The Steganography In Inactive Frames Of Voip

This paper describes a novel high-capacity steganography algorithm for embedding data in the inactive frames of low bit rate audio streams encoded by G.723.1 source codec, which isused extensively in Voice over Internet Protocol (VoIP). This study reveals that, contrary to existing thought, the inactive frames of VoIP streams are more suitable for data embedding than the active frames of the streams; that is, steganography in the inactive audio frames attains a larger data embedding capacity than that in the active audio frames under the same imperceptibility. By analysing the concealment of steganography in the inactive frames of low bit rate audio streams encoded by G.723.1 codec with 6.3 kb/s, the authors propose a new algorithm for steganography in different speech parameters of the inactive frames. An improved voice activity detection algorithm is suggested for detecting inactive audio frames taking into packet loss account. Experimental results show our proposed steganography algorithm not only achieved perfect imperceptibility but also gained a high data embedding rate up to 101 bits/frame, indicating that the data embedding capacity of the proposed algorithm is very much larger than those of previously suggested algorithms.

Keywords— Audio streams, inactive frames, steganography,

Voice over Internet Protocol (VoIP).

Introduction

Voice over Internet Protocol (VoIP) is a form of communication that allows you to make phone calls over a broadband internet connection instead of typical analog telephone lines. Basic VoIP access usually allows you to call others who are also receiving calls over the internet. Interconnected VoIP services also allow you to make and receive calls to and from traditional landline numbers, usually for a service fee. Some VoIP services require a computer or a dedicated VoIP phone, while others allow you to use your landline phone to place VoIP calls through a special adapter.

Streaming media, such as Voice over Internet Protocol (VoIP) streams, are broadcast live over the Internet and delivered to end-users. Security remains one of the main challenges with this new technology. With the upsurge of VoIP applications available for use in recent years, VoIP streams become one of the most interesting cover objects for modern steganography. 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. For example, Wu et al. [1] suggested a G.711-based adaptive speech information hiding approach.Aoki [2] proposed a technique of lossless steganography in G.711 encoded speeches. Ma et al. [3] framed a steganography method of embedding data in G.721 encoded speeches. 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.

However, 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. Chang et al. [4] embedded information in G.729 and MELP audio streams. Huang et al. [5] proposed a steganography algorithm for embedding information in low bit rate audio streams. But these steganography algorithms have constrains on the data embedding capacity; that is, their data embedding rates are too low to have practical applications.

Thus the main focus of this study was to work out how to increase the data embedding capacity of steganography

in low bit rate audio streams. The rest of this paper is organized as follows. Section II summarizes some related work, discussing the possibility of embedding data in the inactive frames of low bit rate audio streams. In Section III, the imperceptibility of the steganography algorithm for embedding data in the inactive audio frames is analyzed. Our

proposed steganography algorithm is presented in Section IV.

Section V details the experimental setup and performance evaluation results. Finally, the paper ends with conclusions and directions for future work in Section VI.

RELATED WORK

The analysis by synthesis (ABS)-based speech information hiding approach was adopted to embed speech data in an original speech carrier, with good efficiency in steganography and good quality of output speech.

Krätzer, Dittmann, and Vogel [8] argued that the inactive

voice of a speech was not suitable for being used as a cover object for steganography owing to an obvious distortion of the original speech. By contrast, Huang et al. [9] suggested an algorithm for embedding information in some parameters of the speech frame encoded by ITU G.723.1 codec, without leading to distinction between inactive voices and active voices. It seems that Krätzer, Dittmann, and Vogel's opinions [8] and Huang and coworkers' results [9] contradict each other. Such a contradiction can be attributed to the different speech codecs that were used to compress and encode audio signals. In [8], audio streams were encoded by a pulse-code modulation (PCM) codec; but an ITU G.723.1 source codec was used to encode audio streams in [9]. The PCM codec is based on the waveform model that samples, quantizes, and encodes audio signals directly; the sample value represents the original volume of the signal. In this case, the inactive voice cannot be used to embed information since it will lead to obvious distortion. However, the source codec is a hybrid codec, which is based on the

source model. This codec compresses the speech at a very low bit rate and performs on a frame-by-frame basis; each frame is encoded into various parameters rather than the sample volumes. Thus the volume of the speech does not change imperceptibly even though their inactive audio frames contain hidden information.

The theoretical analysis above suggests that steganography in the inactive frames of low bit rate audio streams would attain a larger data embedding capacity if an appropriate steganography algorithmwere used. The rest of this paper details our successful effort on such a new steganography algorithm for embedding data in the inactive frames of low bit rate audio streams encoded by ITU G.723.1 source codec.

PRINCIPLE OF STEGANOGRAPHY IN INACTIVE

AUDIO FRAMES

Hangover Algorithm for Detecting Active Voices

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 [10]. The VAD 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 (1)

(1)

means the frame is an inactive voice; otherwise, the frame is an active voice.

The energy of the current frame is computed by

(2)

where is the output signal of the finite impulse response (FIR) filter whose input signal is the current frame The FIR filter computes using (3)

(3)

where is the autocorrelation coefficient vector of the filer.

The threshold in (1), Thr, is given by

(4)

where is the noise size of the current frame, and is updated by its previous value, the energy of the previous frame , and the self-adaptive flag . is defined as follows:

(5)

where , and is limited to a value between 128 and 131 071. In general, the Hangover algorithm is used for detecting inactive voices to avoid noise peaks being extended [10]. If an audio frame is determined to be an inactive voice, the frame is encoded into a silence insert description (SID) frame by using the silence compression algorithm. Having received the SID frame, the decoder generates a comfortable noise at the receiving end.TheHangover algorithm is illustrated in Fig. 1.

Fig. 1. Illustration of Hangover algorithm

The first row in Fig. 1 shows the classification of voice duration before silence compression. Hcnt is the Hangover-frame number of inactive voiceswhen an active voice begins to change to an inactive voice in the speech. The second row is an estimate of the energy and the third rowincludes the corresponding codec algorithms.

An audio stream is actually divided into frames before being encoded. For instance, with G.723 codec the audio stream is divided into frames 30ms in length. Suppose the audio stream contains frames, .If the energy (Enr) of the frame is less than the threshold value (Thr), the frame is the first frame of an inactive voice. This frame is defined as aHcnt frame in Hangover algorithm and is then encoded by using the normal codec algorithm rather than the silence compression algorithm. If subsequent frames are still inactive voices, the Hangover algorithmwill not perform silence compression until the sixth frame. In other words, the Hangover algorithm starts to encode the sixth frame of the inactive voice into a SID frame until the next active voice emerges. The first five frames (first to fifth) of the inactive voice are still encoded into Hangover frames, denoted by . The active voice of the audio stream is encoded into active frames by using the normal codec algorithm. According to the Hangover algorithm, audio frames are classified into three types, active voice frame , Hangover frame , and silence compression frame . The audio speech can be expressed as

(6)

The speech is then encoded into by using Hangover algorithm, which can be written as

(7)

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 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, without using the silence compression algorithm.

Definition 1: The active frame is encoded by the source

codec from the active voice of the speech. It is expressed as

(8)

Definition 2: The inactive frame is encoded by the source codec from the inactive voice of the speech. It is expressed as

(9)

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. Combining (6)–(9) yields

(10)

C. Bit Distribution Patterns of Inactive Frames

This section discusses whether the “1/0” distribution pattern

of an inactive frame is similar to that of an active frame if the

inactive voice of a speech is encoded into inactive frames.

First, we analyzed the statistical probability of “1/0” presentation in inactive frames. Assuming an audio stream is

divided into frames, among them there are inactive

frames and active frames, i.e.,

The audio stream is denoted by

.

Suppose each frame consists of bits, namely

The average probability of “1” presentation in all the inactive frames is computed by using

(11)

where denotes the th “1” in the th inactive frame of the

stream. So the average probability of “0” presentation in all the inactive frames is given by

(12)

TABLE I

AVERAGE PROBABILITIES OF “1” PRESENTATION IN INACTIVE AND ACTIVE FRAMES

Table I lists the experimental results of the statistical probabilities of “1” presentation in the inactive frames and active frames encoded by G.723.1 codec, respectively. Ten speech sample files were used for the experiments, with each file being tested six times in order to work out the average probabilities of “1” presentation in the inactive and active frames.

Finally, we studied the run-length statistical character of

“0/1” in inactive frames. The run-length statistical method

was used to calculate the run-lengths of continuous “0” or “1”

presentation in inactive frames. Assuming the audio stream is

denoted by

and , it satisfies the following equation:

(13)

where , and denotes the run-length of in an inactive frame.

Then the run-length of in the inactive frame is equal to the number of bits from to . The distribution pattern of the run-length in inactive frames is defined as the probability of various run-lengths presenting in all inactive frames of the speech file, given by

(14)

(15)

where , denotes the percent of the run-length of the bit “0” or “1” being equal to in all inactive frames, and and denote the numbers of the run-length of the bit “0” or “1” being equal to in all inactive frames, respectively.

Table II describes the distribution patterns of the run-lengths

of “0” and “1” in all inactive frames and active frames, and the M-W-W test results for comparing the probability distributions between the inactive frames and active frames for four speech samples, respectively. Since for all the cases, we conclude that the probability distributions for both inactive frames and active frames do not differ, indicating that the inactive frames and active frames had a similar run-length pattern for each speech file.

To summarize, the above three experiments on bit distribution patterns indicate that the bit distribution of inactive frames is similar to that of active frames for the same speech files. Otherwise stated, it is highly unlikely to use the “1/0” distribution pattern to distinguish the inactive frames from active frames of low bit rate audio streams.

TABLE II

RUN-LENGTH PATTERNS IN INACTIVE AND ACTIVE FRAMES

D. Steganography in Inactive Frames

The source codec like ITU G.723.1 is operated on a frame-by-frame basis. Each frame encoded by G.723.1 codec has 240 audio samples that are encoded according to PCM. First of all, each frame is filtered by a high-pass filter to remove the dc component and is then divided into four subframes of 60 samples each. A tenth order linear predictive coding (LPC) filter is computed using the unprocessed input signal for every

subframe, and the last subframe is quantized using a predictive split vector quantizer. For every two subframes (120 samples), the weighted speech signal is used to compute the open-loop pitch period. A harmonic noise shaping filter is then constructed using the open-loop pitch period computed previously, and a closed-loop pitch predictor is constructed according to theimpulse response created by the noise shaping filter.

Finally,both the pitch period and the differential value are transmitted to the decoder and the nonperiodic component of the excitation is approximated. After completion of these operations, all speech parameters such as LPC, Pulse sign (Pamp), and Pulse position (Ppos) and so on, are obtained. The next step is to determine which speech parameters of inactive frames are suitable for data embedding. All the speech parameters are sorted into three imperceptibility levels of steganography in terms of the distance of signal-to-noise ratio (DSNR), which is defined as the difference in signal-to-noise ratio (SNR) between the original speech and stego speech. Close analysis of the data in Table V shows the imperceptibility levels of steganography for different parameters of

the inactive frames are widely different. So it is possible to choose different parameters and various parameter bits to embed data on demand of practical applications. In short, the parameters marked with level 1–2 are suitable cover objects for steganography.

To sumup, the steganography process has three subprocesses, voice activity detection, data embedding, and extracting. The corresponding algorithms are detailed below.

Improved VAD Algorithm

Hangover algorithm is normally used for voice activity detection in the speech coding process. To synchronize the embedding and extraction in steganography, it is very important to keep the VAD result consistent between the sender and receiver because an inconsistent VAD result will result in errors in the extracting process. Some factors, such as packet loss, steganography and so on, may have an impact on the VAD result. So an improved VAD algorithm called the residual energy method is suggested below.

The residual energy method adopts the autocorrelation coefficient, which is not affected by the state of the codec, to detect the inactive voice in the speech.

Fig. 2. Flowchart of steganography in inactive and active frames.:

The embedding process is shown in Fig. 2.

. PERFORMANCE ANALYSIS OF STEGANOGRAPHY

IN

INACTIVE FRAMES

Fig. 3. DSNR for steganography in various parameters of inactive frames

Fig. 3 shows the results of experiments on the 20 speech files listed in Table VI, with the horizontal axis representing the number of bits of the parameter that are replaced by secret information. Experiments indicate that in most instances the DSNR value between the original speech and the stego speech was so small that the distortion of the stego speech was unlikely to be perceived as long as appropriate parameter bits of inactive frames were used to embed the secret information. The overall trend in DSNR was upward with increasing bit numbers of embedding. The parameters with DSNR values of less than 0.5 dB were chosen to embed information.

.

Fig. 4. Spectrum comparisons in the time- and frequency-domain.

Objective Quality: To evaluate further the imperceptibility of the stego speech, we compared the spectrum between the original speech and the stego speech in the frequency and time domain. For instance, the spectrums of theMC1 speech file having 183 inactive frames with and without hidden information are shown in Fig. 4.

Careful analysis of Fig. 4 shows very little distortion occurred in the time domain as a result of data embedding in inactive frames; however, we could not perceive any differences between the original speech and the stego speech in the frequency domain. This suggests steganography in inactive frames at a data embedding rate of 101 bits/frame had no or very little impact on the quality of the original speech.

Fig. 5. Comparisons in data embedding rates between the proposed algorithm

“HiF” and other algorithms

As Fig. 5 shows, the data embedding rate of our proposed algorithm “HiF” was much higher than those of the other algorithms. This is because the proposed steganography algorithm made good use of the redundancy in the inactive frames of low bit rate audio streams.

It is worth mentioning that the data embedding capacity of steganography in inactive frames is limited by the number of inactive frames of the original speech file. Research found 30%–50% of a VoIP session were inactive frames, so steganography in the inactive frames could attain a higher data

embedding rate than other algorithms, which is in agreement with our experiment results.

CONCLUSION

In this paper, we have suggested a high-capacity steganography algorithm for embedding data in the inactive frames of low bit rate audio streams encoded by G.723.1 source codec. The experimental results have shown that our proposed steganography algorithm can achieve a larger data embedding capacity with imperceptible distortion of the original speech, compared with other three algorithms. We have also demonstrated that the proposed steganography algorithm is more suitable for embedding data in inactive audio frames than in active audio frames. However, before the proposed algorithm comes into practical use in covert VoIP communications, it is necessary to explore how to assure the integrity of hidden messages in the case of packet loss, which shall be the subject of future work.