INFORMATION HIDING FOR G.711 SPEECH BASED ON SUBSTITUTION OF LEAST SIGNIFICANT BITS AND ESTIMATION OF TOLERABLE DISTORTION

Akinori Ito

Graduate School of Engineering
Tohoku University
6-6-05 Aramaki aza Aoba, Aoba-ku,
Sendai 980-8579, Japan

Shun'ichiro Abe, Yôiti Suzuki

Research Institute for Electrical Communication
Tohoku University
2-1-1 Katahira, Aoba-ku,
Sendai 980-8577, Japan

ABSTRACT

In this paper, we propose a novel data hiding technique for G.711 speech based on the LSB substitution method. The novel feature of the proposed method is that a low-bitrate encoder, G.726 ADPCM, is used as a reference for deciding how many bits can be embedded in a sample. Experiments showed that the method outperformed the simple LSB substitution method and the selective embedding method proposed by Aoki. We achieved 4-kbit/s embedding with almost no subjective degradation of speech quality, and 10 kbit/s while keeping good quality.

Index Terms— Speech coding, Data hiding, G.711, ADPCM, LSB substitution

1. INTRODUCTION

Recently, VoIP (Voice over IP) technology has been extensively used as a new infrastructure of the public phone network. Although several codecs are available for VoIP, G.711 [1], the simplest codec, is the most common one at present and is expected to remain so for the immediate future.

There have been several attempts to enhance G.711-coded speech such as packet loss concealment [2] and bandwidth expansion [3, 4]. These enhancement methods require side information to be added to the speech data coded by the G.711 codec. When adding side information, it is desirable that the speech data containing the additional data for enhancement is downward compatible with ordinary speech data coded by G.711.

An approach based on data hiding solves this problem. Data hiding is a technique to embed certain additional data into the original media data (or host signal) such as image and speech, without significantly degrading the quality of the original media data. Although the techniques used in this approach are similar to those for steganography and watermarking, there is a significant difference. In the data hiding, the embedded information is used to realize certain addional value and is not necessarily kept secret. From the same reason, there is no need to consider any attack against the data hiding. In the present study, the embedded information is mere side information for enhancing the host signal and just hidden for compatibility purpose. Thus secrecy is not required for the embedded information.

Aoki proposed a data-hiding-based speech enhancement approach for the G.711 codec [5], based on LSB substitution for packet loss concealment and bandwidth expansion.

In this paper, we propose a novel data hiding method for G.711-coded speech. Our approach is similar to Aoki's method but with superior performance: our method can embed more data into the speech with less degradation.

2. LSB SUBSTITUTION FOR G.711-CODED SPEECH

2.1. Information hiding based on LSB substitution

In this section, we briefly explain the conventional data hiding method based on LSB substitution, and its improvement by Aoki [5]. The LSB substitution [6] is the simplest method for embedding information into speech data.

Let $x_1, \ldots, x_n$ be a sequence of scalar values, where x_i is encoded into M bits. Let $b_1, \ldots, b_n$ be the data we want to embed, where $b_i \in \{0, 1\}$. Let $m(x)$ is a bitmask function, which clears the least significant bit (LSB) of x. If M is sufficiently large, we can expect that there is no big difference between x and $m(x)$. Therefore, we can use the LSB of x_i as the payload for embedding information. Now let y_i be

$$y_i = m(x_i) + b_i. \tag{1}$$

Then we can use the data $y_1, \ldots, y_n$ as new data where the information $b_1, \ldots, b_n$ is embedded. The embedded information b_i can be recovered by

$$b_i = y_i - m(y_i). \tag{2}$$

One problem of applying the simple LSB substitution method to G.711-coded speech is that distortion of the decoded speech depends on the magnitude of the input speech. When linear quantization is applied, it is ensured

that $0 \leq (x_i - y_i)^2 \leq 1$ for any x_i. On the other hand, a nonlinear quantization method such as G.711 requires a nonlinear transformation for restoring the speech. Let o_i be a sample of the input speech, and $f(o)$ be a nonlinear function. Then, when no information is embedded, encoding and decoding of speech is like this:

$$x_i = f(o_i) \tag{3}$$

$$\hat{o}_i = g(x_i) \tag{4}$$

In G.711, two kinds of nonlinear functions are applied: the μ-law and the A-law. Although we apply the μ-law in this paper, the proposed framework can be also applied to the A-law. Now, the nonlinear function of the μ-law is

$$f(o) = \left\lceil 128 \frac{\ln\left(1 + \frac{255|o|}{8192}\right)}{\ln 256} \right\rceil \mathrm{sgn}(o) \tag{5}$$

$$g(x) = \mathrm{sgn}(x) \left\lceil 8192 \frac{256^{\frac{|x|}{128}} - 1}{255} \right\rceil \tag{6}$$

where $-8192 \leq o < 8192$.

The distortion of a sample by this kind of nonlinear quantization by embedding information is calculated as

$$\varepsilon_i^2 = (g(x_i) - g(y_i))^2, \tag{7}$$

which depends on the magnitude of x_i. When x_i is large, we obtain larger ε_i^2. Therefore, when the magnitude of the signal is large, the distortion audibly degrades the sound quality of the embedded speech signal.

2.2. Selective embedding for G.711

To improve the quality of embedded speech, Aoki invented a selective embedding method[5], which embeds a fixed number of information for samples that have small magnitudes.

Let $i_1, i_2, \ldots, i_n$ be a permutation of $1, 2, \ldots, n$ such that

$$|m(x_{i_1})| \leq |m(x_{i_2})| \leq \cdots \leq |m(x_{i_n})|. \tag{8}$$

Here, $i_j < i_{j+1}$ when $m(x_{i_j}) = m(x_{i_{j+1}})$.

On embedding information, we first define a number $k \leq n$. The key feature of this method is to embed only k bits of information into samples that have small magnitude. We calculate the embedded signal y_i such that

$$y_{i_j} = \begin{cases} m(x_{i_j}) + b_j & \text{if} \quad j \leq k \\ x_{i_j} & \text{otherwise} \end{cases} \tag{9}$$

By using samples with smaller magnitude for embedding, this method reduces the total distortion of the embedded signal. Moreover, the bitrates of the embedded data can be controlled up to 8 kbit/s (i.e. the bitrate of LSB) by changing k in balance with the speech quality.

3. ENHANCEMENT OF LSB SUBSTITUTION BASED ON ESTIMATION OF TOLERABLE DISTORTION

Although Aoki's method can reduce total distortion of the embedded signal, it has two problems. First, reducing distortion does not necessarily improve subjective quality. We could thus further improve the subjective quality of the embedded signal by considering some kind of subjective measure. Second, Aoki's method (and the original LSB substitution) only uses the least significant one bit for embedding. We could use more than one bit, i.e. multiple lower significant bits including LSB, if subjective degradation is small. Although LSB substitution for multiple bits was proposed for hiding data in images [7], such a technique for audio signals has not yet been proposed.

We therefore propose a novel data-hiding algorithm that is based on estimation of tolerable distortion. Here, "tolerable" means that the degradation of the signal is assured to be better than some limit of subjective quality.

To judge the limit of subjective quality, we propose employing another encoder as a reference, which encodes the input speech at less than 64 kbit/s by taking human hearing characteristics into account. Here, we assume that the low-bitrate encoder encodes the input speech sample-by-sample, considering the past input. Therefore, a frame-based encoder like CELP cannot be applied in our framework.

Let $m_j^-(x)$ be a function that clears the least j bits of x, and $m_j^+(x)$ be a function that sets the least j bits of x. Next, let $I(o_i|o_1, \ldots, o_{i-1})$ be an output of the i-th sample using a low-bitrate encoder, given the past input $o_1, \ldots, o_{i-1}$. Generally speaking, an existing low-bit encoder is designed to maximize subjective quality. Therefore, if we determine the number of bits for embedding considering the output of a low-bitrate encoder, we can ensure that the quality of the embedded speech is as good as that by the low-bitrate encoder. According to this idea, we determine the number of embeddable bits e_i as follows. First, we define $I_0(i, j)$ and $I_1(i, j)$ as follows.

$$\hat{o}_k = g(m_{e_k}^-(x_k)) \tag{10}$$

$$I_0(i, j) = I(g(m_j^-(x_i))|\hat{o}_1, \ldots, \hat{o}_{i-1}) \tag{11}$$

$$I_1(i, j) = I(g(m_j^+(x_i))|\hat{o}_1, \ldots, \hat{o}_{i-1}) \tag{12}$$

Then we define e_i as

$$e_i = \max_j \{j | I_0(i, j) = I_1(i, j)\}. \tag{13}$$

Here, e_i means the maximum number of bits where $I_0(i, ei)$ and $I_1(i, e_i)$ give the same result. Provided we use no more bits than e_i, we can ensure that the embedded information does not affect the result of the low-bitrate encoder. Note that Eqs. (11) and (12) contains $e_1, \ldots, e_{i-1}$. This means that we have to determine $e_1, \ldots, e_{i-1}$ and e_i in order. On extracting the embedded information from the signal, Eq. (13) can be

```
for i ← 1 to n
  for j ← 1 to 8
    e_i ← j − 1
    I_0(i,j) ← I(g(m_j^−(x_i))|ô_1, ..., ô_{i−1})
    I_1(i,j) ← I(g(m_j^+(x_i))|ô_1, ..., ô_{i−1})
    if I_0(i,j) ≠ I_1(i,j) then
      break
    end if
  end for
  Embed data into the least e_i bits of x_i
  ô_i ← g(m_{e_i}^−(x_i))
end for
```

Fig. 1. Procedure for embedding

used for determining the embedded bits, because e_i is determined using the past signal whose embedded bits are cleared, meaning that the determination of e_i is not affected by the embedded content. Figure 1 shows a procedure for determining e_i and embedding data into x_i.

In the present proposal, we chose G.726 ADPCM [8] as a low-bitrate encoder on implementing this algorithm, since G.726 has several features that are useful in our framework. First, it determines the code sample-by-sample. Next, G.726 has several bitrate configurations, which allow us to control the amount of embedded information.

In addition, we make an exceptional rule for determining the number of embeddable bits. Let N be the bit length of a sample coded by ADPCM. The number depends on the bitrate of ADPCM; $N = 5$ for the 40 kbit/s configuration, $N = 4$ for 32 kbit/s and $N = 3$ for 24 kbit/s. When $|I_0(i,0)| = 2^{N-1}-1$, it means that the distortion between the true sample and the predicted value is maximum. In this case, $I_0(i,j)$ and $I_1(i,j)$ coincide with high probability, whose absolute value is $2^{N-1}-1$. This happens because the absolute value of the code cannot become any larger. In this case, we cannot trust the fact that $I_0(i,j) = I_1(i,j)$ as an index of subjective quality, so we use the following e'_i instead of e_i, which is defined as

$$e'_i = \begin{cases} 0 & \text{if } |I_0(i,0)| = 2^{N-1}-1 \\ e_i & \text{otherwise} \end{cases} \quad (14)$$

4. EXPERIMENT

4.1. Experimental conditions

We conducted experiments to compare the bitrate of embedded information and the degradation of the original signal by the proposed method and the conventional methods. We used 10 utterances extracted from the NTT-AT phone-balanced speech database for testing.

Table 1. Experimental results using 32 kbit/s ADPCM

Method	Bitrate [bit/s]	MOS-LQO (st.dev.)
LSB substitution	8000	4.05 (0.07)
Aoki's method	2000	4.36 (0.04)
	4000	4.32 (0.04)
	5500	4.14 (0.05)
	6000	4.10 (0.06)
Proposed	5585	4.40 (0.05)

Table 2. Ratio of embeddable bits

e'_i	0	1	2	3	≥ 4
Ratio [%]	51.6	30.0	13.1	4.1	1.3

As an index of subjective quality of embedded speech, we used MOS-LQO (Mean Opinion Score-Listening Quality Objective) [9], which is an objective index of speech quality that is compatible with the MOS value of a subjective listening test. The MOS-LQO value is calculated from ITU-T P.862 PESQ (Perceptual Evaluation of Speech Quality) [10]. We used OPTICOM OPERA for calculating PESQ.

4.2. Experimental results

First, we compared the proposed method with the simple LSB substitution and Aoki's method. As Aoki's method can control the bitrate of embedded information, we tested Aoki's method at several bitrates. When using the proposed method, we used the 32 kbit/s configuration for the ADPCM encoder and we embedded at most 3 bits in one sample.

Table 1 shows the result. Compared with Aoki's method, our method achieved higher MOS-LQO, proving its effectiveness.

Table 2 shows the ratio of embeddable bits for a sample. From this result, no information is embedded in more than half of the sample. On the other hand, we can embed more than two bits of information into almost 20% of the samples

Next, we used 24 kbit/s, 32 kbit/s and 40 kbit/s ADPCM configurations, changing the maximum number of bits for embedding. The experimental result is shown in Fig. 2. The numerals within the graph indicate the maximum number of bits for substitution. We could embed more than 10 kbit/s of information while the quality of the embedded speech remained above 4.0 MOS-LQO. In contrast, we could embed 3 to 4 kbit/s of information with a MOS-LQO value of more than 4.45. Considering that the maximum value of MOS-LQO is 4.5, the speech with 4 kbit/s of hidden data is considered to be almost indistinguishable from the original speech.

Instead of changing the maximum number of bits for substitution, we can use a different way of controlling the balance between bitrate and speech quality for a given ADPCM configuration. The tolerable distortion D_i is calculated as $D_i = (g(m^+_{e_i}(x_i)) - g(m^-_{e_i}(x_i)))^2$. Then we can calculate a

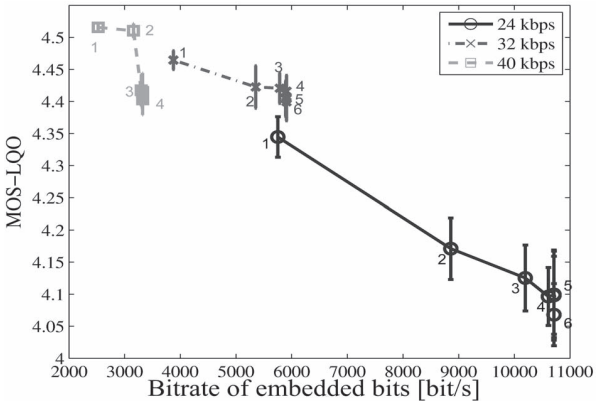

Fig. 2. Bitrate vs. MOS for various configurations

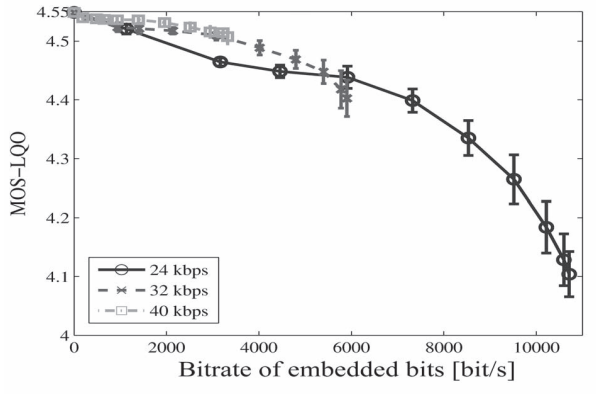

Fig. 3. Bitrate vs. MOS when using α.

restricted number of bits as follows:

$$\tilde{e}_i(\alpha) = \max_j \{j | g(m_j^+(x_i)) - g(m_j^-(x_i))^2 < \alpha^2 D_i\} \quad (15)$$

Here, α is a control parameter where $0 < \alpha \leq 1$. The experimental result is shown in Fig. 3. Compared with Fig. 2, the method that uses α for bitrate control showed better performance.

5. CONCLUSION

In this paper, we proposed a novel data hiding technique for G.711-coded speech based on enhancing the conventional LSB substitution, while employing a low-bitrate codec as a reference for speech quality for deciding if a certain number of bits can be embedded in a sample. Experiments showed we could embed about 4 kbit/s of information with almost no degradation, and more than 10 kbit/s of information while maintaining "good" quality.

One drawback of the proposed method is that the bitrate of the hidden information is not constant. As it is often desirable to ensure a fixed bitrate for embedding information in some fixed format, we are going to improve the proposed method for achieving constant bitrate embedding.

6. ACKNOWLEDGMENT

This study was partly supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) No. 051302004 of the Ministry of Internal Affairs and Communications of Japan.

7. REFERENCES

[1] International Telecommunication Union, "G.711 : Pulse code modulation (PCM) of voice frequencies," 1988.

[2] Noriko Komaki, Naofumi Aoki, and Tsuyoshi Yamamoto, "A packet loss concealment technique for VoIP using steganography," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E86-A, no. 8, pp. 2069–2072, 2003.

[3] Naofumi Aoki, "A band extension technique for G.711 speech using steganography," *IEICE Trans. Communications*, vol. E89-B, no. 6, pp. 1896–1898, 2006.

[4] Akitoshi Kataoka, Takeshi Mori, and Shinji Hayashi, "Bandwidth extension of G.711 using side information," *IEICE Trans. Inf&Syst.*, vol. J91-D, no. 4, pp. 1069–1081, 2008 (in Japanese).

[5] Naofumi Aoki, "Potential of value-added speech communications by using steganography," in *Proc. 3rd Int. Conf. Intelligent Information Hiding and Multimedia Signal Processing*, 2007, vol. 2, pp. 251–254.

[6] Stefan Latzenbeisser, *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House, 2000.

[7] Shao-Hui Liu, Tian-Hang Chen, Hong-Xun Yao, and Wen Gao, "A variable depth LSB data hiding technique in images," in *Proc. Int. Conf. on Machine Learning and Cybernetics*, 2004, vol. 7, pp. 3990–3994.

[8] International Telecommunication Union, "G.726 : 40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)," 1990.

[9] International Telecommunication Union, "Mapping function for transforming of P.862 to MOS-LQO," 2003.

[10] International Telecommunication Union, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.