PRIVACY PRESERVATION IN THE INTERNET TELEPHONY CALLS

M.Karthick[1] | Vaijayanthi Murugan[2]

[1](AP/CSE, SNS College of Technology, Coimbatore, India, vaijayanthimurugan25@gmail.com)
[2](AP/CSE, SNS College of Technology, Coimbatore, India, magukarthik@gmail.com)

Abstract— *Steganography is the hiding of a secret message within an ordinary message and the extraction of secret message at its destination. In digital Steganography, electronic communications may include steganographic coding inside of a transport layer, such as a document file, image file, program or protocol. This paper describes how to transfer the secret message from the source to destination in the audio that are streaming in the Voice over Internet Protocol (VoIP). VoIP is an IP telephony term for a set of facilities used to manage the delivery of voice information over the Internet. The secret data are embedded by a Steganography algorithm in the inactive frames of low bit rate audio streams that are encoded by a source codec. The inactive frame in the audio stream is recognized by Voice activity detection (VAD), it is a technique used in speech processing in which the presence or absence of human speech is detected. The lost or dropped packets may lead to loss of secret data. When a data unit arrives, a check is made on the integrity of the data. If the check fails, this will automatically retransmit the lost packet. To maintain the integrity of the secret data HARQ is used. Experimental results show that the proposed system will not affect the perceptual speech quality between the original speech and the stego speech.*
Index Terms—Audio streams; inactive frames; Steganography; Voice over Internet Protocol (VoIP).

Keywords—Audio streams; inactive frames; Steganography; Voice over Internet Protocol (VoIP).

1. INTRODUCTION

VoIP is one of the most popular services in IP networks and it stormed into the telecom market and changed it entirely. As it is used worldwide more and more willingly, the traffic volume that it generates is still increasing. That is why VoIP is suitable to enable hidden communication throughout IP networks. Applications of the VoIP covert channels differ as they can pose a threat to the network communication or may be used to improve the functioning of VoIP. Digital Steganography in low bit rate audio streams is commonly regarded as a challenging topic in the field of data hiding.

There have been several Steganography methods of embedding data in audio streams.G.711-based adaptive speech information hiding approach [5], a technique of lossless Steganography in G.711 encoded speeches [6], a Steganography method of embedding data in G.721 encoded speeches [7]. All these methods adopt high bit rate audio streams encoded by the waveform codec as cover objects, in which plenty of least significant bits exist. VoIP are usually transmitted over low bit rate audio streams encoded by the source codec like ITU G.723.1 codec to save on network bandwidth. Low bit rate audio streams are less likely to be used as cover objects for Steganography since they have fewer least significant bits than high bit rate audio streams. Little effort has been made to develop algorithms for embedding data in low bit rate audio streams.

2. RELATED WORK

A. G.711-BASED ADAPTIVE SPEECH INFORMATION HIDING APPROACH

These systems presents an adaptive LSB (Least Significant Bit) algorithm to embed dynamic secret speech information data bits into public speech of G.711-PCM (Pulse Code Modulation) [5] for the purpose of secure communication according to energy distribution, with high efficiency in Steganography and good quality in output speech. It is superior to available classical algorithms, LSB. Experiments show that this approach is capable of embedding up to 20 Kbps information data of secret speech into G.711 speech at an average embedded error rate of $10^{-5.}$ It meets the requirements of information hiding, and satisfies the secure communication speech quality constraints with an excellent speech quality and complicating speaker's recognition.

B. Voice Activity Detection based on PWPT and TEO

The voice activity detection is used to distinguish speech from noise and is required in a variety of speech processing systems. In the GSM-based communication system, VAD can save battery power by discontinuing transmission when no voice activity is detected.VAD can be used in a variable bit rate speech coding system in order to control the average bit rate and the over all coding quality of speech. The conventional VAD algorithms are accomplished by applying several parameters extracted from the input speech signal to compare with predetermined thresholds. If the measured parameters exceed the thresholds, then a voice active decision is made. Here it will make use of the [3] perceptual wavelet packets transform (PWPT) and the Teager energy operator (TEO).

C. VAD for Speech Enhancement Applications

An important problem in speech processing applications is the determination of active speech periods within a given audio signal. Speech can be characterized as a discontinuous signal, since information is carried only when someone is speaking. The regions where voice information exists are referred to as 'voice active' segments, and the pauses between talking are called 'voice-inactive' or 'silence' segments. The decision on the class to which an audio segment belongs is based on an observation vector. This is commonly referred to as a ‘feature’ vector [4]. One or many different features may serve as the input to a decision rule that assigns the audio segment to one of these two classes. An algorithm employed to detect the presence or absence of speech is referred to as a voice activity detector (VAD). Generating an accurate indication of the presence or absence of speech is generally difficult, especially when the speech signal is corrupted by background noise or by unwanted impulse noise. Voice activity detection algorithm performance trade-offs are made by maximizing the detection rate of active speech while minimizing the false detection rate of inactive segments.

D. *G.711 Pulse Code Modulation (PCM) of Voice Frequencies*

G.711 uses a sampling rate of 8,000 samples per second. The tolerance on that rate is ± 50 parts per million (ppm). Eight binary digits per sample are used. Two encoding laws are used and these are commonly referred to as the A-law and the mu-law [10]. When using the mu-law in networks where suppression of the all 0 character signal is required, the character signal corresponding to negative input values between decision values numbers 127 and 128 should be 00000010 and the value at the decoder output is -7519. The corresponding decoder output value number is 125. Packet Loss Concealment (PLC) algorithms, also known as frame erasure concealment algorithms, hide transmission losses in an audio system where the input signal is encoded and packetized at a transmitter, sent over a network, and received at a receiver that decodes the packet and plays out the output. Many of the standard CELP-based speech coders, such as Recommendations G.723.1, G.728 and G.729, have PLC algorithms built into their standards.

3. METHODS OF AUDIO STEGANOGRAPHY

A. *LSB coding*

Least significant bit (LSB) coding is the simplest way to embed information in a digital audio file. By substituting the least significant bit of each sampling point with a binary message, LSB coding allows for a large amount of data to be encoded. In LSB coding, the ideal data transmission rate is 1 kbps per 1 kHz. In some implementations of LSB coding, however, the two least significant bits of a sample are replaced with two message bits. This increases the amount of data that can be encoded but also increases the amount of resulting noise in the audio file as well.

B. *Parity Coding*

Instead of breaking a signal down into individual samples, the parity coding method breaks a signal down into separate regions of samples and encodes each bit from the secret message in a sample region's parity bit. If the parity bit of a selected region does not match the secret bit to be encoded, the process flips the LSB of one of the samples in the region. Thus, the sender has more of a choice in encoding the secret bit, and the signal can be changed in a more unobtrusive fashion.

C. *Phase coding*

In phase coding the original sound signal is broken up into smaller segments whose lengths equal the size of the message to be encoded. A Discrete Fourier Transform (DFT) is applied to each segment to create a matrix of the phases and Fourier transform magnitudes. Phase differences between adjacent segments are calculated. Phase shifts between consecutive segments are easily detected. In other words, the absolute phases of the segments can be changed but the relative phase differences between adjacent segments must be preserved. Therefore the secret message is only inserted in the phase vector of the first signal segment as equation (1):

$$\text{Phase new}=\begin{cases}\dfrac{\Pi}{2}, \text{if message bit=0}\\ -\dfrac{\Pi}{2}, \text{if message bit=1}\end{cases} \qquad (1)$$

A new phase matrix is created using the new phase of the first segment and the original phase differences. Using the new phase matrix and original magnitude matrix, the sound signal is reconstructed by applying the inverse DFT and then concatenating the sound segments back together.

4. VoIP COMMUNICATION

The term “VoIP” describes the digitalization, compression and transmission of analogue audio signals (in the majority of cases speech) from a sender to a receiver using IP packets. The receiver applies the reverse process and gets the reconstructed audio signal. After that he can act as the sender. For transmission the size of the used network and the distance between communications partners are of little relevance which means VoIP can be and already is used for worldwide telephony. Many applications of VoIP technology have been developed and are currently under development. For that reason embedding hidden messages in VoIP communication [1] is a very interesting task and may become subject of further studies.

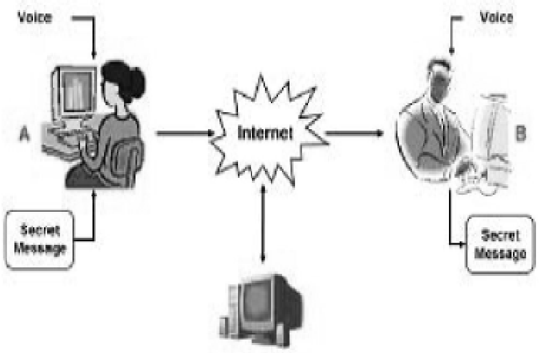


Fig.1 VoIP Communication

For describing that communications in a more formal manner consider the Fig.1 and the following statements. First sender and receiver choose from a set of codecs denoted by SC= {sc1, sc2, sc3... sci i€N} one audio codec for their communication. From the set of steganographic embedding techniques SE={se1,se2,se3,...,sei | i€N} A chooses one algorithm while B selects an according retrieving technique from SR={sr1,sr2,sr3,...,sri | i€N}.

If both techniques match B is able to reconstruct the message from A. In order to increase security the hidden message is encrypted by using a symmetric cryptographic scheme from the set of all cryptographic methods CM= {cm1, cm2, cm3... cmi | i€N}. For applying a cryptographic scheme a secret key K is necessary which is generated from a secret password [7].

The mapping of a password to a secret key of a fixed length is applied by choosing a cryptographic hash function. After encrypting the hidden message the message bits are uniformly distributed and spread over the whole audio stream by using a mixing algorithm. As input the mixing algorithm gets a pseudo random number which is generated by a pseudo random number generator (PRNG).

A. *Communication Flow*

VoIP is a real-time service that enables voice conversations through IP networks. It is possible to offer IP telephony due to four main groups of protocols:

a. *Signalling protocols* – that allow to create, modify, and terminate connections between the calling parties – currently the most popular are SIP, H.323 and H.248

b. *Transport protocols* – the most important is RTP, which provides end-to-end network transport functions suitable for applications transmitting real-time audio. RTP is usually used in conjunction with UDP (or rarely TCP) for transport of digital voice stream

c. *Speech codec's* – e.g. G.711, G.729, G.723.1 that allow compress/decompress digitalized human voice and prepare it for transmitting in IP networks.

d. Other *supplementary protocols* – like RTCP, SDP, or RSVP etc. that complete VoIP functionality. Generally, IP telephony connection consists of two phases: a *signalling phase* and a *conversation phase* [8].In both phases certain types of traffic are exchanged between calling parties.

5. SYSTEM DESIGN

A. *Stream Capturing*

Audio streaming is the method of delivering an audio signal to user system over the Internet. The " Stream Capturing " is designed to involve buffering in the learning experiences by enabling them to record and play back the sent/Received voices in the context - by recording the voice, the data's can be verified for all the sent and received packets, specific sounds like the noises are eliminated in calls, then hearing them as part of the narrative, VOIP can benefit from increased Voice quality and reduced Delays. Since VoIP calls travel digitally on computer networks rather than telecom cables, VoIP recording is done by tapping into the computer network rather than phone lines. Usually this is done by connecting to a router, switch, hub on the network, or through the PC attached to the VoIP phone. One way of doing this is by connecting to the SPAN (Switched Port ANalyzer) port which allows the VoIP recording unit to monitor all network traffic and pick out only the VoIP traffic to record.

B. *The VAD Algorithm*

To reduce network bandwidth in VoIP applications, some source codecs introduce silence compression during the inactive period of audio streams. The silence compression technique has two components: voice activity detection (VAD) and comfort noise generator [9].The VAD algorithm is used to decide whether the current audio frame is an active voice by comparing the energy of the frame (Enr) with a threshold (Thr) as shown in (2)

$$\mathrm{VAD}=\begin{cases} 1, & \mathrm{Enr} \geq \mathrm{Thr} \\ 0, & \mathrm{Enr} \leq \mathrm{Thr} \end{cases} \qquad (2)$$

VAD result = 0 means the frame is an inactive voice,
VAD result = 1 means the frame is an active voice.

The energy of the current frame Enr is computed using (3)

$$\mathrm{Enr}_t=\frac{1}{80}\sum_{60}^{239} e'_t{}^{2}(n) \qquad (3)$$

where e' t (n) is the output signal of the finite impulse response (FIR)filter whose input signal is the current frame {s[n]}n=60,….,239.The FIR filter computes e' t (n) using (4)

$$e'_t=s[n]+\sum_{i=1}^{10} a_{n0}[j].\,s[n-j], n = 60 \rightarrow 239 \qquad (4)$$

The threshold in (2), That, is given by

$$\mathrm{Thr} = \begin{cases} 5.012, & \text{If Nlev=128} \\ 10^{0.7-0.05\log_2 \frac{nlev}{128}}, & \text{If } 128<\text{Nlev}<16384 \\ 2.239, & \text{If Nlev} \geq 16384 \end{cases} \qquad (5)$$

Where Nlev is the noise size of the current frame.

C. *Definitions Of Inactive And Active Frames*

The silence compression technique is an optional function for the source codec. In fact, most source codecs do not use silence compression in VoIP applications. All audio frames are encoded uniformly by using the normal encoding algorithm regardless of whether they are active voices or inactive voices. Thus two types of frames are outputted when the speech stream F is encoded by the source codec. For example, ITU G.723.1 codec encodes the speech into two types of frames, active frames and inactive frames.

Definition 1: The active frame f^{*A} is encoded by the source codec from the active voice of the speech. It is expressed as

$$f_i^{*A}=\emptyset(f_i^{A}),\ i=0,\ldots,N_1 \qquad (6)$$

Definition 2: The inactive frame f^{*S} is encoded by the source codec from the inactive voice of the speech. It is expressed as

$$f_i^{*S}=\emptyset(f_i^{S}),\ j=0,\ldots,N_2 \qquad (7)$$

As the speech is divided into inactive voices and active voices by VAD, all the voices are encoded uniformly by the source codec to form audio frames, in which inactive frames can be distinguished from active frames.

D. Segregation Of Audio Frames

Our audio segregation model is illustrated in Fig. 4, where VAD, data embedding, and audio frame encoding are carried out sequentially in the speech coding process. The sender samples an audio signal and encodes it into a PCM formatted audio stream, F= $\{f_i|i=0,\ldots\ldots, N\}$. The VAD algorithm is then used to detect the inactive voice in the stream.

If the current frame f_i is an inactive voice, the frame is marked with S; otherwise, it is marked with A. As a result, the audio stream is divided into a sequence of frames shown in (8),

$$F=\{f_i^{\,A}, f_j^{\,S}|i=0,.., N_1, j=0,\ldots N_2, N= N_1 + N_2\} \qquad (8)$$

All the frames are then encoded uniformly by G.723.1 codec into a low bit rate stream, which is called the original speech shown in (9),

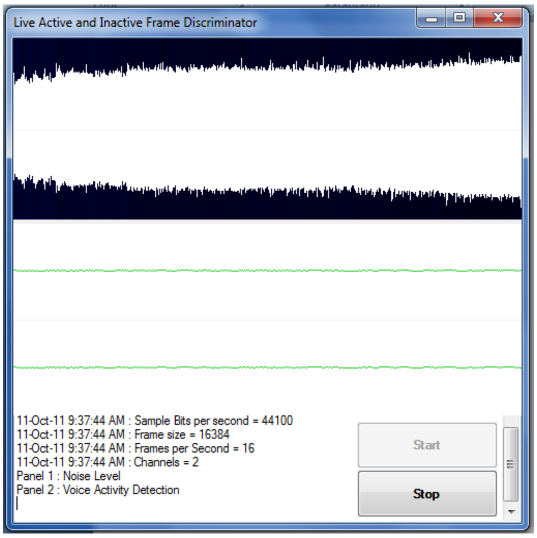
$$F^*=\{f_i^{*A}, f_j^{*S}|i=0,.., N_1 , j = 0,\ldots N_2, N= N_1 +N_2\} \qquad (9)$$

The low bit rate stream contains two types of frames, inactive frames and active frames. According to the frame type, two different steganography algorithms are then used, respectively, to embed the secret information, S in the stream. This study reveals that, VAD result of the inactive frames of VoIP streams are more suitable for data embedding than the active frames of the streams under the same imperceptibility.

E. Voice Active Detection

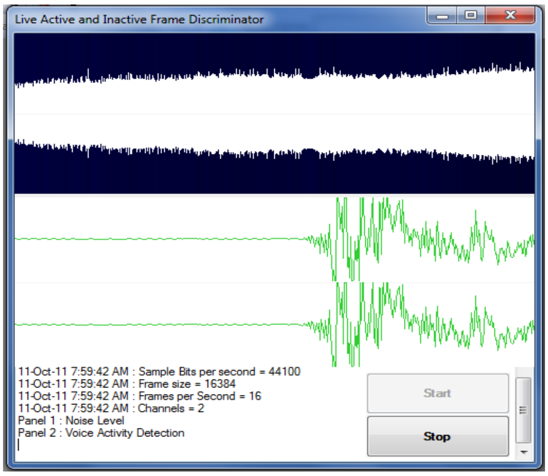
VAD is any important component of speech processing techniques such as speech enhancement, speech coding, and automatic speech recognition. Voice activity detection (VAD), also known as speech activity detection or speech detection, is a technique used in speech processing in which separating conversational speech from silence, music, noise or other non-speech signals. VAD is an important enabling technology for a variety of speech-based applications. Therefore VAD algorithm has been developed that provide varying features and compromises between latency, sensitivity, accuracy and computational cost.

The detection of inactive audio frame by VAD is shown in Fig.2 and active audio frame in Fig.3.The frame discriminator window consisting of two panels the first panel representing the noise level in the surroundings and the second panel representing the voice activity detection it uses 16 fps and a sampling rate of 44100 sample bps.



```
11-Oct-11 9:37:44 AM : Sample Bits per second = 44100
11-Oct-11 9:37:44 AM : Frame size = 16384
11-Oct-11 9:37:44 AM : Frames per Second = 16
11-Oct-11 9:37:44 AM : Channels = 2
Panel 1 : Noise Level
Panel 2 : Voice Activity Detection
```

Fig.2 Inactive audio frame



```
11-Oct-11 7:59:42 AM : Sample Bits per second = 44100
11-Oct-11 7:59:42 AM : Frame size = 16384
11-Oct-11 7:59:42 AM : Frames per Second = 16
11-Oct-11 7:59:42 AM : Channels = 2
Panel 1 : Noise Level
Panel 2 : Voice Activity Detection
```

Fig.3 Active audio frame

F. Steganography In Inactive Frames

Our steganography model is illustrated in Fig. 4. The low bit rate stream contains two types of frames, inactive frames and active frames. According to the frame type, different steganography algorithms are used, respectively, to embed the secret information, $S=(s_1,s_2,\ldots s_i,\ldots.,s_n),s_i€0$ in the stream.

The LSB algorithm is used to embed information in inactive frames.The low bit rate stream with hidden information is called the stego speech, denoted by, F'={f'$_1$,f'$_2$,….,f'$_i$},which is transmitted using VoIP. Afterwards, the receiver receives the stego speech, from which the secret information is finally extracted

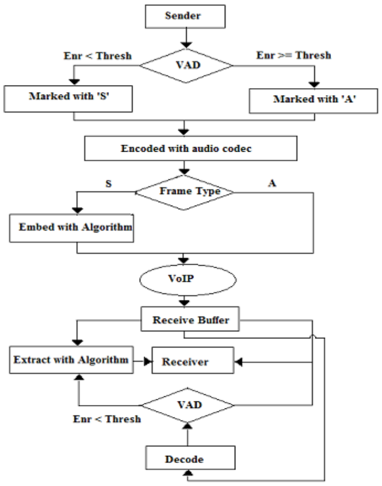


Fig.4.Steganography in inactive frames

1) Stego Data Embedding and Testing

For data embedding LSB (Least Significant Bit) Algorithm is used to encode the message into audio, it performs bit level manipulation to encode the message. Steps:

a. Receives the audio file in the form of bytes and converted in to bit pattern.

b. Each character in the message is converted in bit pattern.

c. Replaces the LSB bit from audio with LSB bit from character in the message

2) Stego Data Extraction

To extract a secret message from an LSB encoded sound file, the receiver needs access to the sequence of sample indices used in the embedding process. Normally, the length of the secret message to be encoded is smaller than the total number of samples in a sound file. The data extraction at the receiver's end is the inverse process of the embedding algorithm, and it is divided into the following three steps.

a. The stego audio with embedded text is received

b. Using LSB algorithm, the text from audio is extracted (otherwise called retrieved) the retrieved text is in the form of binary values

c. The extracted binary values from the audio which is then converted into ASCII value to get the secret text

6. PERFORMANCE EVALUATION

6.1 SPEECH QUALITY MEASUREMENT

The objective is to achieve a good quality of speech after embedding a secret message within the speech. The spectrum between the original speech and the stego speech in the frequency and time domain is compared. Perceptual Speech Quality Measurement of the Mean Cestrum Distortion [6] (MCD) metric is used to measure the quality of the original speech and stego speech. The Fig 6.1 and 6.2 shows the analysis of the original speech and the stego speech, first panel shows the waveform and the second panel shows the spectrogram.

$$\mathrm{MCD} = \frac{1}{N_f}\sum_{k=1}^{N_f}\frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{p}\left(c(i)-\tilde{c}(i)\right)^2}$$

Where

- ✓ N_f is the number of audio frames,
- ✓ c(i) and c˜(i) are the cepstrum coefficients of the original speech and the stego speech
- ✓ p is the order of c(i)

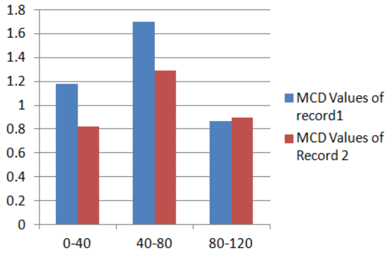


Fig 6.1 Comparison graph of the recorded speech

7. CONCLUSION

The proposed system is to provide a good, efficient method for hiding the data from the hackers and send to the destination in a safe manner. The system will assure the integrity of hidden messages in the case of packet loss. Thus this audio data hiding techniques can be used for a number of purposes other than covert communication or deniable data storage, information tracing and finger printing, tamper detection.This system achieves more suitable for embedding data capacity in inactive audio frames than in active audio frames.The experimental results have shown variances of the stego speech files were relatively small, indicating that the proposed steganography algorithm.

REFERENCES

[1] C. Krätzer, J. Dittmann, T. Vogel, and R. Hillert, "Designand evaluation of steganography for voice-over-ip," in Proc. IEEE Int. Symp.
Circuits Syst., May 2006, pp. 2397–3234.

[2] B. Xiao, Y. F. Huang, and S. Tang, "An approach to information hiding in low bit rate speech stream," in Proc. IEEE GLOBECOM 2008, Dec.2008, pp. 371–375, IEEE Press.

[3] Jhing-Fa Wang and Shi-Huang Chen, " A VAD based on Perceptual Wavelet Packet Transform and Teager Energy Operator"National cheng kung,Tainan,2004

[4] E.Verteletskaya,K.Sakhnov "Voice Activity Detection for Speech Enhancement Applications"ACTA Polytechnica Vol.50 No. 4,2010.

[5] Z.Wu and W. Yang (2006), "G.711-based adaptive speech information hiding approach," Lecture Notes Computer Science, vol. 4113, pp. 1139–1144.

[6] N. Aoki, "A technique of lossless steganography for G.711 telephony speech," in Proc. 2008 4th Int. Conf. Intelligent Inf. HidingMultimedia Signal Process. (IIH-MSP), Harbin, Aug. 2008, pp. 608–611.

[7] L. Ma, Z. Wu, and W. Yang, "Approach to hide secret speech information in G.721 scheme," Lecture Notes Comput. Sci., vol. 4681, pp.1315–1324, 2007.

[8] Wojciech Mazurczyk, Krzysztof, " Steganography of VoIP Streams", Institute of Telecommunications Polland.

[9] Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, ITU-T Recommendation G.723.1 Annex A, 2009.

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