An objective method for evaluating data hiding in pitch gain and pitch delay parameters of the AMR codec

Akira Nishimura[1]

[1]*Department of Media and Cultural Studies, Tokyo University of Information Sciences, Chiba, Japan*

Correspondence should be addressed to Akira Nishimura (akira@rsch.tuis.ac.jp)

ABSTRACT

Audio data hiding technology has several applications in the field of distribution, communication, and audio data trading. Steganographic use of audio data hiding enhances the quality and quantity of audio data communication. On the other hand, embedding hidden data may degrade the perceptual quality of the audio signal. Three methods for hiding data in pitch-related parameters of the advanced multi rate (AMR) narrow-band speech codec were evaluated in terms of the objective quality degradation and the bit rate of the embedding data. Computer simulations of the data hiding system were conducted for the AMR 12.2-kbps and 7.95-kbps modes. The results revealed that the method of replacing the least significant bit (LSB) of the pitch gain parameter with the information bits was superior in terms of embedding bit rate and less sound quality degradation than other methods, which use LSBs of the pitch delay data.

1. INTRODUCTION

The most general application of audio data hiding technology is copyright protection of the audio data, which is called watermarking. Watermarking technology requires robustness with respect to modifications of the watermarked audio signals caused by transmission via various audio media. The modifications are, for example, transcoding by perceptual audio codecs, AD/DA conversions, additive noises, low-pass filtering, and malicious modification attacks for piracy distribution. The size of watermarking data can be small, because the copyright or authentication data is coded efficiently.

Another essential application is steganography, which involves embedding additional data that may or may not be related to the contents of the audio data. Since the embedded data is usually not audible and the human listener is unaware of its existence, the data can be used to enhance the quality and quantity of audio data communication. In such applications, the embedded data can include annotation and semantic description of the audio data, multimedia data, bandwidth extension or packet loss concealment of the speech codec, and hidden channel communication. The size of the additional data is required to be as large as possible in order to increase the range and efficacy of the application.

The most important issue in both watermarking and steganography technologies is the perceptual transparency of the embedded audio signal. In other words, no perceptual quality degradation should be found in the embedded audio signal.

Data hiding in speech data encoded by a speech codec has been considered to be useful for steganography in order to enhance speech communication. A number of methods have been proposed to embed hidden data into encoded speech data.

Most of these studies performed objective measurement of speech quality degradation using segmental signal-to-noise ratio (SNR), which exhibits the level of the reference speech signal relative to the level of the noise components induced by transcoding in short segments. However, modern Code Excited Linear Prediction (CELP) based speech codecs, such as LD-CELP (ITU-T Rec. G.728) , CS-ACELP (ITU-T Rec. G.729), and Advanced Multi Rate (AMR) codecs [1], reconstruct perceptual oriented speech waveforms and have relatively small SNRs of approximately 10 dB. Consequently, a small difference in SNR obtained between the standard codec and the modified codec for data hiding does not truly reflect a small perceptual difference between two codecs.

Perceptual evaluation for speech quality (PESQ) is an alternative method of objective sound quality evaluation for speech codecs that is recommended in ITU-T Rec. P.862 [2]. PESQ compares an original signal with a signal that has been degraded by passing through a communications system. The key to this process is the transformation of both the original and degraded signals to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system, taking into account the perceptual frequency (Bark) and loudness (Sone). The transformed output of PESQ, which is defined in ITU-T Rec. P.862-1, is called the mean opinion score listening quality objective (MOS-LQO) and corresponds to the results of mean opinion score listening quality subjective (MOS-LQS) obtained from human listeners by the subjective experiments.

In the present paper, least significant bit (LSB) based data hiding methods in pitch delay or pitch gain parameters of the AMR codec are evaluated in terms of the capacity of hidden data and the objective quality of decoded speech signals.

2. AMR NARROW-BAND SPEECH CODEC

A large number of 3rd Generation Partnership Project (3GPP) based cellular phones adopt the AMR speech codec. The encoder converts 20 ms of an 8-kHz and 13-bit digital waveform frame into Line Spectral Pair (LSP) parameters, pitch parameters, algebraic code index, and gain parameters. These parameters are transmitted using the selective bit rate mode from 4.75 to 12.2 kbps. The coding scheme for the multi-rate coding modes is the Algebraic Code Excited Linear Prediction (ACELP) coder [3].

A simplified block diagram of the encoding process is depicted in Fig. 1. At first, spectral features of the framed speech signal are quantized as LSP parameters. Then, pitch analysis extracts the pitch delay of the waveform and the gain of the periodical excitation. Finally, the combination of the algebraic pulse positions, their polarities, and a gain are suitably selected by minimizing the residual excitation of the remainder of the periodical pitch excitation in the speech waveform.

Table 1 shows the bit allocation of the AMR coding algorithm for the three typical modes. LSP parameters are encoded once for every 20-ms frame and the other parameters are encoded once for every 5-ms subframes. In the 12.2-kbps and 7.95-kbps modes, the pitch gain and the codebook gain are separately quantized. In other modes, moving averaged prediction from the previous frames and vector quantization are applied to the combined pitch and codebook gain parameters (see the bottom of Figure 1). Except for the 4.75-kbps and 5.15-kbps modes, the pitch delay parameters of the second and fourth subframes are represented as the difference from the nearest integer value of the pitch delay of the previous subframe.

Mode		subframes			
(kbps)	Parameter	1st	2nd	3rd	4th
12.2	2 LSP sets	38			
	Pitch delay	9	6	9	6
	Pitch gain	4	4	4	4
	Algebraic code	35	35	35	35
	Codebook gain	5	5	5	5
10.2	LSP set	26			
	Pitch delay	8	5	8	5
	Algebraic code	31	31	31	31
	Gains	7	7	7	7
7.95	LSP set	27			
	Pitch delay	8	6	8	6
	Pitch gain	4	4	4	4
	Algebraic code	17	17	17	17
	Codebook gain	5	5	5	5

Table 1: Bit allocation of the AMR codec.

3. DATA HIDING IN ENCODED SPEECH DATA

3.1. General methods for embedding

Several methods have been proposed to embed hidden data into the encoded speech parameters. Although embedding data into the LSP parameters [4] may be robust against DA/AD conversion, the sound quality is severely degraded. Embedding data into the fixed codebook index by selecting a labeled codebook table is effective in the high-bit- rate mode [5, 6], because several bit allocations for the codebook table make the fixed pulse positions redundant. These techniques inevitably require integration of the embedding unit and the standard speech encoder.

In the present study, the embedding methods in pitch delay and pitch gain parameters are examined. Quantized pitch delay and pitch gain parameters simply correspond to the physical quantities, the fundamental period (inverse of frequency), and the intensity of the voiced part of the speech signal. Therefore, embedding in the bit

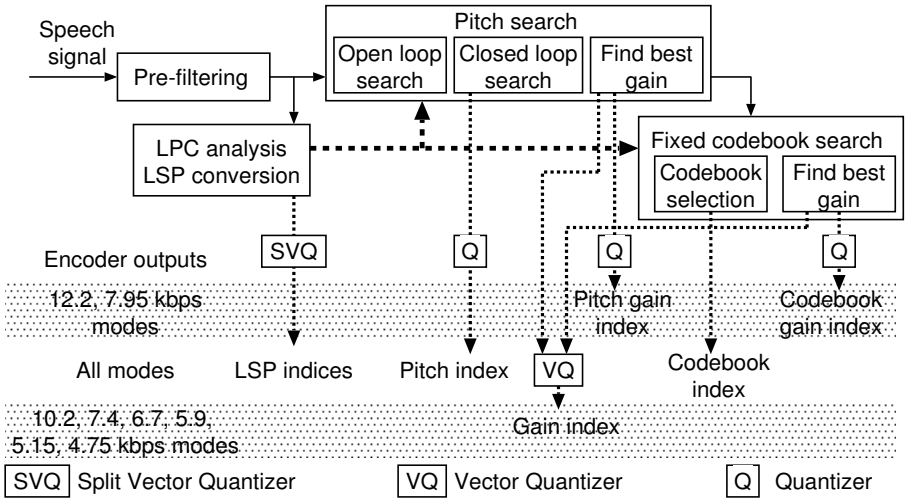

Fig. 1: Simplified block diagram of the AMR encoder.

stream output of the standard AMR encoder can be implemented while presuming quality degradation caused by modifying the bit value at a suitable location. This takes advantage of the use of the standard encoder and an additive embedding unit posterior to the standard encoder.

3.2. Methods for embedding in pitch parameters

Three methods of data hiding into the pitch related data of the AMR codec are evaluated in terms of embedding data capacity and objective sound quality.

Iwakiri proposed a method of hiding data in a LSB of the pitch delay parameter in an ITU-T Rec. G.723.1 based speech codec. Replacing the LSB of the quantized pitch delay data with hidden data for every subframe achieved an embedding rate of 134 bps. If the voice activity detection (VAD) and discontinuous transmission (DTX) functions of the AMR codec are not activated, embedding all 5-ms subframes achieves a maximum embedding rate of 200 bps. This method is hereafter referred to as the pitch LSB (PLSB) method.

Sasaki et al. proposed data hiding in the pitch delay parameter based on the pitch gain value in a CELP based speech codec [7]. If the pitch gain value is less than the threshold value, the apparatus embeds hidden data by replacing the LSBs of the pitch delay data. Assigning a higher threshold value and a wider bit width of the LSB increases the capacity of the embedding data. This method is hereinafter referred to as the gain threshold pitch LSB (GTPLSB) method. The GTPLSB method can be simply applied to the AMR encoder for the 12.2-kbps and 7.95-kbps modes because these modes have separate pitch gain parameters. In other modes, however, vector quantization is performed jointly using the pitch gain parameter and the codebook gain parameter, which is predicted from previous frames. The embedding algorithm may be rather complex and computationally overloaded, except for the 12.2-kbps and 7.95-kbps modes. For this reason, only the 12.2-kbps and 7.95-kbps modes were evaluated in the present paper.

Another simple method of embedding data into the pitch data is LSB replacement of the pitch gain data. In the same way as for the PLSB, embedding all subframes achieves a maximum embedding rate of 200 bps. This method is hereinafter referred to as the pitch gain LSB (PGLSB) method. For the same reason as for the GTPLSB, the 12.2-kbps and 7.95-kbps modes were tested.

Figure 2 shows the data hiding system for a speech codec via a phone network. The above three methods can be implemented to either modify the standard encoding algorithm (see Fig.2 Integrated implementation) or modify the output bit stream of the standard encoder (see Fig.2 Separated implementation). The latter case has the advantage of a simple structure. The former case is considered to be advantageous for the PLSB. Implementing the embedding algorithm in the closed pitch search section allows the fixed codebook search section to optimize and

reduce residual errors caused by modification to the pitch delay value. Other methods have this advantage only in the 12.2-kbps mode because other modes determine the quantized pitch gain values depending on the process of the fixed codebook and gain search (not shown in Fig. 1 for simplicity). In the following section, which describes the computer simulation, the effect of this optimization is examined by comparison between the results of integrated and separated implementations.

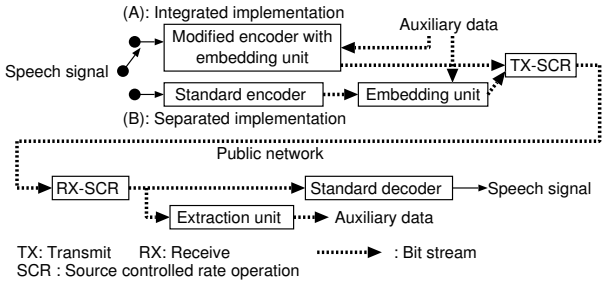

Fig. 2: Data hiding system for a speech codec via a phone network..

3.3. Extraction of embedded data from encoded speech data

Extraction of the embedded data from the encoded speech data is rather simple compared with extracting the robust watermark from the music signals. The bit stream of the input of the speech decoder is analyzed to find hidden bit locations according to the embedding rule. No modification to the standard speech decoder is required (Fig.2). The bit stream including hidden data bits is sent to the standard speech encoder at the same time. This may cause quality degradation of the decoded speech signals. Therefore, an important method to implement data hiding in encoded speech data is to locate bit locations that are not significant from a perceptual standpoint.

4. COMPUTER SIMULATION

4.1. Measurement of quality degradation using PESQ

PESQ was adopted to evaluate the objective quality degradation caused by data hiding in the reference speech signals obtained by 16-bit quantization and 8-kHz sampling. A total of 550 phonetically balanced sentences spoken by 22 Japanese speakers (12 men and 10 women) were fed into the input of the AMR encoder with an embedding unit. These sentences were generated by concatenating two sentences from 1,100 sentences selected from the Continuous Speech Database for Research (Vol. 1) published by the Acoustical Society of Japan. The duration of the speech ranged from 6 to 12 seconds, including silence intervals. The overall level of each input speech signal was –26 dBov. Then, the output bit stream of encoded speech data was fed into the standard AMR decoder. PESQ software distributed by ITU-T was applied to the decoded speech signal.

In addition, the reference speech signals were fed into the standard AMR encoder and decoder, which is distributed by 3GPP organization partners [8], and the PESQ software. The MOS-LQO difference in the decoded speech signal between data hiding and the standard AMR transcoding is considered to be a measure of the quality degradation. Negative values indicate the amount of quality degradation induced by data hiding.

The embedding methods tested herein were the PLSB, GTPLSB, and PGLSB methods. The embedding bit rate was set to two levels, which were below 100 bps and 200 bps. The embedding bit rate is able to be roughly controlled by selecting appropriate values of the embedding parameters, that is, the number of embedding subframes, the width of the LSB, and the pitch gain threshold. The parameter values are shown in Table 2. Since activating the VAD and DTX functions in the encoder resulted in no data being embedded in the frames of the non-speech signal, the embedding bit rate depended somewhat on the length of the silence intervals in each speech sound file.

	Parameters			Embedding
Method	LSB	Subframes	Threshold	bit rate [bps]
PLSB	1	2, 4	—	max. 100
	1	1,2,3,4	—	max. 200
GTPLSB	2	1,2,3,4	3	50 — 120
	3	1,2,3,4	4	100 — 220
PGLSB	1	2, 4	—	max. 100
	1	1,2,3,4	—	max. 200

Table 2: Simulation parameters.

4.2. Results

The results of the computer simulation are shown as a two-dimensional map, where the abscissa and the ordinate denote the rate of embedding bit and the difference of MOS-LQO, respectively. Each dot in the figure represents a speech signal used for testing.

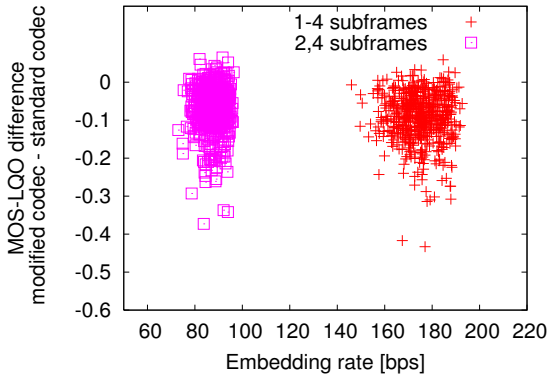

Fig. 3: Quality degradation induced by data hiding versus embedding data bit rate. Integrated PLSB was employed for the 12.2-kbps AMR mode.

Figures 3, 4, and 5 show the results of integrated embedding implementation for PLSB, GTPLSB, and PGLSB, respectively, at a speech data bit rate of 12.2 kbps. Figure 6 shows the result of integrated PLSB at the 7.95-kbps mode.

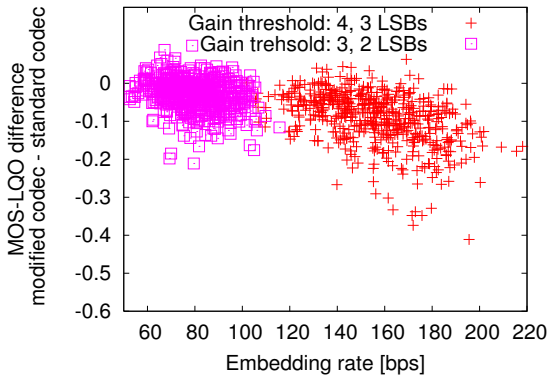

Fig. 4: Quality degradation induced by data hiding versus embedding data bit rate. Integrated GTPLSB was employed for the 12.2-kbps AMR mode.

Figures 7, 8, and 9 show the results of separated embedding implementation for PLSB, GTPLSB, and PGLSB, respectively, at a speech data bit rate of 12.2 kbps. Compared with the integrated implementation, the amount of sound quality degradation was clearly increased for the separated PLSB method. Comparison between integrated and separated implementation was also conducted for GTPLSB and PGLSB in the 12.2-kbps mode.

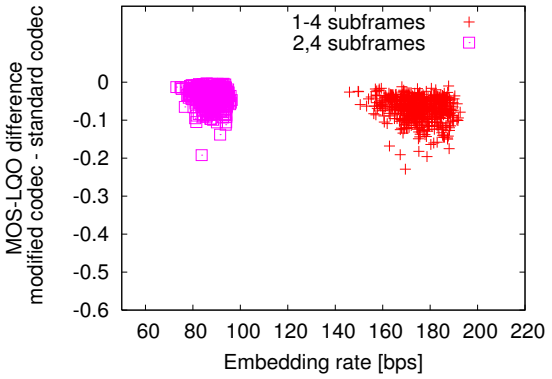

Fig. 5: Quality degradation induced by data hiding versus embedding data bit rate. Integrated PGLSB was employed for the 12.2-kbps AMR mode.

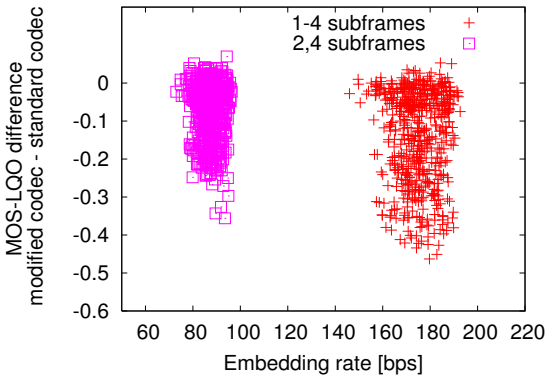

Fig. 6: Quality degradation induced by data hiding versus embedding data bit rate. Integrated PLSB was employed for the 7.95-kbps AMR mode.

There was no significant difference observed between integrated implementation and separated implementation. These results show that the integrated implementation is advantageous only for the PLSB method.

The range of quality degradation of PGLSB is limited and small. The corresponding t-test also shows that the mean MOS-LQO difference was the smallest for PGLSB, as compared to the other methods of separated implementation, in both data bit rate modes and higher embedding bit rate conditions. In addition, the integrated PGLSB at the 12.2-kbps mode showed the smallest MOS-LQO difference among the other methods of the integrated implementation. In order to express the av-

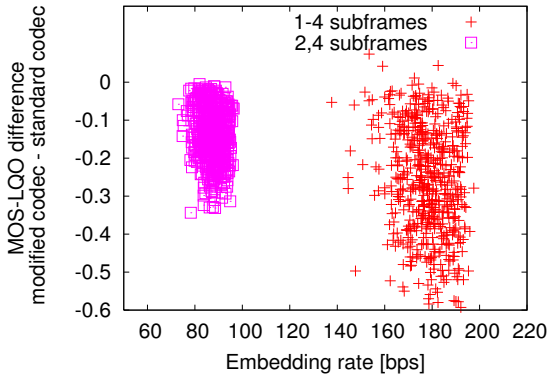

Fig. 7: Quality degradation induced by data hiding versus embedding data bit rate. Separated PLSB was employed for the 12.2-kbps AMR mode.

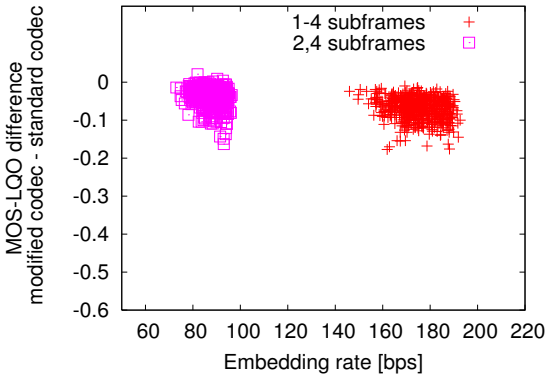

Fig. 9: Quality degradation induced by data hiding versus embedding data bit rate. Separated PGLSB was employed for the 12.2-kbps AMR mode.

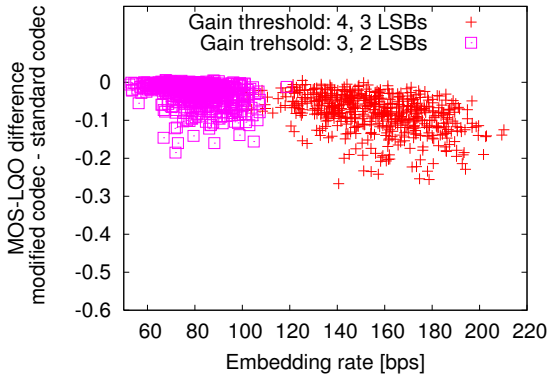

Fig. 8: Quality degradation induced by data hiding versus embedding data bit rate. Separated GTPLSB was employed for the 12.2-kbps AMR mode.

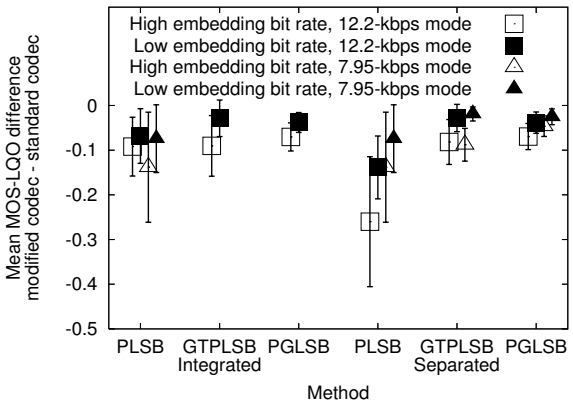

Fig. 10: Mean MOS-LQO difference for all conditions. Error bars denote $\pm$ 1 standard deviation.

erage differences among all conditions at a glance, Fig. 10 shows the mean MOS-LQO difference and $\pm$ 1 standard deviation for all conditions.

The range of quality degradation of GTPLSB is comparable to that of PGLSB. However, the range of embedding bit rate is diverse and presents a disadvantage for practical data hiding applications.

In summary, PGLSB yields superior results in both the separated and integrated implementations.

5. DISCUSSION

The embedding bit rates for PGLSB and PLSB depend on the duration of silence or non-speech intervals in the speech signal. If the noisy condition is simulated, the embedding bit rate will increase slightly for both PLSB and PGLSB. The embedding bit rates of GTPLSB in the noisy condition will clearly increase, because the ratio of periodic components in the speech waveform decreases in the noisy condition. Informal simulation revealed that embedding bit rate increases from 10% to 20% depending on the SNR.

The integrated implementation is advantageous only for the PLSB method. The reason why the integrated implementation is not effective for the other methods is as follows: The errors induced by PLSB embedding are reduced by optimizing the three parameters in the encoder,

the pitch gain, the fixed codebook index, and the fixed codebook gain. On the other hand, the number of the optimized parameters are two, the fixed codebook index and the fixed codebook gain, for GTPLSB and PGLSB. Balancing optimization between the pitch gain and the codebook-related parameters may be effective to reduce errors in the pitch delay data. Another reason is localization of the pitch delay errors of GTPLSB. The GTPLSB method replaces LSBs of the pitch delay data where the pitch gain is small, that is, where non-periodic speech signal is observed. The errors of the pitch delay data in such region do not affect the perceptual quality of the speech signal.

The present study dealt with objective evaluation of the three data hiding methods for the AMR narrow-band speech codec. Subjective evaluation is also useful for confirming the present results, whereas a great deal of effort is required for subjective experiments. Most subjective evaluations for data hiding in speech codecs conducted in previous studies used the absolute category rating (ACR) method, in which the listeners performed evaluation using absolute categories of excellent, good, fair, poor, and bad, which corresponds to nominal values of five to one. The ACR method is not suitable for discovering subtle sound quality degradations. A general method for measuring perceptual transparency is the double-blinded AXB discrimination test. Giving that the perceptual difference is clear between the standard codec and the modified codec, however, it does not mean that the sound quality of the modified codec is inferior to that of the standard codec. An adequate method to rate the sound quality of the modified codec compared with that of the standard codec is pair or multiple degradation comparison test, such as MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) method, as specified in ITU-R Rec. BS.1534-1.

The simple algorithms of the GTPLSB and PGLSB methods are limited to the AMR 12.2-kbps and 7.95-kbps modes. PGLSB is difficult to extend to other data bit rate modes in the present form because the pitch gain parameter is not separately quantized in the output bit stream of the encoder. The advantages of the LSB based data hiding method are simplicity and a computationally light load. Extension and improvement of the LSB methods for other bit rate modes, while maintaining the advantages of these methods, should be examined in the future.

6. SUMMARY

Three methods for data hiding in pitch-related parameters of the AMR narrow-band speech codec were evaluated in terms of the objective quality degradation and bit rate of embedding data. Computer simulation of the data hiding system revealed that the method of replacing the LSB of the pitch gain parameter in information bits was far superior to the other methods, which use the LSBs of the pitch delay data. The present method and simulation were conducted for the AMR 12.2-kbps and 7.95-kbps modes. Extension to other bit rate modes should be examined in the future.

Acknowledgments

The present research was supported in part by the Collaboration Research Program No. 4 of Tokyo University of Information Sciences 2007, 2008 and by KAKENHI, 20560365.

7. REFERENCES

[1] 3rd Generation Partnership Project, "Mandatory Speech Codec speech processing functions AMR Speech Codec; General Description," **26.071**, (2001).

[2] ITU-T Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," **P.862**, (2001).

[3] 3rd Generation Partnership Project, "Mandatory Speech Codec speech processing functions AMR Speech Codec; Transcoding Functions," **26.090**, (2001).

[4] HATADA Mitsuhiro, SAKAI Toshiyuki, KOMATSU Naohisa, and YAMAZAKI Yasushi, "A Study on Digital Watermarking Based on Process of Speech Production," IPSJ CSEC SIG Notes, **2002**, No. 43, 37—42 (2002).

[5] Munetoshi Iwakiri and Kineo Matsui, "Embedding a Text into Conjugate Structure Algebraic Code Excited Linear Prediction Audio Codecs," Journal of IPSJ, **39**, No. 9, 2623–2630 (1998).

[6] B. Geiser and P. Vary, "Backwards Compatible Wideband Telephony in Mobile Networks: CELP

Watermarking and Bandwidth Extension," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. IV, 533—536, (2007).

[7] Shigeru Sasaki, Masakiyo Tanaka, Yoshiteru Tsuchinaga, Masanao Suzuki, and Yasuji Ota, "Method and system for embedding and extracting data from encoded voice code," United States Patent 7310596 (2007).

[8] 3rd Generation Partnership Project, "ANSI-C code for the Adaptive Multi Rate speech codec," **26.073**, (2001).