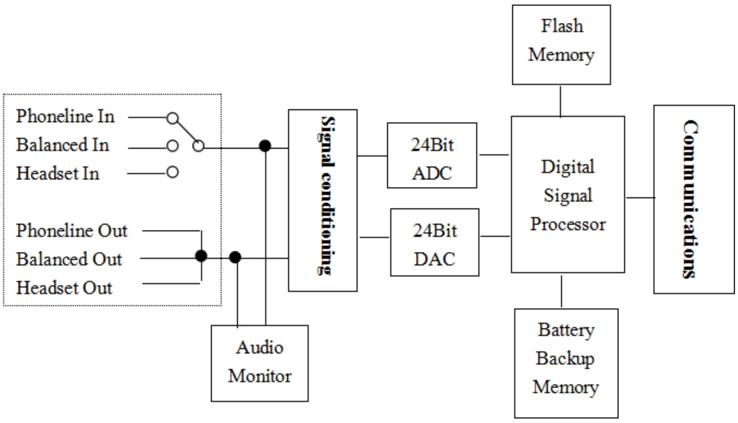


Figure 6 Simplified block diagram of DSLA.

Figure 6 explains the internal components of the DSLA used in our experiments. The DSLA has two channels. From the user interface application, each channel can be switched to use one of three connector types: phone line, balanced, and handset. The DSLA has a 24-bit analogue to digital converters (ADC) and 24-bit digital to analogue converters (DAC). This provides approximately 104 dB of dynamic range. The converters are optimally placed to ensure signals have enough head room and a low enough noise floor. DSLA incorporates bespoke software to provide command and control functionality as well as real-time signal detection and processing [19].

4 Results and Discussion

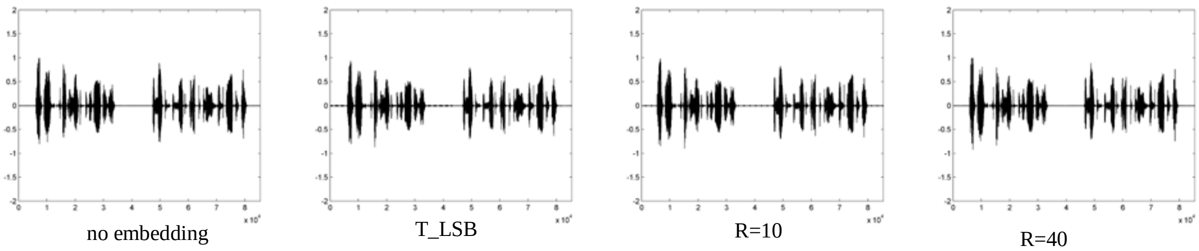
We implemented two sets of tests in the experiments, resulting in two groups of experimental results which can be divided into two categories: Female speech results and Male speech results. From the experimental results, we analyzed the waveform in the time-domain of speech samples, the spectrum energy in the frequency-domain of speech samples, and the signal-to-noise ratio (SNR) of audio signals, and performed statistical analysis of PESQ scores of female speech samples and male speech samples, respectively.

4.1 Female speech results

Figure 7 shows comparisons in the time-domain and frequency-domain of different stego-speech samples containing secret messages as well as the “no embedding” speech sample. As Figure 7(a) shows, some distortions between the waveforms of “T_LSB” and “no embedding” speech samples existed. There were little distortions between the “no embedding” speech and “R=value” stego-speech samples, and some distortions were caused by the different background noises in each call. Listening tests indicated that our human ear could not distinguish the differences between the “no embedding” cover-speech and the stego-speech with the hidden secret message. Figure 7(b) shows the spectrograms of the stego-speech samples received at the VoIP receiver. The spectrogram was the result of calculating the frequency spectrum of sound. As can be seen from these figures, there were slight differences in these spectrograms between the “no embedding” cover-speech and the stego-speech samples. This means that the proposed steganographic algorithm had no or little impact on the time-domain and frequency-domain of the original cover-speech.

We also tested the perceptual evaluation speech quality value to analogize the subjective mean opinion score. Table 2 lists the average PESQ values of female speech samples, the variances between “no embedding” speech and stego-speech samples with the hidden message, and the SNR values of stego-speech samples. Figure 8(a) shows the distributions of PESQ scores of “no embedding” samples and stego-speech samples based on the data in Table 2. The black line is PESQ values for “no embedding” speech samples, and the red curve represents PESQ values for the stego-speech samples containing secret messages. When the values of R increased, the PESQ values of stego-speech samples changed little, and the PESQ values were 3.5 approximately. It is obvious that the proposed encryption and embedding operations had caused little degradation in speech quality, indicative of effective covert VoIP communication.

The SNRs of “no embedding” samples and stego-speech samples were also measured. SNR is a measure of signal strength relative to background noise. The ratio is usually measured in decibels. Figure 8(b) shows comparisons of SNR values between original “no embedding” speech samples and stego-speech samples. The black line is SNR values for “no embedding” speech samples; the red curve represents SNR values for stego-speech samples with the hidden message. With the values of R increasing, the SNR values showed an upward trend until the value of R reached 35 around, indicating that the proposed steganographic algorithm with larger R had no or less impact on the real-time VoIP communication.



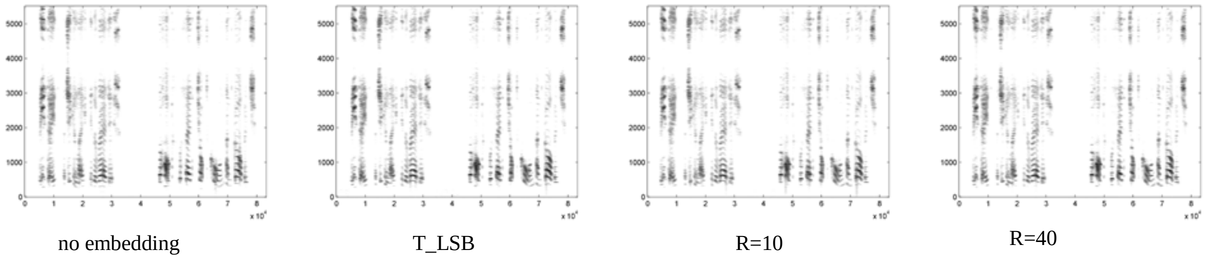
no embedding

T_LSB

R=10

R=40

(a) comparisons in time-domain of female speech-samples



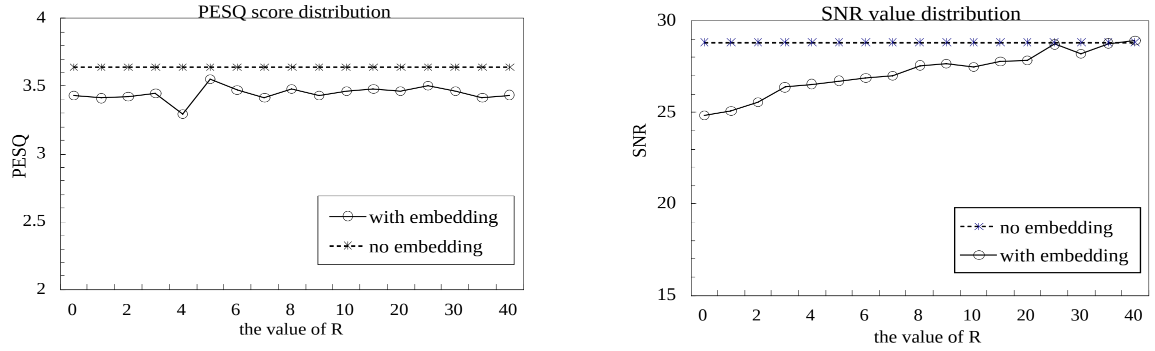
no embedding T_LSB R=10 R=40

(b) comparisons in frequency-domain of female speech-samples

Figure 7 Comparisons in time-domain and frequency-domain of female speech samples.

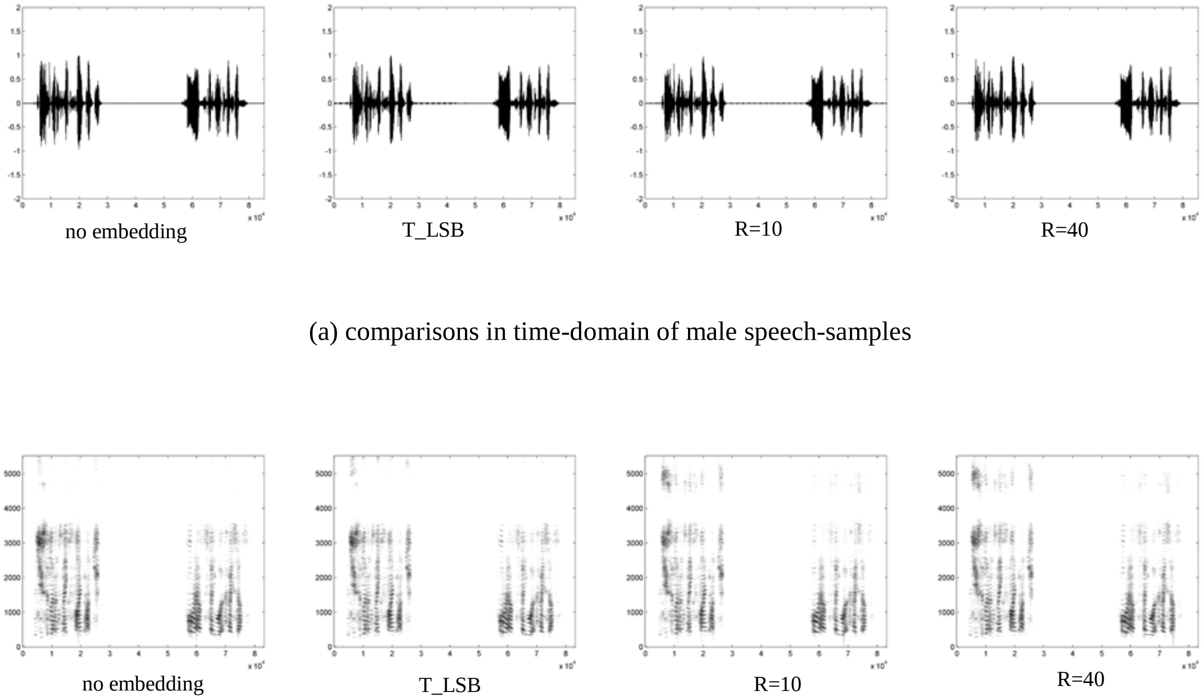
Table 2 Testing results for female speech samples using DSLA II

Female speech samples	PESQ				SNR	
	Max	Min	Mean	Variance	Mean	Variance
no embedding	3.65	3.63	3.6400	---	28.8333	---
T_ LSB	3.47	3.41	3.4330	0.2070	24.8333	4
R=1	3.45	3.35	3.4125	0.2275	25.1000	3.2333
R=2	3.48	3.39	3.4250	0.2150	25.5750	2.7583
R=3	3.54	3.44	3.4500	0.1900	26.3750	1.9583
R=4	3.37	3.21	3.2950	0.3450	26.5500	1.7833
R=5	3.70	3.48	3.5500	0.0900	26.7250	1.6083
R=6	3.48	3.46	3.4700	0.1700	26.9000	1.4333
R=7	3.43	3.39	3.4150	0.2250	27.0250	1.3083
R=8	3.52	3.41	3.4800	0.1600	27.5750	0.7583
R=9	3.48	3.36	3.4325	0.2075	27.6750	0.6583
R=10	3.47	3.46	3.4625	0.1775	27.4750	0.8583
R=15	3.53	3.46	3.4800	0.1600	27.8000	0.5333
R=20	3.54	3.42	3.4625	0.1775	27.8500	0.4833
R=25	3.55	3.45	3.5025	0.1375	28.7250	-0.3917
R=30	3.54	3.33	3.4650	0.1750	28.2000	0.1333
R=35	3.59	3.33	3.4150	0.2250	28.7750	-0.4417
R=40	3.51	3.34	3.4350	0.2050	28.9250	-0.5917



(a) (b)

Figure 8 Comparisons of PESQ and SNR values for female speech samples.



(a) comparisons in time-domain of male speech-samples

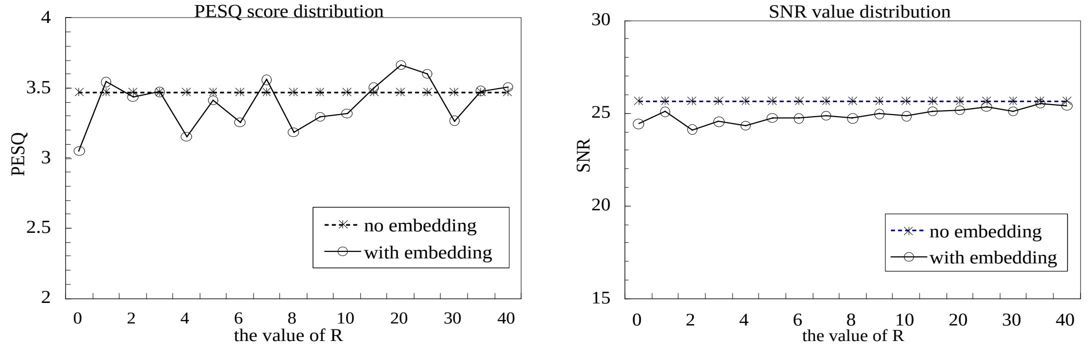
(b) comparisons in frequency-domain of male speech-samples

Figure 9 Comparisons in time-domain and frequency-domain of male speech samples.

Table 3 Testing results for male speech samples using DSLA II

Male speech	PESQ				SNR	
samples	Max	Min	Mean	Variance	Mean	Variance
no embedding	3.61	3.29	3.4700	---	25.66	---
T_ LSB	3.16	2.87	3.0500	0.4200	24.42	1.2400

R=1	3.60	3.48	3.5433	-0.0733	25.08	0.5800
R=2	3.86	3.18	3.4366	0.0334	24.12	1.5400
R=3	3.62	3.25	3.4733	-0.0033	24.54	1.1200
R=4	3.20	3.08	3.1533	0.3167	24.34	1.3200
R=5	3.91	3.07	3.4133	0.0567	24.76	0.9000
R=6	3.28	3.24	3.2567	0.2133	24.72	0.9400
R=7	3.88	3.24	3.5600	-0.0900	24.88	0.7800
R=8	3.25	3.10	3.1867	0.2833	24.72	0.9400
R=9	3.57	2.96	3.2933	0.1767	24.96	0.7000
R=10	3.58	3.15	3.3200	0.1500	24.84	0.8200
R=15	3.71	3.26	3.5033	-0.0333	25.12	0.5400
R=20	3.75	3.53	3.6667	-0.1967	25.16	0.5000
R=25	3.62	3.58	3.6033	-0.1333	25.32	0.3400
R=30	3.65	3.00	3.2667	0.2033	25.10	0.5600
R=35	3.68	3.30	3.4800	-0.0100	25.52	0.1400
R=40	3.58	3.47	3.5100	-0.0400	25.38	0.2800



(a) (b)

Figure 10 Comparisons of PESQ and SNR values for male speech samples.

4.2 Male speech results

Figure 9 shows comparisons in the time-domain and frequency-domain of original male speech (no embedding) samples and stego-speech samples, respectively. Pictures in Figure 9(a) are the waveforms of original male speech (no embedding) samples and stego-speech samples. Pictures in Figure 9(b) are the spectrograms of original male speech (no embedding) samples and stego-speech samples. As these pictures show, there were small changes among them. It is obvious that the proposed steganographic algorithm had little impact on the time-domain and frequency-domain of the original cover-speech.

Table 3 lists changes in PESQ value and SNR value between the “no embedding” speech samples and stego-speech samples with data embedding. Comparatively, the PESQ values of male speech samples were larger than the PESQ values of female speech samples. The maximum values of male stego-speech samples were also larger than those of female stego-speech samples. However, as can be seen from Figure 10(a) that the alterations among the male stego-speech samples were more apparent than the female stego-speech samples.

Figure 10(b) shows the changes in SNR values of "no embedding" speech samples and stego-speech samples with data embedding at different values of R. The SNR values of stego-speech samples increased as the values of R increased. The signal-to-noise values of male stego-speech samples were smaller than those of female stego-speech samples, increased with R increasing, and the changes were more slowly. Overall, these results indicated that the proposed steganographic algorithm had little impact on the real-time VoIP communication.

4.3 Security analysis using Mann-Whitney-Wilcoxon test

The primary goal of steganography is to hide the very fact that a covert communication is taking place alone with an innocuous communication. The security performance of a given steganographic system is normally evaluated based on statistical undetectability, which describes how difficult it is to reliably determine the existence of a hidden message in a cover object. In other words, a secure steganographic system means a statistically undetectable system. Instead of conventional statistical tests, the M-W-W method [20] was employed to conduct security analysis in this study.

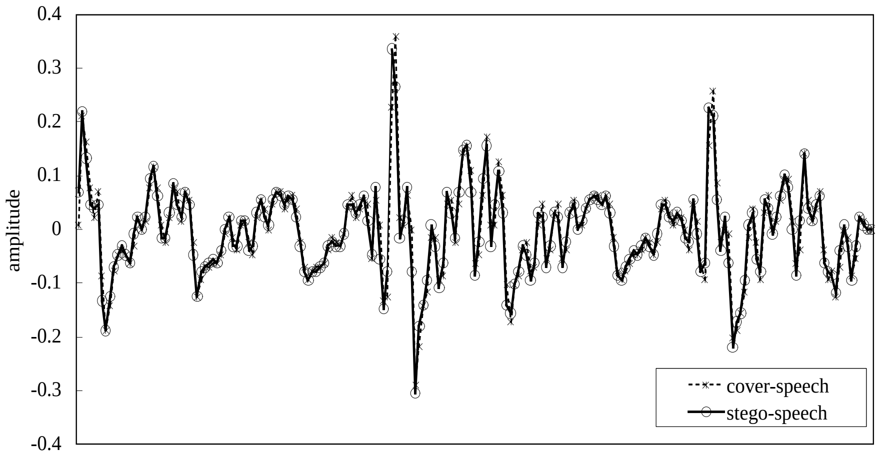


Figure 11 Comparisons of waveforms between male stego-speech samples (when R=10) and cover-speech samples.

This section details how the security of the proposed steganographic algorithm for covert VoIP communications was evaluated using the M-W-W test instead of normal statistical analysis. The M-W-W test is a non-parametric test for assessing whether two independent samples of observations come from the same distribution, and this test is one of the best-known non-parametric significance tests [20]. Comparing the probability distributions between the stego VoIP streams (stego-speech) and the normal VoIP streams (cover-speech) show whether the differences are almost indistinguishable.

When the sample sizes are sufficiently large (at least 12 each), the M-W-W test is based on the standardized test statistic [20]:

$$z^* = \frac{S_2 - E\{S_2\}}{\sigma\{S_2\}} \ , \tag{5}$$

where $E\{S_2\}$ and $\sigma\{S_2\}$ are the mean and square root of variance of the sampling distribution S_2 that is the combination of the two samples of observations to be assessed. To have 95% confidence, i.e., with a confidence coefficient (1-α) of 0.95, where α is called the level of significance, we therefore require $z(1 - \frac{\alpha}{2}) = z(0.975) = 1.960$, where z is the percentile of the standard normal distribution. Hence, the decision rule for the test is as follows:

If $|z^*| \leq 1.960$, conclude H_0 (two distributions do not differ).

If $|z^*| > 1.960$, conclude H_1 (two distributions differ).

To perform the security analysis of the proposed steganographic algorithm, the test statistic for the M-W-W test was calculated as follows:

1. Combined the n_1 sample observations from Population 1 (original VoIP streams) and the n_2 sample observations from Population 2 (stego VoIP streams with embedding), and arrayed the combined data in an ascending order.
2. Assigned ranks to the combined observations (starting with 1 for the smallest observation).
3. Summed the ranks for the n_2 sample observations from Population 2, and denoted this sum by S_2.

Table 4 M-W-W test results and parameters used

Sample size, n_1	201
Sample size, n_2	201
Sum, S_2	40933
Mean of variance, $E\{S_2\}$	40501.50
*S*quare root of variance, $\sigma\{S_2\}$	1356800.25

Test statistic, z^*	0.3704

Figure 11 shows the waveforms of 2 ms of male cover-speech and stego-speech samples when the values of R were set to 10. Table 4 contains the M-W-W test results and the parameters used for comparing the probability distribution drawn from the original VoIP streams (cover-speech) to that drawn from the stego VoIP streams with the hidden messages (stego-speech), as shown in Figure 11. For our proposed steganographic algorithm, the computed test statistic (z^*) was equal to 0.3704, given the sample sizes n_1 and n_2 being 201 and 201, respectively. Since $| z^*| \leq 1.960$, we concluded H_0 – that the probability distributions for the stego VoIP streams and the original VoIP streams without embedding did not differ. This means the proposed steganographic algorithm is undetectable in terms of statistical analysis, which has been used widely to assess the security of steganographic systems.

5 Conclusion

In this study we proposed a VoIP steganographic algorithm with AES and key distribution to achieve a real-time covert VoIP communication. The encryption and embedding processes were almost synchronized. And we also carried out experiments on a large amount of female and male speech VoIP samples. As the experimental results shown, whereas the cover speech could be clearly heard, the covert VoIP communication using the proposed steganographic algorithm was undetectable in terms of statistical analysis. Performance evaluation with state-of-the-art network equipment DSLA and security tests conducted using the M-W-W method indicated that the proposed VoIP steganographic algorithm is secure, effective, and robust.

In future, we should investigate other speech codecs applicable to VoIP, such as iLBC, G.722, G.728, etc., and design new steganographic algorithms which have no or less impact on the quality of speech.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61272469, Grant 61303237, and Grant 61379126, and the Wuhan Scientific Research Program under Grant 2013010501010144.

References

1 C. H. Yang, C. Y. Weng, S. J. Wang, etc. “Adaptive data hiding in edge areas of images with spatial LSB domain systems,” *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*, vol. 3, pp. 488-497, Sep. 2008.

2 Y. K. Lee, L. H. Chen, “High capacity image steganographic model,” *IEE PROCEEDINGS-VISION IMAGE AND SIGNAL PROCESSING*, vol. 147, pp. 288-294, Jun. 2000.

3 L. M. Marvel, C. G. Boncelet, C. T. Retter, “Spread spectrum image steganography,” *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 8, pp. 1075-1083, Aug. 1999.

4 R. Darsana, A. Vijayan. “Audio Steganography Using Modified LSB and PVD,” *TRENDS IN NETWORKS AND COMMUNICATIONS*, vol. 197, pp. 11-20, 2011.

5 N. Cvejic, T. Seppanen, “Increasing the capacity of LSB-based audio steganography,” Proc. 5th IEEE Workshop on Multimedia Signal Processing (MMSP 2002), ST THOMAS, 2002, pp. 336-338.

6 O. Cetin, F. Akar, A. T. Ozcerit, etc. “A blind steganography method based on histograms on video files,” *IMAGING SCIENCE JOURNAL*, vol. 60, pp. 75-82, Apr. 2012.

7 W. Mazurczyk, K. Szczypiorski, “Steganography of VoIP streams,” *Lecture Notes in Computer Science*, vol. 5332, pp. 1001-1018, 2008.

8 Y. F. Huang, S. Tang, C. Bao, Y. J. Yip. “Steganalysis of compressed speech to detect covert voice over Internet protocol channels,” *IET INFORMATION SECURITY*, vol. 5, iss.1, pp. 26-32, 2011.

9 R. Miao, Y. F. Huang, “An Approach of Covert Communication Based on the Adaptive Steganography Scheme on Voice over IP,” Proc. IEEE Int. Conf. on Communications, Kyoto, Japan, JUN 05-09, 2011.

10 C. Kratzer, J. Dittmann, T. Vogel, R. Hillert, “Design and evaluation of steganography for Voice-over-IP,” Proc. IEEE Int. Symp. on Circuits and Systems, Kos, GREECE, 21-24 May 2006, pp. 2397-2340.

11 H. Tian, K. Zhou, Y. F. Huang, etc. “A Covert Communication Model Based on Least Significant Bits Steganography in Voice over IP,” Proc. 9th Int. Conf. for Young Computer Scientists, Zhangjiajie, P. R. China, Nov 18-21, 2008, pp. 647-652.

12 C. Y. Wang, Q. Wu, “Information Hiding in Real-time VoIP Streams,” Proc. 9th IEEE Int. Symp. on Multimedia, Taichung, Taiwan, Dec 10-12 2007, pp. 255-262.

13 C. H. Wang, Y. S. Liu. “A dependable privacy protection for end-to-end VoIP via Elliptic-Curve Diffie-Hellman and dynamic key changes,” *JOURNAL OF NETWORK AND COMPUTER APPLICATIONS*, vol. 34, pp. 1545-1556, 2011.

14 R. W. Yu, J. Yuan, G. Du, P. Li, “An identity-based mechanism for enhancing SIP security,” Proc. 2012 IEEE 3rd International Conference on Software Engineering and Service Science (ICSESS), Beijing, 22-24 June 2012, pp. 447-451.

15 Daemen, Joan; Rijmen, Vincent (9/04/2003). “AES Proposal: Rijndael”. National Institute of Standards and Technology. p. 1. Retrieved 21 February 2013.

16 FIPS 197, Announcing the advanced encryption standard (AES), 2001.

17 ITU-T Rec. P.10/g.100, Vocabulary for performance and quality of service, 2006.

18 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.

19 DSLA II Getting Started Guide, revision 3.0, Malden Electronics Ltd, United Kingdom, 2012.

20 J. Neter, W. Wasserman, G. A. Whitmore, Applied Statistics, 4th edition, Simon & Schuster, Inc., 1993, pp. 435-450.