Real-Time Scheme for Covert Communication Based VoIP

Haider Ismael Shahadi, Muayad S. Kod, Baneen Qasem and Hameed R. Farhan

1Department of Electrical and Electronic Engineering, University of Kerbala, 56001, Karbala, Iraq.

Abstract. Covert communication is an important necessity for governments and companies due to the ease of eavesdropping and access to information. A private channel with robust encryption can provide the required protection for the communications. However, a public channel is cheaper and less restrictive (from any place and any communication devices) than private channels. This paper proposes a real-time covert communication scheme by creating a secret speech channel in a public speech channel based on audio steganography. The scheme compresses the secret speech using a very fast and simple process based on lifting wavelet transform. Also, the cover speech is compressed by G.711 encoder at synchronous time periods with respect to the secret speech to meet the networks of voice over internet protocol (VoIP). Subsequently, the secret data hides in a compressed cover speech data in real-time. The embedding procedure employs the only high strength data for embedding to maintain the quality of the cover data after embedding at greater than 40 dB in terms of signal to noise ratio (SNR). The secret data is retrieved 100% without error in case of a lossless channel.

1. Introduction

With the rapid development of the internet, steganography becomes widely used in various fields and applications to provide higher security levels by hiding secret data in digital media such as images, text, audio, or video files as covers. Audio steganography is a challenger to other media due to the higher sensitivity of the human auditory system (HAS) compared to the human visual system (HVS) [1][2]. If the secret data is also audio, the challenge is increased due to the large amount of data that pose some difficulties in carrying it on a cover. The challenge will be in peak if audio steganography requires to process in real-time communication. This is because of the contradictory requirements of the efficient audio steganographic system, including high embedding capacity, imperceptibility, robust scheme, high level of security, low complexity, and suitable latency proper for real-time communication.

The developed version of audio steganography uses Voice over Internet Protocol (VoIP) steganography calls as cover audio to embed secret data. The reason for using VoIP is the providing of global services with a low cost or even free services and high reliability. Also, it gives the security for the confident data, where it is uneasy for attackers to reveal that there are secret embedded data in VoIP streams or not [3][4]. However, to meet real-time communication requirements, there is difficulty in introducing the embedding algorithm with more secure operations to hide and protect the secret information. Therefore, this paper attempts to provide a scheme that meets real-time voice
communications by creating an embedded channel for voice communication in another public channel as a carrier or cover voice.

Recently, several steganography techniques are proposed based on real-time VoIP. However, the embedded data rates for most of these techniques are unsuitable for hiding speech into speech in real-time communications. Most techniques employ the least significant bit (LSB) to obtain low complexity and high embedding rate. In [5], a covert communication model had been proposed based on the LSB method in VoIP, which embedded encrypted secret data into compressed cover audio (compressed by G.729a codec). The model can provide good transparency, maintain security, and produce short-latency, but the LSB algorithm used has no robustness to additive noise, and the nonstandard encryption algorithm was used for encryption. The researchers in [6] proposed a real-time steganographic technique to hide secret message encrypted with Advanced Encryption Standard (AES) algorithm and symmetric key into cover audio encoded by PCM Codec. The model has variable embedding capacities due to using different embedding location intervals for LSB replacements. However, the capacity is suitable for hiding simple secret data, where only the cover in real-time (or online), while the secret message should entirely exist (offline).

In [7], the authors proposed a lossless audio steganography model based on integer lifting wavelet transform and the LSB method. It used integer values to eliminate any error that may occur in the recovered data in the rounding process using non-integer values. The proposed algorithm is fast, simple, and has adequate security. However, it is not suitable for VoIP because the embedding before compress speech, so that the hidden speech may lose during the compression process in VoIP. The confident data is embedded by adjusting the length of silence intervals of cover audio in [8]. The introduced algorithm increased robustness to MP3 compression and noise, but the embedding capacity is much reduced.

Another technique has been introduced in [9], which uses the inactive frames of low bit rate audio streams compressed with G.723.1 source Codec to embed the secret data with high embedding capacity. However, the bit rate is unsuitable for speech hiding as a message signal.

In this study, a lossless scheme for real-time covert VoIP steganography is proposed. The proposed scheme has a very high embedding rate (25% from the cover speech size) with imperceptibility above 40 dB in terms of signal-to-noise ratio (SNR). Moreover, it is robust against additive noise. The scheme is designed in order to employ VoIP public channel as a carrier for secret speech data. As it is known, every channel, whether cable or audio has limited capacity, so the great challenge within our work is the capacity of the public channel. Moreover, the processing time of the embedding algorithm should be proportional to the real-time communication requirements.

2. The Proposed Real-time Audio Steganography Based VoIP

The proposed real-time VoIP communication approach is implemented with a secure channel that is created into a public channel. As shown in figure 1, the call between two important persons can be hidden into another call between two ordinary persons who communicate in a VoIP network. The speech of the first important person is considered as a secret speech and embedded in the cover speech (the first normal person speech) using the proposed embedding algorithm. The embedding algorithm is based on module operator and high energy cover samples to provide high imperceptibility to the resulted signal (stego-speech) by allowing minimum distortion. All eavesdropping people or even the second ordinary person at the receiver end do not have any authentication to access the secret speech at the VoIP network. The second important person is the only one that can recover the secret speech after applying the proposed recovery algorithm. The steps of the proposed embedding and recovery algorithms are demonstrated in the following sub-sections.

2.1. Embedding Algorithm Steps

The main steps of the embedding algorithm are shown in figure 2 (a), and illustrated as follows:

1. An input cover speech (C) of 32 kHz sampling rate is framed without overlap into N frames of 512 samples/frame to realize synchronization between the cover and secret speech signals. Each
sample is implemented with 16 bits of resolution. Also, an input secret speech (S) of 8 kHz is framed into N frames without overlap, where each frame has 128 samples/frame with a resolution of 8 bits/sample. The outputs of this step are two frames (C_i and S_i, i= 1,2, ..., N) in a period of 16 mile-second.

2. Each cover frame is compressed using G.711 Codec to get the frame (C_i,comp) with 512 samples, where each sample comprises 8 bits (b_7 b_6 b_5 b_4 b_3 b_2 b_1 b_0). Also, each secret frame is decomposed using one level of Haar integer to integer lifting wavelet transform (Haar-int2int-LWT) to obtain an approximation sub-band and a detail sub-band, where each of which has 64 samples. After that, the average value for the detail samples is calculated and inserted in the approximation sub-band as one sample. As a result, the compressed secret frame (S_i,comp) will have 65 samples/frame, and each sample has 8 bits of resolution. The resulted compressed secret frame (S_i,comp) is permuted randomly according to the Pseudo-Random Generator (PRN). The output of PRN is dependent on a certain secret key (K) that is entered by the user.

3. After the permutation stage, each secret sample (S_{ij,comp}) is split into four digits of two bits to obtain (S_{ir,comp,P}), where r = 1, 2, 3, 4.

4. In the embedding stage, selecting high energy cover samples from C_i,comp according to a specific input threshold (Thr). So, only the candidate cover samples will be used in the embedding process, as illustrated in (1).

\[ C_{ik,Candidate} = C_{ij} \text{ if } C_{ij} > Thr, \text{ where } k = 1, 2, \ldots, M, M < 512 \]  \hspace{1cm} (1)

Hold bit value (HB) is chosen between 0 and 3 to specify the first position of embedding after LSB in \( C_{ik,Candidate} \), in case of \( C_{ik,Candidate} \) is chosen as a suitable sample for embedding. This means that the first bit of embedding will be increased with increasing HB to increase the hidden data robustness against adding noise. For example, if HB is 3, then the first embedding bit will be the fourth one from candidate \( C_{ik,Candidate} \) and the second bit of embedding will be the fifth one.

5. Subsequently, extracting the bits after the final hold bit from the candidate cover samples to get \( C_{ik,Candidate,HB} \) to be used for embedding the secret digits of \( S_{ir,comp,P} \). Equations (2) and (3) are used to perform the embedding process.

\[ R = C_{ik,Candidate,HB} \mod(4) \]  \hspace{1cm} (2)

\[ \hat{C}_{ik,HB} = C_{ik,Candidate,HB} - (R - S_{ir,comp,P}) \]  \hspace{1cm} (3)

6. The hold bits of stego-sample are combined with it (in case of HB>0) to obtain \( \hat{C}_{ik} \). Also, rebuilt the compressed cover frame after the embedding process by replacing each candidate sample used in embedding with its corresponding stego-sample (\( \hat{C}_{ik} \)). The resulted stego-cover frame (\( \hat{C}_i \) with a length of 512 samples) carries the secret frame (S_i with a length of 65 samples).
Finally, the N stego-frames are combined into stego-data ($\hat{C}$). Now, the stego-data represents compressed data of a speech signal. Any person has access to this data; he/she can apply G.711 decoder to hear a speech signal similar to the cover without perceptible difference compared to the original one.

Figure 1. The general structure of the proposed real-time covert VoIP communication.

2.2. Recovery Algorithm Steps
The main steps of the message recovery algorithm are shown in figure 2 (b), where they are explained as follows:

1. An input stego data ($\hat{C}$) is framed into N frames of 512 samples/frame without overlapping.
2. The samples that have a strength greater than (Thr) are selected to extract only the samples that carry secret information. Then, removing the HB from each selected sample (in case of HB>0). The resulted samples are $\hat{C}_{i,HB}$.
3. Retrieve the secret information starting from the first sample from $\hat{C}_{i,HB}$ until completing $65^*4$ digits using the following formula:

   $$S_{IP,comp,P} = \hat{C}_{i,HB} \mod (4)$$  \hfill (4)

4. Sequentially, every four digits are combined into one secret sample, then reconstructing the frame i from the secret compressed data that have 65 samples ($S_{i,comp,P}$).
5. Decrypt the resulted frame by using the inverse of the permutation used in the embedding process at the transmitter side. The resulted frame is $S_{i,comp}$.
6. Subsequently, $S_{i,comp}$ is decompressed using the inverse of one level-Haar-int2int –LWT. This is achieved by creating 64 samples of the detail coefficients with equal values (all the coefficients have the same retrieved average value). The other approximation coefficients represent the reminder 64 samples from $S_{i,comp}$. The resulted frame is $S_i$, which represents a retrieved secret speech frame.
7. Finally, the frames of the retrieved secret speech are combined according to the synchronous time to obtain the final secret message.

The retrieved compressed secret message data is error-free in case of a lossless channel. This is because the robust method used in terms of its operations is invertible without losing any information from the embedded data.

3. Results and discussion
In this section, some of the experimental results for the proposed scheme as well as comparison with related work are listed. The obtained results are based on embedding 64000 bps secret speech into a compressed cover speech of 256000 bps (512000 bps before the compression). Several real-time input speech signals as covers and real-time input speech as secret signals have been tested. MATLAB (2017a) programming is employed to implement and evaluate the performance of this work.
Figure 2. The proposed Scheme steps: (a) Embedding steps, (b) Secret message recovery steps.

3.1. Perceptual quality and embedding capacity tests

There is a trade-off between the quality of stego-Speech and embedding capacity. This means that increasing an embedding capacity leads to a decrease in perceptual quality. Typically, a suitable algorithm can maintain quality while increasing capacity. The quality test of stego-speech is calculated using Signal to Noise Ratio (SNR) according to (5) [10] and Perceptual Evaluation of Speech Quality (PESQ).

\[
SNR = 10 \times \log_{10} \frac{\sum_{i=1}^{n} C_i^2}{\sum_{i=1}^{n} |\hat{C}_i - C_i|^2}
\]  

(5)

Where \( n \) is the number of samples in cover (\( C \)) or stego (\( \hat{C} \)) speech.

Table 1 shows several tests for perceptual quality in terms of SNR and PESQ. Three cover and secret speeches have been tested with threshold values, 64 and 128. The tests show that the proposed scheme has excellent imperceptibility for different input cover and secret speeches. This is because the secret data is embedded only in the high energy cover samples. Therefore, the difference between the original and modified samples is reduced to minimum values. As a result, maintaining the cover speech quality. The given quality is achieved with a very high embedding capacity that is equal to 12.5\% in the case of...
measuring the exact data hidden after compressing the secret speech. Alternatively, the embedding rate is 25% of the cover speech size if we consider the input secret speech size with respect to the compressed cover speech size. All the results in table 1 have been achieved for HB = 0. However, in the case of a noisy channel, increasing the robustness is required to avoid any loss in hidden data, which is achieved by increasing the value of HB factor, as illustrated in the next subsection.

3.2. Robustness tests against additive noise
The similarity between the original and the retrieved secret speech is measured using NC and BER, according to (6) [10] and (7) [11], respectively.

\[
NC = \frac{\sum_{i=1}^{m} S_i \times \hat{S}_i}{\sqrt{\sum_{i=1}^{m} S_i^2} \times \sqrt{\sum_{i=1}^{m} \hat{S}_i^2}}
\]  

(6)

\[
BER = \frac{1}{k} \times \sum_{i=1}^{k} \left\{ 0, \hat{S}_i = S_i \right\} \times 100\%
\]  

(7)

Where: m is the total number of samples in the secret speech and k is the total number of bits in the secret speech.

The robustness of the proposed approach increases with embedding insertion depth that is controlled by the HB value. However, table 2 shows the system robustness tests against adaptive Wight Gaussians noise (AWGN) with versus HB value, where the increase of HB value leads to an increase in the secret speech immunity. Although, the stego-speech quality is degraded, but it is still within a good or acceptable range. The robustness results of both the embedding stage and the overall system (embedding and compression stages) in terms of BER and NC are shown in table 3 where the proposed system has good robustness with an error-free retrieved secret speech in the case without adding noise.

| Table 1. Tests of quality performance for the Stego-key. |
|---------------------------------|
| Secret signals | Cover signals | Threshold | SNR      | PESQ     |
|----------------|---------------|-----------|----------|----------|
| S1             | C1            | 64        | 41.4580  | 4.42     |
|                |               | 128       | 41.5586  | 4.46     |
|                | C2            | 64        | 41.3846  | 4.39     |
|                |               | 128       | 41.4277  | 4.41     |
|                | C3            | 64        | 41.3681  | 4.37     |
|                |               | 128       | 41.4194  | 4.40     |
| S2             | C1            | 64        | 41.7019  | 4.56     |
|                |               | 128       | 41.8358  | 4.58     |
|                | C2            | 64        | 41.6385  | 4.51     |
|                |               | 128       | 41.6534  | 4.55     |
|                | C3            | 64        | 41.6024  | 4.50     |
|                |               | 128       | 41.6529  | 4.53     |
| S3             | C1            | 64        | 41.5320  | 4.45     |
|                |               | 128       | 41.5959  | 4.48     |
|                | C2            | 64        | 41.4879  | 4.43     |
|                |               | 128       | 41.5145  | 4.45     |
|                | C3            | 64        | 41.4150  | 4.40     |
|                |               | 128       | 41.4913  | 4.44     |
3.3. Comparison with related work

In this section, the proposed approach is compared to some related work. There is an important issue here in the comparison, where the most existing research in the area of VoIP steganography is dedicated to hiding offline secret data (text, image, or speech) within real-time cover speech [6][12][13][14]. Thus the hiding rates that have been used are unsuitable for embedding real-time secret speech. In our approach, a real-time secret speech is hidden within a real-time cover speech and fixed high hiding rate of 64 kbps, which is more beneficial than most other work. Furthermore, the suggested approach is superior in terms of quality performance, as indicated in table 4, where the average values for SNR and PESQ are not less than 41dB and 4.5 respectively in the case of HB=0.

### Table 2. Some tests of secret speech immunity against AWGN for the proposed approach.

| HB | SNR for stego-speech  | NC for different AWGN in dB |
|----|------------------------|----------------------------|
| 0  | 41.5328                | 50 | 0.974 | 0.931 | 0.887 | 0.826 | 0.764 |
| 1  | 34.8962                | 40 | 0.983 | 0.955 | 0.917 | 0.865 | 0.779 |
| 2  | 30.0853                | 30 | 0.985 | 0.974 | 0.957 | 0.919 | 0.856 |
| 3  | 25.3505                | 30 | 1     | 0.991 | 0.984 | 0.938 | 0.924 |

### Table 3. Test results for quality performance of the retrieved secret speech.

| Cover signals | Threshold | Embedding Stage | Overall system (compression and embedding) |
|---------------|-----------|-----------------|------------------------------------------|
|               |           | NC  | BER % | NC  | BER % |
| C1            | 64        | 1   | 0     | 0.9958 | 0.3775 |
|               | 128       | 1   | 0     | 0.9100 | 0.5090 |
| S1            | C2        | 64  | 1    | 0     | 0.9958 | 0.3775 |
|               |           | 128 | 1    | 0     | 0.8764 | 0.5096 |
| C3            | 64        | 1   | 0     | 0.9958 | 0.3775 |
|               | 128       | 1   | 0     | 0.9958 | 0.3775 |
| S2            | C2        | 64  | 1    | 0     | 0.9974 | 0.3604 |
|               |           | 128 | 1    | 0     | 0.9005 | 0.4938 |
| C3            | 64        | 1   | 0     | 0.9974 | 0.3604 |
|               | 128       | 1   | 0     | 0.8930 | 0.4795 |
| S3            | C2        | 64  | 1    | 0     | 0.9989 | 0.3233 |
|               |           | 128 | 1    | 0     | 0.8987 | 0.4605 |
| C3            | 64        | 1   | 0     | 0.9989 | 0.3233 |
|               | 128       | 1   | 0     | 0.9102 | 0.5058 |

3.3. Comparison with related work

In this section, the proposed approach is compared to some related work. There is an important issue here in the comparison, where the most existing research in the area of VoIP steganography is dedicated to hiding offline secret data (text, image, or speech) within real-time cover speech [6][12][13][14]. Thus the hiding rates that have been used are unsuitable for embedding real-time secret speech. In our approach, a real-time secret speech is hidden within a real-time cover speech and fixed high hiding rate of 64 kbps, which is more beneficial than most other work. Furthermore, the suggested approach is superior in terms of quality performance, as indicated in table 4, where the average values for SNR and PESQ are not less than 41dB and 4.5 respectively in the case of HB=0.
References

Table 4. Comparative results of the perceptual quality and hiding rate with some related approaches.

|                | Zhijun [12] | Jiang [13] | Peng [14] | Tang [6] | The proposed |
|----------------|-------------|------------|-----------|----------|--------------|
| SNR            | Not reported| 38.7       | Not reported| 27.5     | 41.5         |
| PESQ           | 3.11        | 4.04       | 4.4       | 3.5      | 4.5          |
| Hiding rate (bps) | 2400       | 8000       | 800       | 3968     | 64000        |

In addition to the superiority of the quality performance, the proposed approach has a full recovery of secret speech in a lossless channel. Also, it has good robustness against AWGN due to embedding the secret speech data in higher positions (3rd, 4th, and 5th levels), whereas [6], [5], and most of the existing methods used only the LSB algorithm, which is very sensitively to additive noise.

4. Conclusions

In this paper, real-time communication of covert speech system based VoIP has been proposed. The system hides a real-time secret speech (conducted in a secure channel) into a real-time cover speech (conducted in a public channel). In order to increase the perceptual quality of stego speech, firstly, the secret speech is compressed using Haar int2int LWT to reduce the required size of embedding. Secondly, the high energy cover samples are selected to embed the secret data. Also, random permutation based PNR and stego-key is achieved on secret data before embedding to increase the security level. Moreover, HB factor is employed to increase the immunity of the hidden data against the AWGN by increasing the depth of embedding. The experimental results show that the secret speech is fully retrieved in case of a lossless channel and with good performance quality in case of adding noise. Also, the embedding algorithm requires low computational complexity with suitable processing time for real-time VoIP requirements. Furthermore, a high data embedding rate up to 64 kbps with better imperceptibility have been achieved. The proposed scheme can be used in several applications that need a high level of security based on public and cheap channel (VoIP), such as military and intelligence communications.

References

[1] El-Khamy SE Korany NO and El-Sherif MH 2016 A security enhanced robust audio steganography algorithm for image hiding using sample comparison in discrete wavelet transform domain and RSA encryption Multimed. Tools Appl. 76 22 p 24091–24106.
[2] Tan D Lu Y Yan X and Wang X 2019 A simple review of audio steganography Proc. 2019 IEEE 3rd Inf. Technol. Networking. Electron. Autom. Control Conf. ITNEC 2019 p 1409–1413.
[3] Deepikaa S and Saravanan R 2019 VoIP steganography methods, a survey Cybern. Inf. Technol. 19 1 p 73–87.
[4] Jiang Y Zhang L Tang S and Zhou Z 2013 Real-time covert VoIP communications over smart grids by using AES-based audio steganography Proc. 2013 IEEE Int. Conf. Green Comput. Commun. IEEE Internet Things IEEE Cyber, Phys. Soc. Comput. p 2102–2107.
[5] Tian H Zhou K Huang Y Feng D and Liu J 2008 A covert communication model based on least significant bits steganography in voice over IP Proc. 9th Int. Conf. Young Comput. Sci. ICYCS 2008 p 647–652.
[6] Tang SY Jiang YJ Zhang LP and Zhou ZB 2014 Audio steganography with AES for real-time covert voice over internet protocol communications Sci.China Inf.Sci. 57 3 p 1–14.
[7] Shahadi H I Jidin R and Way W H 2014 Lossless audio steganography based on lifting wavelet transform and dynamic stego key Indian J. Sci. Technol. 7 3 p 322–334.
[8] Shirali-Shahreza M H and Shirali-Shahreza S 2010 Real-time and MPEG-1 layer III compression resistant steganography in speech IET Inf. Secur. 4 1 p 1–7.
[9] Huang YF Tang S and Yuan J 2011 Steganography in inactive frames of VoIP streams encoded by source codec IEEE Trans. Inf. Forensec Secur. 6 2 p 296–306.
[10] Shahad HI Jidin R and Way WH 2015 Concurrent hardware architecture for dual-mode audio steganography processor-based FPGA Comput. Electr. Eng. 49 p 95–116.
[11] Anjana KA Satheesh CC Kamal S and Supriya M H 2017 Spread spectrum based encrypted audio steganographic system with improved security Proceeding Second Int. Conf. Circuits, Control. Commun. IEEE p 109–114.

[12] Zhijun W Haijuan C and Douzhe L 2015 An approach of steganography in G.729 bitstream based on matrix coding and interleaving Chinese J. Electron. 24 1 p 157–165.

[13] Jiang Y and Tang S 2018 An efficient and secure VoIP communication system with chaotic mapping and message digest Multimed. Syst. Springer. 24 3 p 355–363.

[14] Peng J and Tang S 2020 Covert Communication over VoIP Streaming Media with Dynamic Key Distribution and Authentication IEEE Trans. Ind. Electron. 68 p 3619–3628.