Journal of Information Hiding and Multimedia Signal Processing ©2016 ISSN 2073-4212
Ubiquitous International Volume **7**, Number **5**, September **2016**

A Synchronization Scheme for Hiding Information in Encoded Bitstream of Inactive Speech Signal

Rong-San Lin

Dept. of Computer Science and Information Engineering,
Southern Taiwan University of Science and Technology
No.1, Nan-Tai Street, Yungkang Dist., Tainan City 710, Taiwan
rslin@stust.edu.tw

Received February, 2016; revised April, 2016

Abstract. *This paper proposes a high-capacity hiding algorithm for embedding data in the inactive frames of low-bit rate speech bit-stream encoded by the G.723.1 speech codec. An improved voice activity detection (VAD) algorithm is implemented to accurately detect inactive speech frames. We propose a hiding scheme that can correctly extract secret information in the receiver. The scheme uses the VAD result as a flag to synchronize the embedding and extraction process in stenography. We also show that the encoded bits of the stochastic excitation pulse parameters are more suitable for data embedding than the encoded bits of other speech parameters. The results of an imperceptibility evaluation indicate that the average perceptual evaluation of speech quality (PESQ) score is degraded only slightly, by 0.027, relative to that of original speech that is not hidden data. Experimental results show that our proposed hiding algorithm not only achieves perfect imperceptibility but also produces a high data-embedding capacity, up to 2128 bits per second. The standard objective and subjective quality measurements verify that our proposed hiding algorithm can achieve a larger data-embedding capacity with imperceptible distortion than those of previously suggested algorithms.*

Keywords: Imperceptibility; Information hiding; Speech Bitstream; G.723.1, VAD.

1. Introduction.

In recent years, with the explosive growth of Internet use and multimedia technology, people are enjoying great convenience in their communication. Increasing amounts of information are being converted into digital formats in order to accelerate the distribution and exchange of data. Information hiding is a technique for embedding a secret message into a carrier object, such as an image or audio, for the purposes of concealing that secret message. The output of such an operation is called a stego-object, which is transmitted through the channel. The receiver can extract the secret message from the stego-object. To securely transmit the data, the embedded objects must have the characteristic of being imperceptible. Currently, most hiding techniques that use a speech signal as a carrier can only use a low compression rate, such as waveform coding. The G.723.1 and G.729 [1, 2] speech codecs can achieve high voice quality and low bit rate. They are already being used extensively on the Internet, such as in the Voice-over-Internet Protocol (VoIP) communication system. Information hiding in a speech signal thus provides a valuable means of secret transmission, as speech communication is so popular in daily life. Information hiding in a low-bit rate speech bitstream is commonly regarded as a challenging topic in the field of data hiding.

Several steganography methods for embedding data in a speech bitstream or images have been proposed in the existing literature. For example, Ito *et al.* [3] presented a data-hiding technique for the G.711 speech codec based on the least significant bits (LSBs) substitution method. Wang *et al.* [4] proposed a method for information hiding in a real-time VoIP stream. Above methods adopted high-bit rate speech bitstreams encoded by waveform coding as cover objects, in which quantities of LSBs exist. However, VoIP are usually transmitted over low-bit rate speech bitstreams encoded by the source codec, such as the ITU-T G.723.1 codec. Chang *et al.* [5] embedded information in a mixed-excitation linear prediction (MELP) and a G.729 encoded bitstream. Liu *et al.* [6] presented a hiding algorithm based on a vector quantization method. However, this two hiding algorithms have constraints on the data embedding capacity; in other words, their data embedding rates are too low to have practical applications. Huang *et al.* [7] suggested an algorithm for embedding data in the inactive frames of VoIP streams encoded by a source codec. Lin [8] proposed an approach for hiding information in encoded bits of speech signal. It must be noted that different speech codecs were used to compress and encode speech signals in the above approaches. In Ref. [3], speech signals were encoded by a pulse-code modulation (PCM) codec, which belongs to a waveform coding structure that samples, quantizes, and encodes speech signals directly; the sample value represents the original volume of the speech signal. In this structure, inactive speech also cannot be used to embed information since it will result in an obvious distortion in speech quality. However, the G.723.1 codec is a hybrid codec, which is based on a source coding structure. This codec compresses the speech at a very low bit rate and on a frame-by-frame basis; each frame is encoded into various parameters rather than sample volumes, and then these parameters are quantized into a bitstream and transmitted to the decoder. Thus, the volume of the speech does not change perceptibly, even though their inactive frames contain hidden information. Consequently, when an appropriate hiding algorithm is used, the inactive frames of the speech signal are more suitable for embedding data, attain a larger data-embedding capacity, and have the least effect on speech quality. In addition to speech signal as cover objects, vector quantization can be applicable to image-based watermarking methods. Yang *et al.* [9] proposed a reversible data hiding scheme based on the integer wavelet transform. Huang *et al.* [10] integrated error-resilient coding into the watermarking algorithm. In this paper, we also propose a hiding scheme that can correctly extract secret information in the receiver.

Thus, the purpose of this study is to realize a high-capacity hiding technique for embedding data in an encoded bitstream of inactive speech frames. The rest of this paper is organized as follows. In Section 2, the G.723.1 codec bit allocations are briefly reviewed. Our proposed hiding algorithm is described in Section 3. The experiments and performance evaluation results are presented in Section 4. Conclusions are presented in Section 5.

2. ITU-T G.723.1 Speech Codec.

This codec operates on frames (30 ms) of 240 samples each. The set of LPC of the last sub-frame is converted to ten line spectrum pair (LSP) parameters; these ten LSP parameters are divided into three sub-vectors with dimensions of 3, 3, and 4. Each sub-vector is vector-quantized using an 8-bit codebook. Every subframe speech signal is then encoded by the adaptive codebook (ACB) and fixed codebook (FCB) search procedures. The closed-loop pitch lag is computed as a small differential value around the open-loop pitch lag estimate or the previous subframe pitch lag. In other words, for subframes 0 and 2, the closed loop pitch lag is selected from around the appropriate open-loop pitch lag in the range of 1 to +1 and coded using 7 bits. For subframes 1 and 3, the closed-loop pitch lag is coded differential using 2 bits,

TABLE 1. Bit allocations of the 6.3 and 5.3 kbit/s coding algorithm

Parameters coded	Subframe 0	Subframe 1	Subframe 2	Subframe 3	Total
LPC (LSP)	-	-	-	-	24
Pitch lags (Olp)	7	2	7	2	18
Total gain	12	12	12	12	48
Pulse positions (Ppos)	20(12)	18(12)	20(12)	18(12)	73(48)
Pulse signs (Pamp)	6(4)	5(4)	6(4)	5(4)	22(16)
Grid index	1	1	1	1	4
Total encoded bits					189(158)

and may differ from the previous subframe lag only by 1, 0, 1, or 2. For the FCB search procedures, the codec adopts multi-pulse maximum likelihood magnetization (MP-MLQ) and the algebraic CELP (ACELP) for high rate (6.3 kbit/s) and low rate (5.3 kbit/s) codings, respectively. Therefore, a speech frame is encoded using G.723.1 codec with the encoded bits 189 and 158 bits for 6.3 kbit/s and 5.3 kbit/s coding algorithm, respectively. Thus, this codec has two bit rates and their bit allocations are listed in Table 1.

3. The Proposed Information Hiding Scheme.

To accurately detect inactive speech frames and correctly extract secret information, we implement an improved VAD algorithm and use the VAD result (*Vad*) as a flag, to indicate whether the current bitstream packet is hidden information or not in our hiding scheme.

3.1. Improved VAD Algorithm used in the G.723.1 Codec.

The G.723.1 codec employs a silence compression technique to reduce network bandwidth in VoIP applications and it is an optional function for the G.723.1 codec. The silence compression technique uses a VAD algorithm [1] to determine whether the current speech frame is an active voice frame by comparing the energy of the frame (*Enr*) with a threshold (*Thr*):

$$Vad = \begin{cases} 1, & Enr \geq Thr \\ 0, & Enr < Thr \end{cases} \tag{1}$$

where *Vad=0* means the frame is an inactive voice frame; otherwise, the frame is an active voice frame.

The VAD algorithm of the G.723.1 codec merely computes the residual energy of the current frame and then compares it with a threshold value as the deciding criterion. To further improve the accuracy of the G.723.1 VAD algorithm, we implemented in the G.723.1 codec a VAD algorithm that had been used in the G.729 codec [2] in our experiments. It can accurately detect inactive speech frames and increase the capacity of the embedding data. The implemented VAD algorithm is called an improved VAD algorithm. We note that the frame size and coding algorithm of two codecs are not the same. The improved VAD algorithm calculates four difference measures of the speech parametric features. Four difference measures are generated from the current frame vector and the running averages of the background noise as follows: The spectral distortion measure is generated as the sum of the squares of the difference between

the current frame $\{LSF_i\}_{i=1}^{p}$ vector and the running averages of the background noise $\{\overline{LSF_i}\}_{i=1}^{p}$:

$$\Delta S = \sum_{i=1}^{p} \left(LSF_i - \overline{LSF_i} \right)^2 \tag{2}$$

where *p=10* is the order of LPC, . The full-band energy difference measure:

$$\Delta E_f = \overline{E_f} - E_f \tag{3}$$

The low-band energy difference measure:

$$\Delta E_l = \overline{E_l} - E_l \tag{4}$$

The zero-crossing difference measure:

$$\Delta ZC = \overline{ZC} - ZC \tag{5}$$

With the four difference measures $(\Delta S, \Delta E_f, \Delta E_l, \Delta ZC)$ acquired, a multi-boundary decision regions rule is adopted in the improved VAD algorithm to divide the speech signal into two types of frames: active frames and inactive frames. This decision rule is detailed in Ref. [2]. Equations (2),(3),(4),(5) show that the improved VAD algorithm computes the four parametric features of the current frame as a detection criterion. Therefore, it can achieve high accuracy in detection.

3.2. **Synchronized Embedding and Extraction Process Scheme.** In our proposed hiding method, we first modify the silence compression function of the G.723.1 codec, so all inputted speech frames are encoded uniformly using the normal encoding algorithm, regardless of whether they are active voice frames or inactive voice frames, so they have the same bit allocation. In other words, the VAD algorithm is merely used to classify the speech signal into two types of frames, active frames (*Vad =1*) and inactive frames (*Vad =0*) before being encoded, and our proposed hiding method is performed after the encoded process. Our proposed hiding scheme is illustrated in Figure 1, where VAD, speech frame encoding, and the embedding and extracting algorithms are performed sequentially in the hiding procedure. Firstly, the sender samples a speech signal and encodes it into a PCM-formatted speech frame. The VAD algorithm is then used to detect the inactive voice frame. If the current frame is an inactive voice frame, then the frame is marked with *Vad =0*; otherwise, it is marked with *Vad =1.* As a result, the speech signal is divided into a sequence of frames. All the frames are then encoded uniformly by the G.723.1 codec into a low-bit rate encoded bitstream, which comprises encoded bits of the speech parameters, such as listed in Table 1; LSP, pulse signs (Pamp), and pulse positions (Ppos) and so on. When *Vad =0*, the hiding algorithm is performed to embed the secret information, $S = (s_1, s_2, ..., s_n)$, $s_i \in (0, 1)$ in the encoded bitstream of the inactive frame. When *Vad =1*, the encoded bitstream is an active voice frame and secret data is not embedded. Huang *et al.s* [7] proposed method used the original G.723.1 VAD approach to embed and extract information in the encoded bitstream of the inactive frame. In their method, the encoded bitstream packet is first decoded in the receiver by the G.723.1 decoder to obtain decoded speech, which is then detected by the VAD algorithm to decide whether the current frame is an inactive frame. If the decoded speech is an inactive frame, then secret information is extracted from the encoded bitstream packet. It must be noted that the encoded bitstream of the parameters of the inactive speech frame were substituted with bits (s_i) of the secret information, $S = (s_1, s_2, ..., s_n)$,$s_i \in (0, 1)$ in the embedding procedure. Therefore, it is possible that the inactive speech frame will become an active speech frame in the receiver, resulting in a missed extraction of secret information using Huang *et al.*s proposed approach. Therefore, it is very important to keep the VAD results consistent between the sender and receiver, because an inconsistent VAD result will cause errors in the extraction process.

For the G.723.1 codec, the encoded bits of the speech parameters are encapsulated in a standard bitstream format, as shown in Figure 2. The packet head (Ftyp, 2 bits) is used to instruct the current coding model, which contains four types: high rate, low

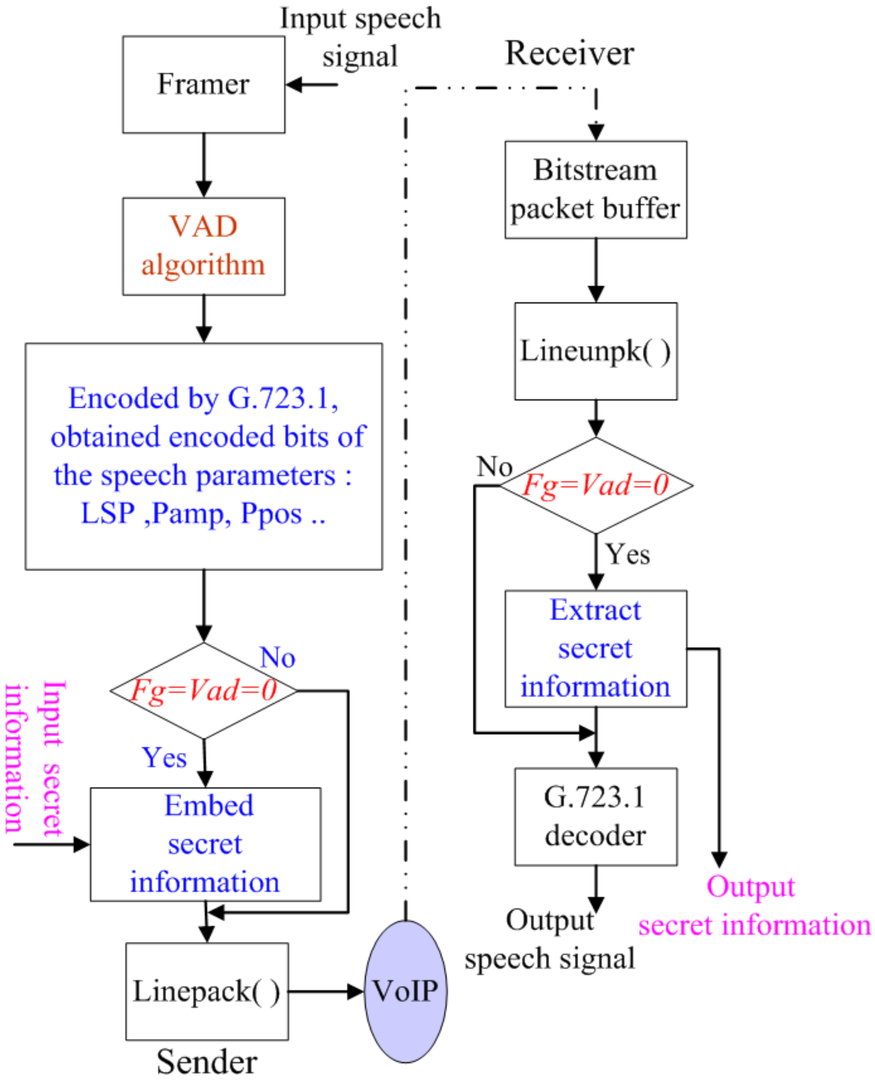


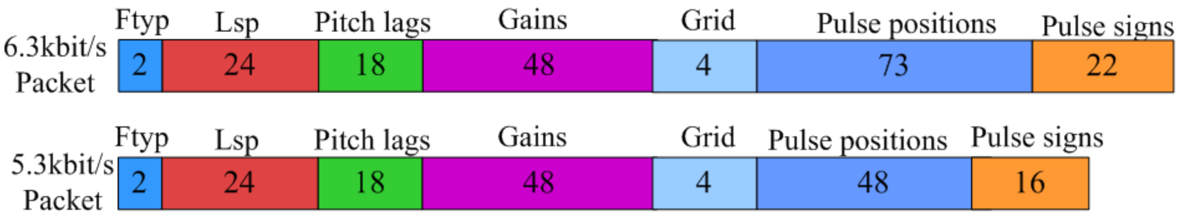
FIGURE 1. Flowchart of the proposed hiding scheme

TABLE 2. Expression of the packet head (Ftyp)

Ftyp	Coding model
00	with hidden data & high rate
01	with hidden data & low rate
10	without hidden data & high rate
11	without hidden data & low rate

rate, silence insertion description (SID) frame, and untransmitted. It is worth mentioning that the silence compression technique is an optional function for the G.723.1 codec. We disable the silence compression function in our hiding method, so the coding model does not contain SID frame, and untransmitted type. Therefore, the packet head has one redundant bit.

To synchronize the embedding and extraction process in steganography, we propose an information-hiding scheme that can correctly extract secret information The scheme uses the VAD result *(Vad)* as a synchronization flag(*Fg*), which is embedded in the packet head (*Ftyp*, redundant bit) to indicate whether the current bitstream packet is hidden information or not.



	Ftyp	Lsp	Pitch lags	Gains	Grid	Pulse positions	Pulse signs
6.3kbit/s Packet	2	24	18	48	4	73	22
5.3kbit/s Packet	2	24	18	48	4	48	16

Figure 2. Bitstream standard packet format of the encoded speech parameters

We modified the above instruction of packet head, as shown in Table 2; high rate, low rate, and with ($Fg = 0$) and without ($Fg = 1$) hidden information. The encoded bitstreams with and without hidden information are called the stego-bitstream and the pure-bitstream, respectively.

3.3. **Embedding and Extraction Algorithms.** The embedding process is divided into four steps, as shown in Figure 1.

Step 1): VAD. The speech signal in PCM format is divided into frames, and each frame is inputted into the VAD detector that adopts the four parametric features algorithm described above. The frame is marked with $Fg = 0(Vad = 0)$ if it is determined to be an inactive voice frame; otherwise, the frame is marked with $Fg = 1(Vad = 1)$.

Step 2): Encode all frames with the G.723.1 codec. Regardless of the frame type, all the frames are then encoded using the standard G.723.1 algorithm with 6.3 or 5.3 kbit/s bit rates. The resulting encoded bits of the speech parameters containing active and inactive frames are then outputted from the speech encoder.

Step 3): Embedding secret messages in an inactive frame. When Fg =0, the proposed hiding algorithm is performed to embed the secret messages, $S = (s_1, s_2, ..., s_n), s_i \in (0, 1)$ in the encoded bitstream of the inactive frame. When *Vad* =1, the encoded bitstream is an active voice frame and secret data is not embedded.

Step 4): Encapsulation and sending. The packet head (*Ftyp*) and the encoded bits of the speech parameters of the inactive and active frames are encapsulated in encoded bitstream packets regardless of whether they are hidden information or not. These encoded bitstream packets are transmitted over the Internet to the receiver.

The linepack and lineunpk functions of the G.723.1 codec are used to encapsulate and de-encapsulate the encoded bits of the speech parameters at the sender and receiver, respectively. The extraction of secret information from the encoded bitstream packet is the inverse process of the embedding algorithm. It is divided into the following two steps.

Step 1): Receiving and de-encapsulation. The encoded bitstream packets are received, buffered, and then de-encapsulated using the lineunpk function in the receiver. Step 2): Extracting secret messages. The receiver receives the encoded bitstream packet, and then detects the packet head. If the packet head (*Ftyp*) is 00 or 01, which indicates the current encoded bitstream packet with hidden information, then the secret information is directly extracted from the stego-bitstream packet, so the proposed method can keep the VAD results consistent between the sender and receiver.

Either the stego-bitstream or pure-bitstream packet is then decoded by the G.723.1 decoder in the receiver to obtain decoded speech, which is called the stego-speech (with embedded data) or the original speech (without embedded data), respectively. In summary, the embedded and extracted information can synchronize processes between the sender and receiver using our proposed algorithm, and thus the proposed hiding scheme can correctly extract secret information. It must be noted that our proposed scheme does not execute a VAD algorithm in the receiver.

To provide a security of secret messages, there are many encryption algorithms can be used, such as DES (Data Encryption Standard) algorithm. It is worth mentioning that our proposed method uses hiding techniques that are fully compatible with the original G.723.1 speech codec. In other words, the encoded bitstreams with and without hidden information have the same packet size and bitstream format. Therefore, any eavesdropper cannot extract the secret messages, even though suspecting the existence of a secret message, because eavesdropper doesn't know bitstream packet format of the speech codec and those encoded bits of the speech parameters with hidden messages.

4. **Performance Analyses.** In our experiments, twenty-three test speech files were employed as cover objects for steganography. They can be downloaded from the website $ftp : //www.itu.int/rec/T - REC - P.862 - 200102 - I/en$. These speech files were tested to evaluate the performance of the proposed methods. We first evaluate the performance of the improved VAD algorithm. The experimental results are shown in Table 3. After comparing the number of inactive frames of the improved VAD method with those of the original G.723.1 VAD approach, it is obvious that the average number of inactive frames of the former algorithm is about 17 higher than that of the latter approach for each speech file. It must be noted that the Huang *et al.*s [7] proposed method used the original G.723.1 VAD approach. Therefore, the improved VAD method can increase the data embedding capacity.

Next, we investigate the importance of the encoded bits of the speech parameters to find the LSBs, which are more suitable for embedding data. To evaluate the stego-speech quality objectively we adopted the ITU-T P.862 recommendation [11]. The recommendation describes an objective method for predicting the subjective quality of narrowband speech codecs. Two parameters, imperceptibility and data-embedding capacity, were used to evaluate the performance of the proposed hiding algorithm.

4.1. **Imperceptibility Evaluation of the Speech Parameters.** In this section, we evaluate the imperceptibility of hidden information in encoded bits of the speech parameters, the same secret information (a text file) was embedded in the encoded bits of the twenty-three speech files, and the degradation of PESQ (DPESQ) values of the resulting stego-speech files were then computed to evaluate the imperceptibility in our experiments. The DPESQ is defined as the difference in PESQ between the stego-speech and the original speech (decoded speech), and is given by:

$$\mathrm{DPESQ} = \mathrm{PESQ_S} - \mathrm{PESQ_O} \tag{6}$$

where $\mathrm{PESQ_S}$ and $\mathrm{PESQ_O}$ are the PESQ values of the stego-speech and the original speech, respectively.

Figure 3 shows the results of experiments on the twenty-three speech files listed in Table 3, with the horizontal axis representing the number of bits of the speech parameter that are replaced by secret information. Experimental results indicate that in most instances the DPESQ value between the original speech and the stego-speech was so small that the distortion of the stego-speech was unlikely to be perceived, provided that appropriate encoded bits of the speech parameters of the inactive frames were used to embed the secret information. As shown in Figure 3, when the bit number of hidden information in the Olp (pitch lags) parameter was not more than 5 bits, the DPESQ value was under 0.005; however, the DPESQ value rose significantly when more than 6 bits of information were embedded. This means that no more than 5 bits of information should be embedded in the Olp parameter. Finally, we only replaced two LSBs of the Olp parameter with secret information (embedding 1 bit in each even sub-frame). For the Gains parameter, when no more than 2 encoded bits were embedded in the LSBs (amounting to 8 bits of

Table 3. Comparison of the number of inactive frames of the improved VAD method with those of the original G.723.1 VAD approach

Speech file name	file_length (s)	number of inactive frames (G.723.1)	number of inactive frames (improved)
or105	8.4	149	168
or109	8.01	131	152
or114	8.58	161	175
or129	7.23	130	148
or134	8.07	158	176
or137	7.05	126	142
or145	8.25	124	141
or149	8.16	134	152
or152	7.17	83	89
or154	7.2	131	147
or155	7.38	131	145
or161	7.89	124	142
or164	7.59	111	120
or166	7.98	127	144
or170	8.22	137	152
or179	8.28	132	148
or221	8.1	121	141
or229	8.01	128	145
or246	7.86	144	160
or272	8.13	132	150
u_am1s01	7.98	160	180
u_am1s02	7.98	151	173
u_am1s03	7.98	147	168
Average	7.89	133	150

Table 4. Bit numbers of the parameters perfectly suited to data embedding

Parameter name	Olp	LSP	Gains	Grid	Ppos	Pamp	Total bits
Number of bits	2	3	8	4	73	22	112

hidden information per frame), the DPESQ value was under 0.007; however, the DPESQ value rose significantly when more than 8 bits of information were embedded. For the LSP parameter, we replace three LSBs in each inactive frame with secret information. In this case, the DPESQ values were under 0.0049. By examining the DPESQ values in Figure 3, we realized that all encoded bits of the $Ppos$, $Pamp$, and Grid parameters of the inactive frames are more suitable for data embedding than the encoded bits of other speech parameters. In other words, the encoded bits of the stochastic excitation pulse parameters are more suitable for data embedding. Therefore, we selected bit numbers of the parameters that were perfectly suited to embedding data, as shown in Table 4, for conducting further hiding experiments. (Ppos = H_Ppos + L_Ppos, 6.3 kbit/s)

Next, we evaluate the imperceptibility of hidden information in various parameters of inactive and active frames. In the following figures (Figures. 4, 5, and 6), the numbers below each bar (ex. 2 bits) indicate the bit numbers of the speech parameters that are replaced by secret information. The "Total" bar represents the total number of replaceable parameter bits in an active or inactive frame. Figure 4 shows the experimental results of using the G.723.1 codec with 6.3 kbit/s, and the average DPESQ values for data embedding in various speech parameters of active frames. Similarly, Figure 5 shows the experimental results for secret information embedded in various speech parameters of inactive frames.

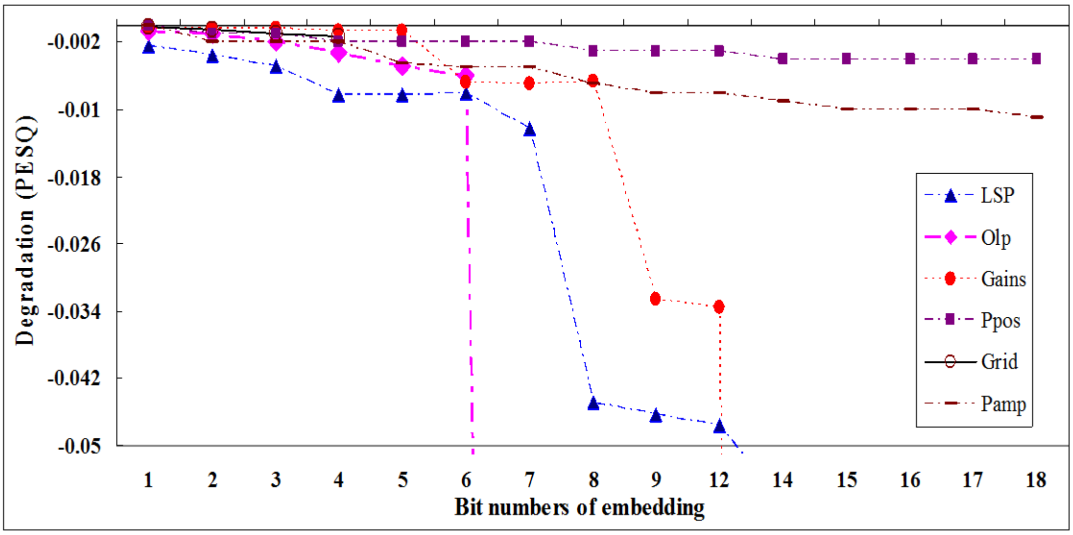


Figure 3. DPESQ values for hidden information in various parameters of inactive frames (at 6.3 kbit/s).

As a frame of G.723.1 with 6.3 kbit/s has 189 bits, and the total number of replaceable speech parameter bits in an inactive frame is 112 bits, the data-embedding capacity ratio EC_r for an inactive frame is:

$$EC_r = \text{Embedding bits/Total bits} = 112/189 = 59.3\% \tag{7}$$

From the experimental results shown in Figures 4 and 5, we found that the average DPESQ values of embedding 112 bits of data in active and inactive frames are 1.023 and 0.027, respectively. It is obvious that the inactive frames of a speech signal are more suitable for data embedding than the active frames of a speech signal; that is, hiding information in the inactive frames attains greater imperceptibility than in the active frames under the same data-embedding capacity ratio. Next, using the 5.3 kbit/s coding as an experimental platform, secret information was embedded in various speech parameters of inactive frames, and the experimental results are shown in Figure 6. A frame of G.723.1 with 5.3 kbit/s has 158 bits, and the total number of replaceable speech parameter bits in an inactive frame is 81 bits. Thus the data-embedding capacity ratio $EC_r = 51\%$ and the average DPESQ value is 0.023.

From the experimental results shown in Figures 5 and 6, we found that the 6.3 kbit/s coding is more suitable for data embedding than the 5.3 kbit/s coding. That is, hiding information in the former attains a larger data-embedding capacity ratio than in the latter with almost the same imperceptibility.

4.2. **Speech Quality Evaluation.** Since evaluating the quality of stego-speech by directly rating the mean opinion score (MOS) is difficult for non-expert participants in this area. Instead, the MOS listening quality objective (MOS-LQO) score specified by ITU-T P.862.1 [12] can evaluate the speech quality. ITU-T P.862.1 has mapped the raw P.862 scores (PESQ values) to MOS-LQO scores. Next, we performed speech quality evaluation

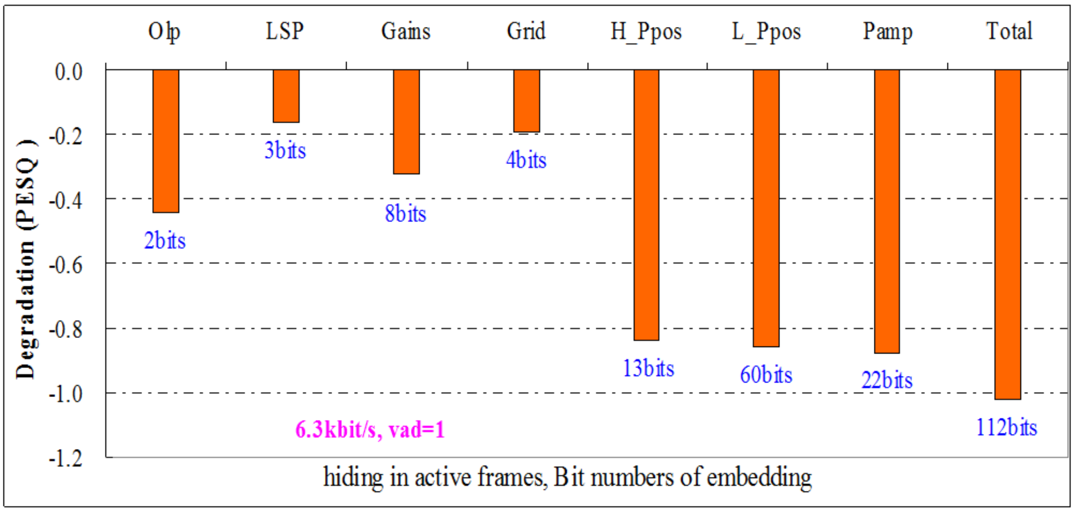


Figure 4. Average DPESQ values for data embedding in various parameters of active frames.
(ex. Olp, 2 bits + LSP, 3 bits + ... + Pamp, 22 bits = Total, 112 bits) .

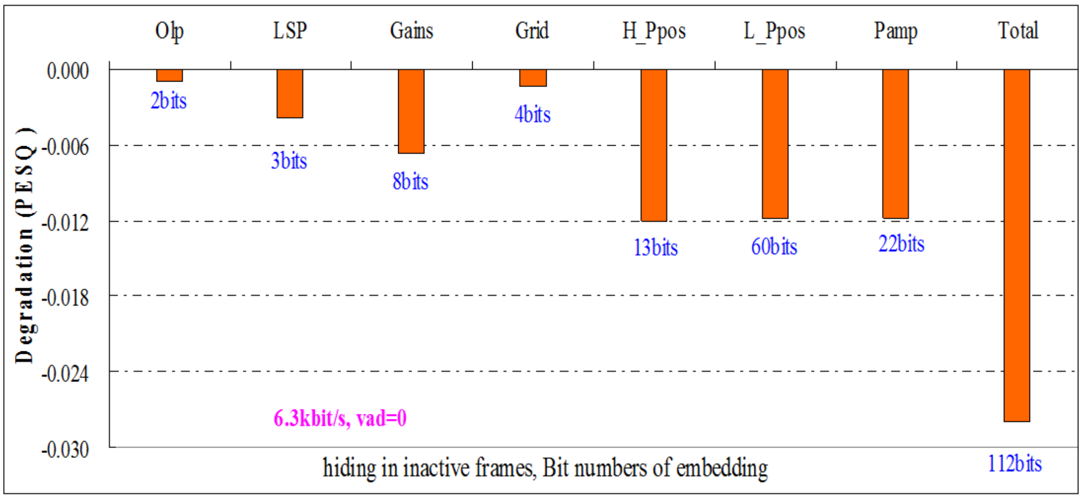


Figure 5. Average DPESQ values for data embedding in various parameters of inactive frames

based on the 6.3 kbit/s bit rate. The testing results of using the ITU-T P.862.1 method are listed in Table 5. It is observed that the difference in MOS-LQO between the original speech and the stego-speech was so minor (3.2%) that distortion resulting from hiding in inactive frames was imperceptible. Comparing the average degradation MOS-LQO values of the proposed method with the Huang *et al*.s [7] approach, we found that both the degradation MOS-LQO values are nearly the same.

We also computed the LPC mean cepstrum distortion (MCD) [13] to measure the objective quality of the stego-speech. The MCD is defined as:

$$MCD = \frac{1}{N}\sum_{j=1}^{N}\frac{10}{\ln(10)}\sqrt{2\sum_{i=1}^{p}\left(c(i)-\hat{c}(i)\right)^2} \quad (8)$$

where N is the number of speech frames, and $c(i)$ and $\hat{c}(i)$ are the LPC cepstrum coefficients of the original speech and the stego-speech, respectively, and p is the order of

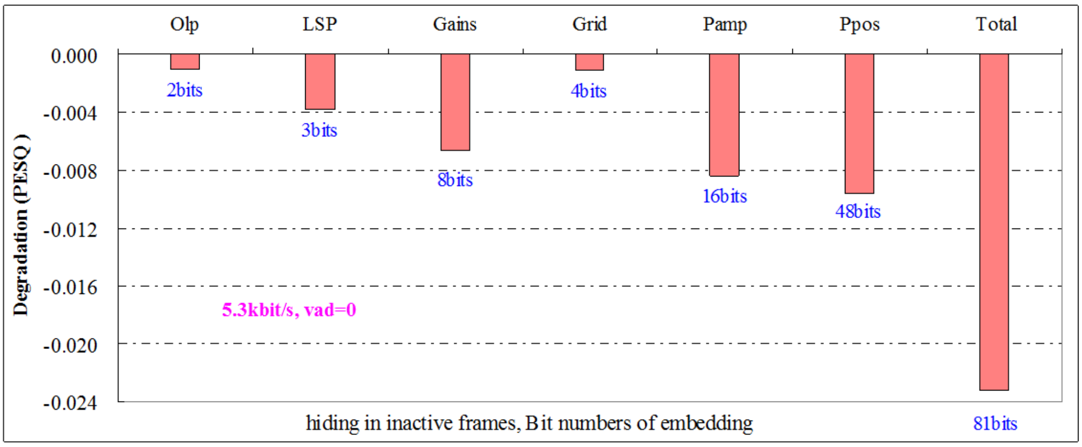


Figure 6. Average DPESQ values for data embedding in various parameters of inactive frames.

Table 5. Testing results with ITU-T P.862.1 and MCD

Speech file name	Original speech MOS-LQO	Stego-speech (proposed) MOS-LQO	Stego-speech MCD
or105	3.808	3.792	1.16
or109	3.778	3.759	0.83
or114	3.807	3.786	1.10
or129	4.034	3.985	0.97
or134	4.033	3.990	1.16
or137	3.996	3.953	1.09
or145	3.590	3.577	1.01
or149	3.942	3.908	0.86
or152	3.698	3.670	0.86
or154	3.913	3.874	0.92
or155	3.964	3.938	1.16
or161	3.920	3.894	0.86
or164	3.865	3.820	0.85
or166	3.932	3.890	0.82
or170	4.002	3.982	0.91
or179	3.758	3.726	0.97
or221	3.807	3.769	0.91
or229	3.906	3.878	0.88
or246	4.058	4.018	1.00
or272	3.965	3.942	0.86
u_am1s01	3.223	3.210	1.15
u_am1s02	3.668	3.629	1.23
u_am1s03	3.713	3.652	0.93
Average	3.843	3.811	0.98

LPC. Table 5 lists the *MCD* results of the twenty-three stego-speech files with information embedded in the inactive frames. It is observed that all the *MCD* values of the stego-speech files are very small, indicating that the proposed hiding algorithm for embedding information in the inactive frames achieved perfect imperceptibility.

Table 6. Percentage of judgment failures

Judgment	Listener 1	Listener 2	Listener 3	Listener 4	Listener 5	Average
failures	50%	46%	47%	49%	53%	49%

We implement a simple and informal subjective quality evaluation called the $A/B/X$ test in this paper. This test method is described as follows: Suppose there are three types of speech files, denoted by A, B and X, respectively. A represents the stego-speech file containing hidden information, B represents the original speech file without hidden information, and X is either A or B. Five non-expert evaluators were invited to listen these speech files listed in Table 5, using a headset. The A and B speech files were first played and then the X speech file was played, these untrained listeners were asked to decide whether *X is A or B*. The X speech file was chosen randomly from the A or B speech file. Each listener made 23 judgments in total. Table 6 shows the percentage of failures that listeners identify the incorrect stego-speech files, and the average percentage of failure judgments was 49%. Results imply that the listeners cannot distinguish the quality of the stego-speech from that of the original speech.

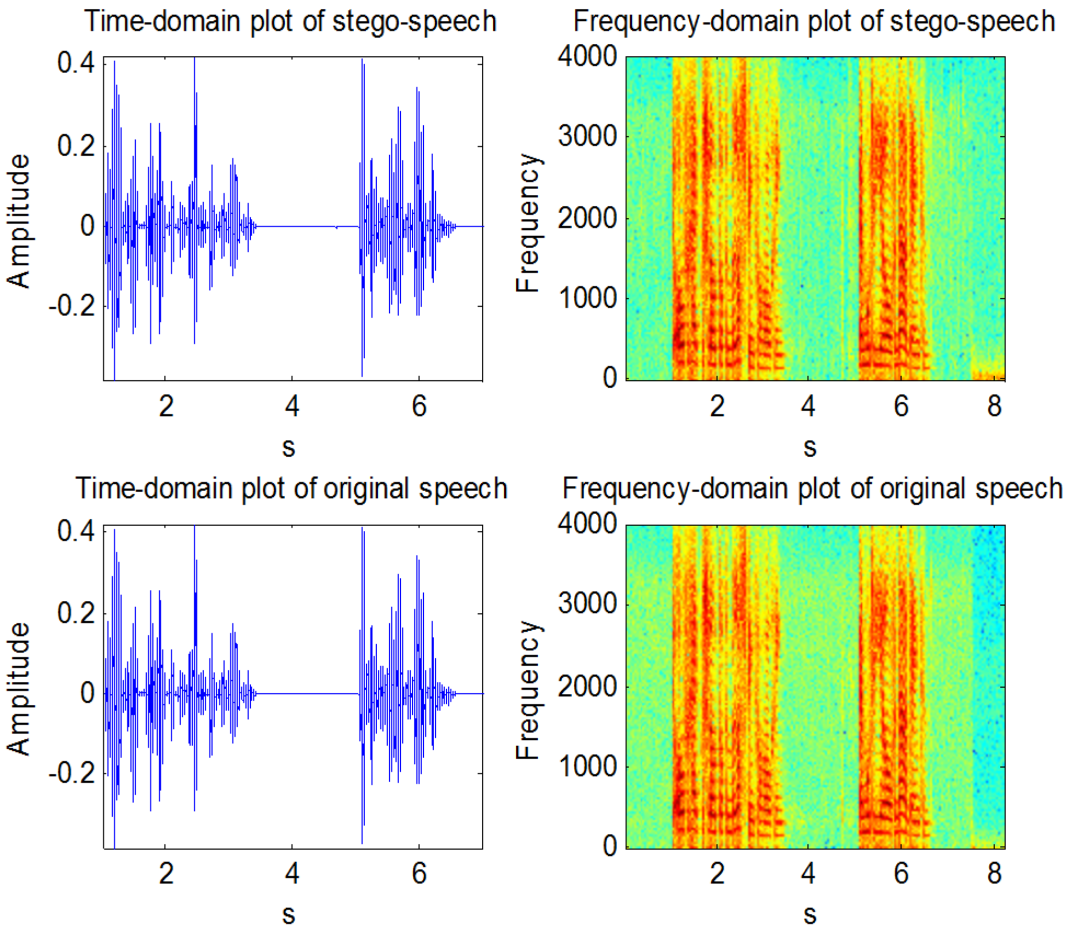


Figure 7. Comparison of spectra in the frequency and time domains

To further evaluate the imperceptibility of the stego-speech, we compared the distortion of the original speech and that of the stego-speech in the frequency and time domains. For example, the spectra of the or145 speech file, having 141 inactive frames with and without hidden information, are shown in Figure 7. We scarcely observe any distortion in the time domain. We also cannot perceive any differences between the original speech and the stego-speech in the frequency domain. This indicates that hiding information in inactive frames at a data-embedding capacity of 112 bits/frame had very little or no impact on the quality of the stego-speech.

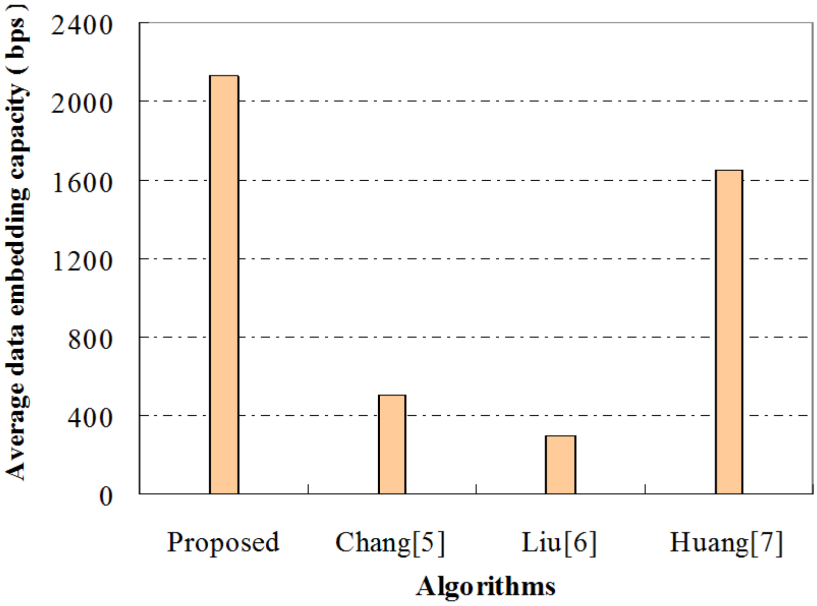


Figure 8. The data embedding capacity of the proposed algorithm in comparison with that of the other existing algorithms

Additionally, we also compared the data embedding capacities of the proposed algorithm and the other algorithms. Suppose the average frame number of the original speech signal per second is L and the number of inactive frames is D, then the number of active frames is $L - D$. The bit numbers of the information embedded in an inactive and an active frame for our proposed algorithm are $B_n = 112$ and $M = 0$ bits, respectively. From Table 3, we obtain the average number of inactive frames as $D = 19(150/7.89)$ per second. The average data embedding capacity of the speech file can be obtained as E_c in bits per second (bps)$E_C = D \times B_n + (L - D) \times M =$ 2128 bits relative to the data-embedding capacity rate, on average 33.8 %. This study compares the proposed method with several hiding information algorithms [5, 6, 7]. Figure 8 compares the data embedding capacities of our proposed algorithm and the other existing algorithms. Note the data embedding capacities of the existing [5], [6] and [7] schemes shown in Figure 8 are referred to their paper. As Figure 8 shows, the data embedding capacity of our proposed algorithm is much higher than that of the other existing algorithms.

To accompany the MOS-LQO and MCD tests described above, we have made the decoded sound files and the demo files available at *http://faculty.stust.edu.tw/ rslin/hiding.htm* for subjective evaluation by listening and demonstration.

5. **Conclusions.** In this paper, we proposed a high-capacity information-hiding algorithm and implemented an improved voice activity detection algorithm that can correctly extract secret information and accurately detect inactive frames. The improved VAD method can increase the average number of inactive frames by about 17, relative to the original G.723.1 VAD approach for each speech file. Additionally, to synchronize the embedding and extraction process in steganography, we propose a hiding scheme that can correctly extract secret information. The scheme uses the VAD result as a synchronization

flag. The difference in the MOS-LQO between the original speech and the stego-speech was minor (3.2%), indicating that the proposed hiding algorithm achieved perfect imperceptibility. For a data-embedding capacity rate 33.8%, the average PESQ value is degraded only slightly, by 0.027, relative to that of the original speech that is not hidden data. The experimental results show that our proposed hiding algorithm can achieve a larger data-embedding capacity with imperceptible distortion of the stego-speech, compared with other three algorithms.

Acknowledgment. This work was supported by the Ministry of Science and Technology of Republic of China under research contract MOST 104-2221-E-218 -018. The author would like to thank the reviewers for their valuable comments and suggestions.

REFERENCES

[1] I. ITU, 723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, *Telecommunication Standardization Sector of ITU*, 1996.

[2] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications, *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997.

[3] Y. SUZUKI *et al.*, Information hiding for g. 711 speech based on substitution of least significant bits and estimation of tolerable distortion, *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 93, no. 7, pp. 1279–1286, 2010.

[4] C. Wang and Q. Wu, Information hiding in real-time voip streams, in *Multimedia, 2007. ISM 2007. Ninth IEEE International Symposium on*, pp. 255–262, IEEE, 2007.

[5] P.-C. Chang and H.-M. Yu, Dither-like data hiding in multistage vector quantization of melp and g. 729 speech coding, in *Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on*, vol. 2, pp. 1199–1203, IEEE, 2002.

[6] J.-X. Liu, Z.-M. Lu, and H. Luo, A celp-speech information hiding algorithm based on vector quantization, in *Information Assurance and Security, 2009. IAS'09. Fifth International Conference on*, vol. 2, pp. 75–78, IEEE, 2009.

[7] Y. F. Huang, S. Tang, and J. Yuan, Steganography in inactive frames of voip streams encoded by source codec, *IEEE Transactions on information forensics and security*, vol. 6, no. 2, pp. 296–306, 2011.

[8] R.-S. Lin, An imperceptible information hiding in encoded bits of speech signal, in *2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, pp. 37–40, IEEE, 2015.

[9] C.-Y. Yang, C.-H. Lin, and W.-C. Hu, Reversible data hiding for high-quality images based on integer wavelet transform, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 3, no. 2, pp. 142–150, 2012.

[10] H.-C. Huang, S.-C. Chu, J.-S. Pan, C.-Y. Huang, and B.-Y. Liao, Tabu search based multi-watermarks embedding algorithm with multiple description coding, *Information Sciences*, vol. 181, no. 16, pp. 3379–3396, 2011.

[11] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs, in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2, pp. 749–752, IEEE, 2001.

[12] R. P. ITU-T, 862.1: Mapping function for transforming p. 862 raw result scores to mos-lqo, 2003.

[13] N. Kitawaki, H. Nagabuchi, and K. Itoh, Objective quality evaluation for low-bit-rate speech coding systems, *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 242–248, 1988.