Steganographic Band Width Extension for the AMR Codec of Low-Bit-Rate Modes

Akira Nishimura

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

akira@rsch.tuis.ac.jp

Abstract
This paper proposes a bandwidth extension (BWE) method for the AMR narrow-band speech codec using steganography, which is called steganographic BWE herein. The high-band information is embedded into the pitch delay data of the AMR codec using an extended quantization-based method that achieves increased embedding capacity and higher perceived sound quality than the previous steganographic method. The target bit-rate mode is below 7 kbps, the level below which the previous steganographic BWE method did not maintain adequate sound quality. The sound quality of the steganographic BWE speech signals decoded from the embedded bitstream is degradation in the quality relative to speech signals decoded from the same bitstream by the legacy AMR decoder.

Index Terms: speech codec, telecommunication, objective quality evaluation, PESQ, data hiding, steganography

1. Introduction
Bandwidth extension (BWE) for narrow-band speech signals, typically 8-kHz sampled digital speech signals transmitted over a telephone network, is considered to be an effective service for the current generation of telecommunication devices. The simplest approach for communication using wide-band speech signals is adoption of a wide-band speech codec. Although the latest wide-band speech codec, the G.711-WB code, is interoperable with the legacy narrow-band G.711 codec through a small modification to the transcoder in the network, most wide-band codecs (G.722, AMR-WB, SILK, etc.) have no backward compatibility with the legacy narrow-band codecs. Therefore, a bitstream generated by the wide-band encoder on the transmitter side cannot be directly fed into the narrow-band decoder on the receiver side.

A number of BWE algorithms that maintain backward compatibility with the legacy codecs have been developed to extend the upper frequency range of the narrow-band speech codec, typically from 3.4 kHz to 7.5 kHz. In other words, a bitstream generated by the encoder is decoded by the BWE decoder into the wide-band speech signal, while the identical bitstream can be decoded by the legacy decoder into the narrow-band speech signal.

Conventional approaches of bandwidth extension artificially generate high-band signals using the statistical characteristics and extracted features of the low-band signals on the receiver side [1, 2, 3]. This does not require any modifications to the legacy narrow-band encoder. However, such approaches are limited in terms of accurately representing the accurate high-band sounds emitted from various types of speakers in noisy environments.

Another approach extracts the high-band information on the transmitter side and transmits it together with the narrow-band speech data. The high-band information corresponds to the spectral and temporal information of the speech signal in the upper frequency region. On the receiver side, the high-band signal is synthesized and added to the decoded narrow-band signal.

One such approach, steganographic BWE, in which out-of-band information is embedded into the narrow-band speech data, has been introduced by several researchers [4, 5]. A bitstream including embedded high-band information is intended to have little effect on the sound quality decoded by the legacy narrow-band decoder. Vary and Geiser have investigated steganographic wide-band telephony for several legacy narrow-band speech codecs, namely, G.711, G.726, G.729, GSM FR, and GSM EFR [6]. Their method is based on a simplified version of the Time-Domain Band Width Extension module from the ITU-T G.729.1 codec. At the encoder, 5 sub-band gains representing the spectral envelope of the high-band signal are quantized and embedded into each 20-ms signal data frame. Their study indicated that the upper-band information can be embedded at a bit rate of 600 bps into the legacy narrow-band speech codecs with little objective sound quality degradation. The wide-band signals decoded by the steganographic BWE decoder have sound quality comparable with the legacy wide-band speech signals (AMR-WB and G.722). However, embedding side information at 600 bps into the speech data at a total bit rate of less than 8-kbps undoubtedly affects sound quality.

In the present study, a steganographic BWE method for the low-bit-rate modes of the AMR narrow-band speech codec is investigated. The previous method of embedding hidden data into the pitch delay data is extended to increase embedding capacity and improve perceived sound quality. Side-information at 220 bps in the high-band signal is utilized to synthesize speech signals above 3.9 kHz. The source of the high-band signal is automatically selected as either the output of the full-wave rectifier of the middle-band signal or the noise signal depending on the high-band level. Objective evaluation of sound quality for the decoded narrow-band signals and the BWE signals was conducted by using the Perceptual Evaluation for Speech Quality (PESQ) program defined by ITU-T P.862.1 and P.862.2.

2. Data hiding in AMR speech codec
2.1. AMR codec
A large number of Third-Generation Partnership Project (3GPP)-based cellular phones use the AMR speech codec,
which is one of the latest derivatives of the code-excited linear prediction (CELP)-based speech codecs. The AMR speech encoder converts 20 ms of an 8-kHz and 13-bit digital waveform frame into line spectral pair (LSP) parameters, pitch parameters, an algebraic code index, and gain parameters. These parameters are transmitted using a selective bit rate mode from 4.75 to 12.2 kbps [7].

2.2. Data hiding methods

Several methods have been proposed for embedding hidden data into the CELP-based speech parameters. Embedding data into the algebraic codebook index by selecting a labeled codebook table is effective in high-bit-rate modes [8, 9], because several bit allocations in the codebook table make the fixed pulse positions redundant. However, sound quality may degrade in the low-bit-rate modes of the AMR codec, since the number of bit allocations for the algebraic codebook is quite small, for example, 14 bits per subframe in the AMR 6.7-kbps mode.

Iwakiri (2002) proposed a method of hiding data by quantization of the pitch delay parameter in the ITU-T G.723.1 speech codec. Embedding bit information is accomplished by quantizing the pitch delay of the subframe signal, where the lowest n-bit of the pitch delay data corresponds to the hidden bit information. This is implemented by restricting the search range of the pitch delay value in 2^n steps. The benefit of this method is that the optimum search algorithm of the algebraic codebook table compensates for the quality degradation induced by errors in pitch delay data to some extent.

The present study extends this embedding method based on quantization of pitch delay data. The sensitivity of the human auditory system with respect to pitch difference is known to be roughly proportional to the ratio of the frequency difference. In other words, a human listener has difficulty detecting a small pitch delay difference relative to a large pitch delay value. Therefore, if the pitch delay value of the odd subframe exceeds one-half of the pitch delay range, that is, 256 for the 12.2-kbps mode and 128 for other modes, the embedding bit width of quantization is extended by an additional 1 bit for the next even and odd subframes. In addition, if the pitch delay value of the odd subframe is less than 32 with the exception of the 12.2-kbps mode, no data is embedded in the next even subframe. As a result, the number of embedded bits varies from 2 bits to 8 bits per 20-ms voiced frame. The embedding bit rate depends on the amount of silent intervals and the pitch delay value of the speech signal. The average embedding bit rate is approximately 220 bps.

3. BWE algorithm

Vary and Geiser utilized noise excitation as the source of the high-band signals at the decoder [6]. This was almost successful based on the evaluation by the PESQ program. However, the quality degradation is relatively conspicuous for speech signals of high pitch since higher harmonics are actually present in the high-frequency region. In addition, embedding side-information at 600 bps into the encoded speech data at a total bit rate of less than 8 kbps seriously affects sound quality. For example, 600-bps embedding degrades the mean opinion score listening quality objective (MOS-LQO) value of the 8-kbps G.729-encoded speech signals by approximately 0.17 [6].

The present study adopts a simpler representation of the spectral envelope as well as the selection of harmonic signals and noise-like signals as the high-band signal. The gains of the two sub-band signals in the upper frequency region are quantized and embedded by the steganographic AMR encoder. Harmonic high-band signals, which are generated by full-wave rectification of the middle-band signals [1, 10], and noise excitation become sources of the synthesized high-band signal according to the gain parameter value.

3.1. Encoder

A block diagram of encoding and decoding is shown in Fig. 1. Since the power of the upper frequency band is correlated to that of the middle frequency band (MB: 1.6 kHz to 3.8 kHz), the root-mean-square (RMS) levels of the high frequency sub-bands (first high band (HB1): 3.9 kHz to 5.0 kHz; second high band (HB2): 5.0 kHz to 7.0 kHz) relative to the level of the MB are quantized as a gain parameter. The gains are calculated once per 20-ms data frame. The quantization step sizes of HB1 and HB2 are 3 dB and 4 dB, respectively. The quantization function \( Q(x) \) is shown in Eq. 1, where \( x \) is the high-band gain, \( C \) is a constant to offset the gain, and \( D \) is the quantization step size. \( Q(x) \) ranges from 0 to 15 and is converted to a 4-bit binary. The values of \( C \) for HB1 and HB2 are 10 dB and 14 dB, respectively.

\[
Q(x) = \left( \frac{x + C}{D} \right) + 7
\]

The two 4-bit binary data of the quantized gain parameter are partly interleaved before embedding. Table 1 shows the bit interleave table where \( d1 \) denotes the bit of the HB1 gain, and \( d2 \) denotes the HB2 gain. The suffix numbers from 1 to 4 denote the location relative to the highest bit. The bit information to be embedded is selected from the left side of Table 1 in 2 to 8 bits according to the capacity of hidden data, which is determined by the pitch delay value.

Table 1: Bit interleave table. \( d1 \) denotes a bit of HB1 gain and \( d2 \) denotes that of HB2 gain. Suffix numbers denote locations relative to the highest bit.

| \( d1_1 \) | \( d1_2 \) | \( d2_1 \) | \( d2_2 \) | \( d1_3 \) | \( d2_3 \) | \( d1_4 \) | \( d2_4 \) |

3.2. Decoder

The AMR decoder decodes a narrow-band speech signal from the steganographic bitstream data. The output of the AMR decoder is filtered by the band-pass filter whose frequency range corresponds to that of the MB. The filtered waveform is full-wave rectified to generate high-band signals of the current data frame. Then, it is filtered by the band-pass filter whose frequency range corresponds to that of HB1. A noise signal is also filtered by the band-pass filters whose frequency range corresponds to those of HB1 and HB2. The HB2 noise is the source of the synthesized HB2 signal.

The dequantization functions of the gain parameter \( DQ(y) \) are defined for each number of the extracted \( bn \)-bit in Eq. 2, where \( y \) is a decimal gain converted from a \( bn \)-bit binary. Eq. 2 is designed to minimize quantization errors in the dequantization process from the available \( bn \)-bit gain data. When no bit information is available for HB2, the gain of HB1 is utilized instead. The dequantized gain values are used in conjunction with the RMS levels of HB1, HB2, and the MB at the decoder to restore the relative levels of HB1 and HB2 at the encoder.
4. Objective quality measurements of embedded and BWE speech signals

The PESQ program distributed by ITU-T was used to evaluate the quality degradation of the narrow-band speech signals decoded by the legacy AMR decoder from the embedded bitstream and the wide-band speech signals decoded by the steganographic BWE decoder from the same embedded bitstream. The legacy wide-band speech codec, AMR-WB, was also tested for comparison.

PESQ is a perception-based method of objective sound quality evaluation for speech codecs. Using a psychophysical representation of audio signals, it compares the original signal with a signal that has been degraded by being passed through a communication system. The transformed output of PESQ (ITU-T P.862.1) is MOS-LQO and corresponds to the results of the mean opinion score listening quality subjective (MOS-LQS) obtained from human listeners through subjective experiments. In addition, P.862.2 defines MOS-LQO for wide-band (16-kHz sampling) speech signals.

A total of 550 phonetically balanced sentences spoken by 22 Japanese speakers (12 men and 10 women) of 16-bit quantization and 16-kHz sampling were input into the BWE encoder. These sentences were generated by concatenating two sentences from among 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 s, including intervals of silence. The overall level of each input speech signal was –26 dB. The output bitstream of the encoded speech data was fed into the standard AMR decoder and the BWE decoder. In addition, the speech signals of the database were downsampled to 8 kHz and fed into the standard AMR encoder and decoder.

Figure 3 shows the average MOS-LQO values of the decoded narrow-band speech signals for several typical bit-rate modes of the AMR codec. The quality degradation of the steganographic speech data decoded by the legacy AMR decoder is less than 0.1 for all bit-rate modes.
not only clean speech signals but also speech signals with background noises, for example, bubble noises or environmental noises. The present BWE algorithm has not been tested quantitatively under such conditions. Furthermore, subjective evaluation of sound quality for the steganographic narrow-band signals and the BWE signals should be used to verify the objective evaluation of the present BWE system.

6. Summary

The proposed steganographic BWE system embeds high-band information into the pitch delay data of the AMR codec using an extended quantization-based method. The quality degradation of the narrow-band signals decoded by the legacy AMR decoder from the steganographic bitstream is relatively small. The sound quality of the steganographic BWE speech signals is comparable to that of the AMR-WB codec for bit-rate modes below 6.7 kbps. The proposed BWE method is backward compatible with the legacy AMR codec and low-bit-rate modes of wide-band telephony, while maintaining a reasonable level of quality.

7. Acknowledgments

The present study was supported in part by a Grant-in-Aid for Scientific Research (C) No. 20560365 (2009).

8. References


