5

Lossless Steganography for Speech Communications

Naofumi Aoki
Graduate School of Information Science and Technology, Hokkaido University
Japan

1. Introduction

Transmitting supplementary data by steganography, new functions can be added to communications systems without changing conventional data format. Based on this concept, several applications have been proposed for enhancing the speech quality of telephony communications. These applications secretly transmit side information along with speech data itself for enhancing the performance of signal processing such as packet loss concealment and band extension (Aoki, 2003; Aoki, 2006; Aoki, 2007a; Aoki, 2012).

The simplest steganography technique employed in such applications is the LSB (Least Significant Bit) replacement technique (Cox, 2008). It just replaces the LSB of speech data with secret message. Since the LSB of speech data is not very important in perception, the LSB replacement technique is sufficient enough in many practical cases.

However, there is no way to avoid inevitable degradation of speech data by embedding secret message with the LSB replacement technique. In order to mitigate this problem, this article describes an idea of the lossless steganography technique for telephony communications (Aoki, 2007b; Aoki, 2008; Aoki, 2009a; Aoki, 2009b; Aoki, 2010a). The proposed technique exploits the characteristic of the folded binary code employed in several speech codecs, such as G.711 and DVI-ADPCM.

2. LSB replacement technique

The LSB replacement technique is known as one of the simplest steganography technique (Cox, 2008). It just embeds secret message into the LSB of cover data. The embedding procedure of the LSB replacement technique is programmed in C language as shown in Fig. 1. In this procedure, b represents a 1 bit secret message and c represents an 8bit cover data. The LSB replacement technique is categorized as a lossy steganography technique, since it may degrade cover data by embedding secret message.

```
if (b == 0) c &= 0xFE;
if (b == 1) c |= 0x01;
```

Fig. 1. Embedding procedure of the LSB replacement technique programmed in C language.

3. Lossless steganography technique based on the folded binary code

For representing signed integers as binary data, many speech codecs employ the folded binary code instead of the 2's complement, the most common format of binary data. Table 1 shows how an 8 bit speech data is represented by the folded binary code as well as the 2's complement.

decimal number	2's complement	folded binary code
+127	0 1 1 1 1 1 1 1	0 1 1 1 1 1 1 1
~		
+3	0 0 0 0 0 0 1 1	0 0 0 0 0 0 1 1
+2	0 0 0 0 0 0 1 0	0 0 0 0 0 0 1 0
+1	0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 1
+0	0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0
-0	0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0
-1	1 1 1 1 1 1 1 1	1 0 0 0 0 0 0 1
-2	1 1 1 1 1 1 1 0	1 0 0 0 0 0 1 0
-3	1 1 1 1 1 1 0 1	1 0 0 0 0 0 1 1
~		
-127	1 0 0 0 0 0 0 1	1 1 1 1 1 1 1 1
-128	1 0 0 0 0 0 0 0	

Table 1. 2's complement and folded binary code for representing 8 bit speech data.

As shown in this table, an 8 bit speech data encoded in the 2's complement ranges from -128 to +127. On the other hand, an 8 bit speech data encoded in the folded binary code ranges from -127 to +127. Although the folded binary code cannot represent -128, it may represent both +0 and -0 instead. This redundancy can be a container for embedding secret message without any degradation.

The embedding procedure of the proposed technique is programmed in C language as shown in Fig. 2. In this procedure, b represents a 1 bit secret message and c represents an 8bit cover data encoded in the folded binary code. The proposed technique is categorized as a lossless steganography technique, since it does not degrade cover data by embedding secret message.

```
if ((c & 0x7F) == 0)
{
  if (b == 0) c &= 0x7F;
  if (b == 1) c |= 0x80;
}
```

Fig. 2. Embedding procedure of the proposed technique programmed in C language.

4. G.711

G.711 is the most common codec for telephony speech standardized by ITU-T (International Telecommunication Union Telecommunication Standardization Sector) (ITU-T, 1988). It consists of μ-law and A-law schemes designated as PCMU and PCMA, respectively. PCMU is mainly employed in North America and Japan. It encodes 14 bit speech data into 8 bit compression data at an 8 kHz sampling rate. PCMA is mainly employed in Europe. It encodes 13 bit speech data into 8 bit compression data at an 8 kHz sampling rate.

Figure 3 and 4 show the encoding and decoding procedure of PCMU. The compression data of PCMU consists of 1 bit sign, 3 bit exponent, and 4 bit mantissa (ITU-T, 2005). The compression data is encoded in the folded binary code. Table 2 shows some of the compression data and their corresponding speech data decoded with PCMU. As shown in this table, the speech data decoded with PCMU ranges from -0 to -8031, and +0 to +8031.

Figure 5 and 6 show the encoding and decoding procedure of PCMA. The compression data of PCMA consists of 1 bit sign, 3 bit exponent, and 4 bit mantissa (ITU-T, 2005). The compression data is encoded in the folded binary code. Table 2 shows some of the

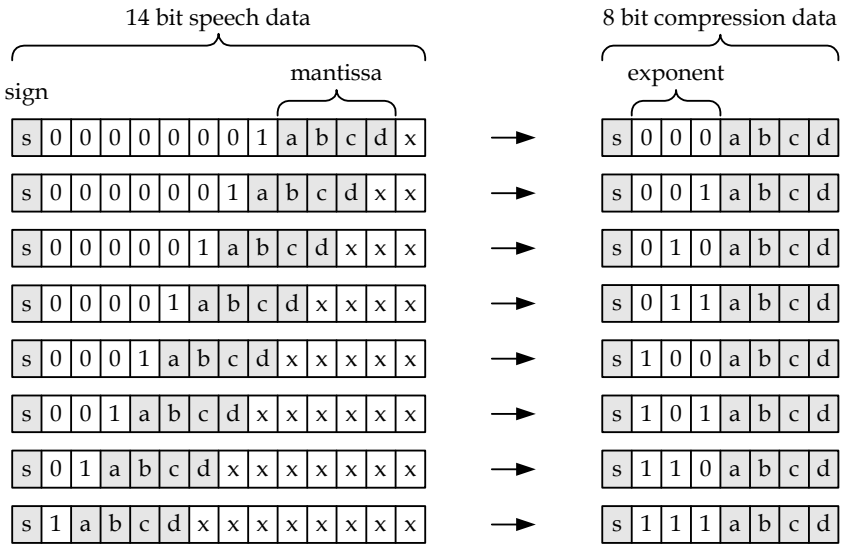

Fig. 3. Encoding procedure of PCMU.

compression data and their corresponding speech data decoded with PCMA. As shown in this table, the speech data decoded with PCMA ranges from -1 to -4032, and +1 to +4032.

Note that there is an overlap in the speech data decoded with PCMU. On the other hand, there is no such an overlap in the speech data decoded with PCMA. This indicates that a lossless steganography technique is available for PCMU, although it is not for PCMA.

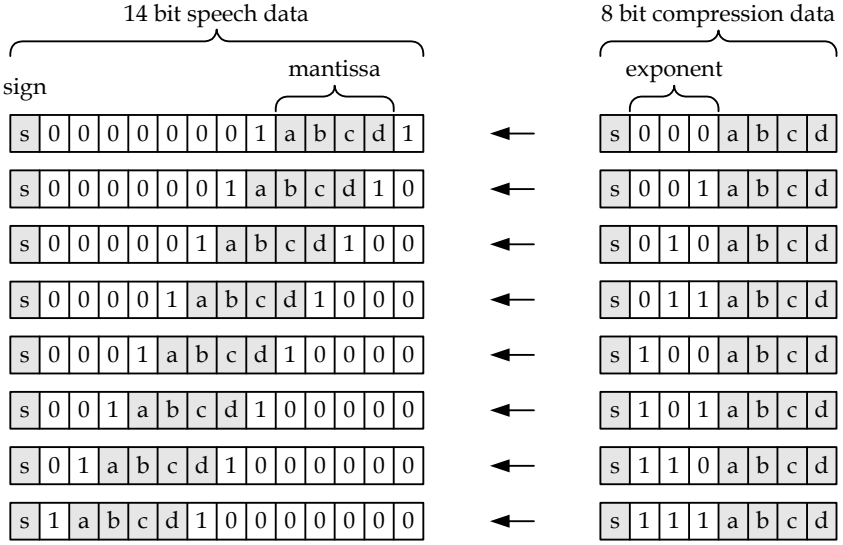

Fig. 4. Decoding procedure of PCMU.

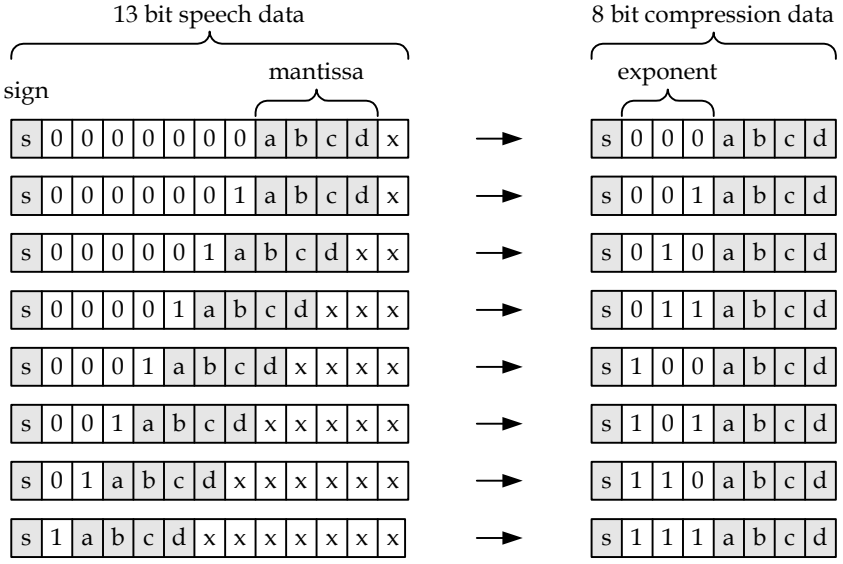

Fig. 5. Encoding procedure of PCMA.

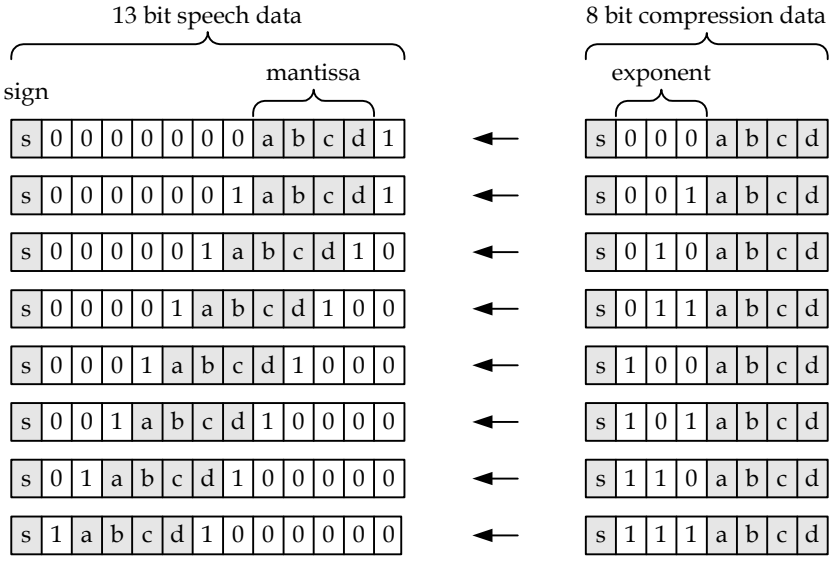

Fig. 6. Decoding procedure of PCMA.

compression data	decoded speech data (PCMU)	decoded speech data (PCMA)
+127	+8031	+4032
~		
+3	+6	+7
+2	+4	+5
+1	+2	+3
+0	+0	+1
-0	-0	-1
-1	-2	-3
-2	-4	-5
-3	-6	-7
~		
-127	-8031	-4032

Table 2. Speech data decoded with PCMU and PCMA.

5. Lossless steganography technique for G.711

Taking account of the characteristic of PCMU, secret message can be embedded into both +0 and -0 in the compression data without any degradation. When 0 is required to be embedded, the sign bit of the compression data is changed to be 0. This means that the compression data is changed to be +0. When 1 is required to be embedded, the sign bit of the compression data is changed to be 1. This means that the compression data is changed to be -0. The embedding procedure of the proposed technique is defined as follows.

$$c = \begin{cases} +0 & (|c| = 0, b = 0) \\ -0 & (|c| = 0, b = 1) \end{cases} \tag{1}$$

where b represents a 1 bit secret message and c represents an 8 bit compression data. This procedure is programmed in C language as shown in Fig. 7.

Figure 8 shows an example of the proposed technique. The compression data is represented as white and black circles according to the sign bit. The sign bit of the compression data represented by white circle is 0. On the other hand, the sign bit of the compression data represented by black circle is 1.

This example shows 4 candidates that can contain 4 bit secret message in total. According to their sign bits, these data originally contain 4 bit secret message represented as (0, 0, 1, 1). In order to embed secret message represented as (0, 1, 0, 1), the proposed technique changes these data as shown in Fig. 8. Since all of these are decoded to be 0 even if their sign bits are changed, the proposed technique does not degrade the speech quality at all.

```
if ((c & 0x7F) == 0)
{
  if (b == 0) c &= 0x7F;
  if (b == 1) c |= 0x80;
}
```

Fig. 7. Embedding procedure of the proposed technique programmed in C language.

6. DVI-ADPCM

The concept of the proposed technique may potentially be applicable to other codecs that also employ the folded binary code. Another example is DVI-ADPCM. Not only G.711 but also DVI-ADPCM is employed in telephony communications as a standard VoIP (Voice over IP) codec (RFC, 1996).

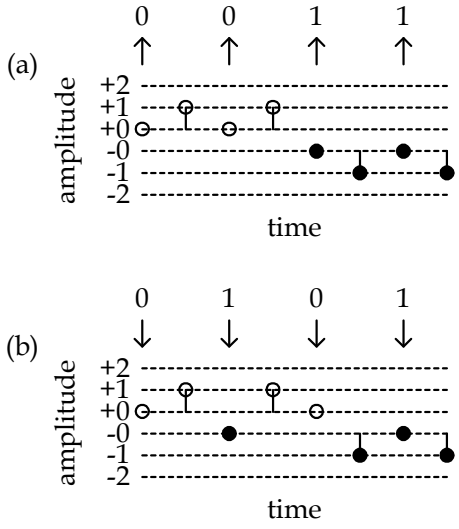

Fig. 8. Example of the proposed technique: (a) compression data before embedding, and (b) compression data after embedding.

DVI-ADPCM is a speech codec based on the ADPCM (Adaptive Differential Pulse Code Modulation) algorithm developed by DVI (Intel's Digital Video Interactive Group) (Microsoft, 1994). The block diagram of the ADPCM algorithm is shown in Fig. 9. In this diagram, $x(n)$ represents a speech data and $c(n)$ represents a compression data at the time of n.

DVI-ADPCM is designated as DVI3 and DVI4 according to the size of compression data. DVI3 encodes 16 bit speech data into 3 bit compression data at an 8 kHz sampling rate. DVI4 encodes 16 bit speech data into 4 bit compression data at an 8 kHz sampling rate. The compression data of DVI3 consists of 1 bit sign and 2 bit magnitude. The compression data of DVI4 consists of 1 bit sign and 3 bit magnitude. Both of these are encoded in the folded binary code.

Figure 10 and 11 show the decoding procedure of DVI3 and DVI4 programmed in C language. In these procedures, c is a compression data, x is a speech data, d is a difference between the previous and the current speech data, and s is a step size (Microsoft, 1994).

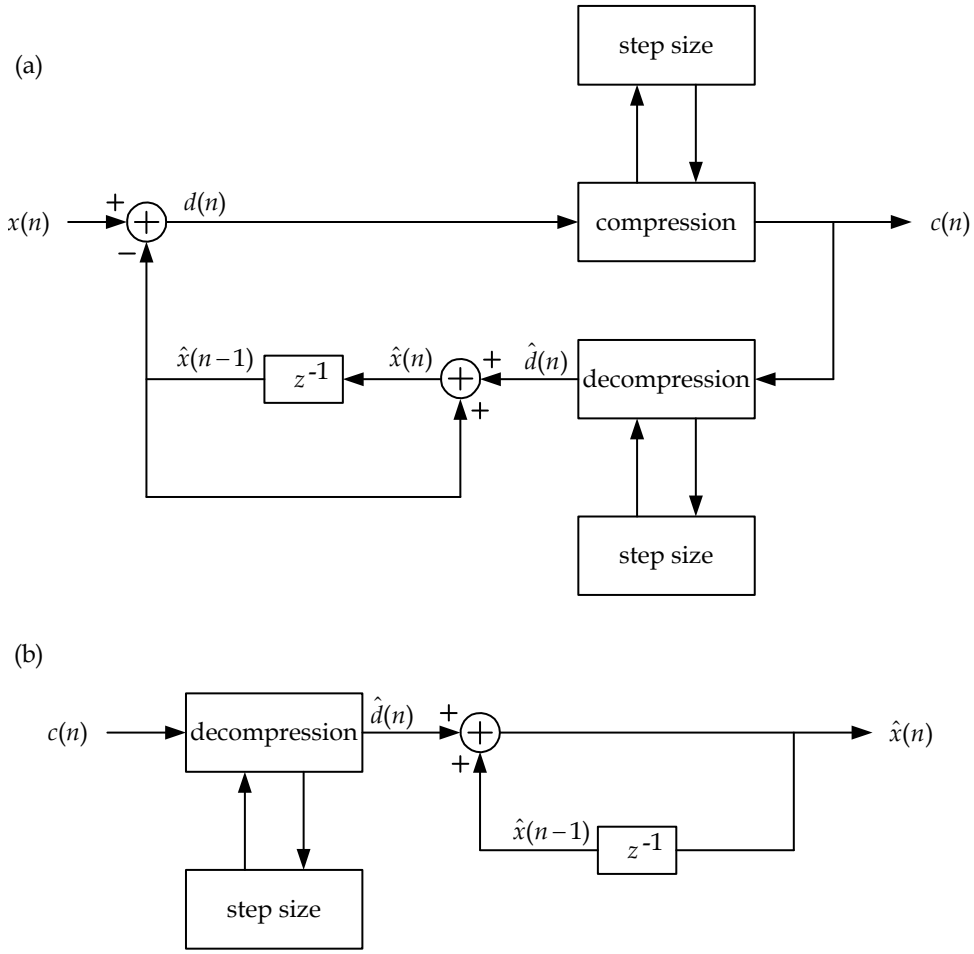

Fig. 9. Block diagram of ADPCM: (a) encoding procedure, and (b) decoding procedure.

```
d = s >> 2;
if (c & 0x1) d += s >> 1;
if (c & 0x2) d += s;
if (c & 0x4) d = -d;
x += d;
```

Fig. 10. Decoding Procedure of DVI3 programmed in C language.

```
d = s >> 3;
if (c & 0x1) d += s >> 2;
if (c & 0x2) d += s >> 1;
if (c & 0x4) d += s;
if (c & 0x8) d = -d;
x += d;
```

Fig. 11. Decoding Procedure of DVI4 programmed in C language.

7. Lossless steganography technique for DVI-ADPCM

The step size of DVI-ADPCM ranges from 7 to 32767 as shown in Fig.12 (Microsoft, 1994). When the magnitude of the compression data is 0 and the step size is 7, the speech data decoded with DVI4 does not depend on the sign of the compression data. This condition allows a lossless steganography technique that can embed secret message without any degradation. The embedding procedure of the proposed technique is defined as follows.

```
int s[89] =
{
  7, 8, 9, 10, 11, 12, 13, 14,
  16, 17, 19, 21, 23, 25, 28, 31,
  34, 37, 41, 45, 50, 55, 60, 66,
  73, 80, 88, 97, 107, 118, 130, 143,
  157, 173, 190, 209, 230, 253, 279, 307,
  337, 371, 408, 449, 494, 544, 598, 658,
  724, 796, 876, 963, 1060, 1166, 1282, 1411,
  1552, 1707, 1878, 2066, 2272, 2499, 2749, 3024,
  3327, 3660, 4026, 4428, 4871, 5358, 5894, 6484,
  7132, 7845, 8630, 9493, 10442, 11487, 12635, 13899,
  15289, 16818, 18500, 20350, 22385, 24623, 27086, 29794,
  32767
};
```

Fig. 12. Step size of DVI-ADPCM.

$$c = \begin{cases} +0 & (|c| = 0, s = 7, b = 0) \\ -0 & (|c| = 0, s = 7, b = 1) \end{cases} \tag{2}$$

where b represents a 1 bit secret message, c represents a 4 bit compression data, and s represents a step size. This procedure is programmed in C language as shown in Fig. 13.

Note that there is no such a condition that allows a lossless steganography technique for DVI3. This means that a lossless steganography technique is not available for DVI3 in the same manner of the proposed technique for DVI4.

```
if ((c & 0x7) == 0 && s == 7)
{
  if (b == 0) c &= 0x7;
  if (b == 1) c |= 0x8;
}
```

Fig. 13. Embedding procedure of the proposed technique programmed in C language.

8. Capacity of the proposed technique

The capacity of the proposed technique was evaluated by using speech data obtained from actual telephony environment, such as a private room, an office room, a cafeteria, and a railroad station. In these conditions, 8 male speech data (m1 – m8) and 8 female speech data (f1 – f8) were collected. As shown in Table 3, the duration of the speech data denoted as L was more than 120 s. The voice activity ratio of the speech data denoted as R was at around 50 %, since telephony speech generally shows the half duplex structure due to the alternate conversation process (Wright, 2001). RMS (Root Mean Square) of the background noise was calculated from voice inactive intervals.

	speech	L (s)	R (%)	RMS (dB)
private room	m1	134	51	-56.55
	m2	126	52	-59.66
	f1	124	57	-56.42
	f2	126	59	-55.72
office room	m3	127	48	-49.73
	m4	133	44	-48.59
	f3	123	49	-49.41
	f4	134	54	-48.04
cafeteria	m5	125	51	-37.48
	m6	136	48	-40.22
	f5	124	61	-37.37
	f6	123	67	-40.31
railroad station	m7	136	51	-33.87
	m8	129	68	-31.02
	f7	123	64	-34.03
	f8	126	54	-30.55

Table 3. Speech data obtained from actual telephony environment.

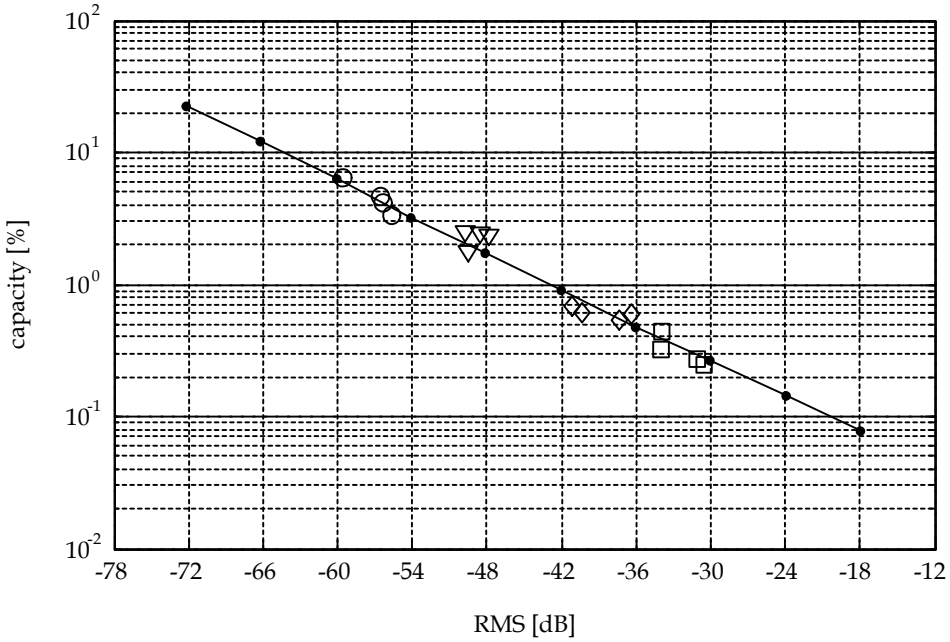

Fig. 14. Capacity of the proposed technique for PCMU: Circles, triangles, diamonds, and squares represent the capacity of a private room, an office room, a cafeteria, and a railroad station, respectively. Solid line represents the average capacity obtained from a simulation using a speech dialogue database.

The capacity of the proposed technique for PCMU is shown in Fig.14. This figure also shows a solid line that represents the average capacity obtained from a simulation using a speech dialogue database (ATR, 1997). It is indicated that the capacity of the proposed technique depends on the background noise in each telephony environment. The capacity ranges from 3.3 % to 6.4 % for the speech data obtained from a private room in which the background noise is almost imperceptible. It is interpreted that the capacity ranges from 264 bps to 512 bps in this condition. On the other hand, the capacity ranges from 0.24 % to 0.44 % for the speech data obtained from a railroad station in which the background noise is very annoying. It is interpreted that the capacity ranges from 19.2 bps to 35.2 bps in this condition.

The capacity of the proposed technique for DVI4 as well as PCMU is shown in Table 4. Compared with PCMU, the capacity for DVI4 is much smaller. Note that the capacity for DVI4 is very small even if the background noise is almost imperceptible. The capacity

ranges from 0.029 % to 0.16 % for the speech data obtained from a private room. It is interpreted that the capacity ranges from 2.32 bps to 12.8 bps in this condition. On the other hand, there is no capacity for the speech data obtained from a cafeteria and a railroad station.

	speech	PCMU (%)	DVI4 (%)
private room	m1	4.7	0.058
	m2	6.4	0.16
	f1	4.3	0.038
	f2	3.3	0.029
office room	m3	2.5	0.0032
	m4	2.4	0.0013
	f3	1.7	0.0017
	f4	2.4	0.0012
cafeteria	m5	0.58	0
	m6	0.67	0
	f5	0.54	0
	f6	0.61	0
railroad station	m7	0.44	0
	m8	0.27	0
	f7	0.32	0
	f8	0.24	0

Table 4. Capacity of the proposed technique for PCMU and DVI4.

9. Semi-lossless steganography

Semi-lossless steganography technique is an idea for increasing the capacity of the proposed technique (Aoki, 2010b). This article describes how the capacity of the lossless steganography technique for PCMU can be increased by the semi-lossless steganography technique.

Figure 15 shows how the semi-lossless steganography technique embeds secret message. In the embedding procedure, this technique modifies an 8 bit compression data as follows.

$$c' = \begin{cases} c + j & (+0 \le c \le +127 - j) \\ c - j & (-127 + j \le c \le -0) \end{cases} \tag{3}$$

where j (≥0) represents the amplitude modification level.

The amplitude modification may cause undesirable clipping in the 8 bit compression data, if its magnitude exceeds 127-*j*. Consequently, this technique can recover the original speech data only when the amplitude of the 8 bit compression data ranges from -127+*j* to +127-*j*.

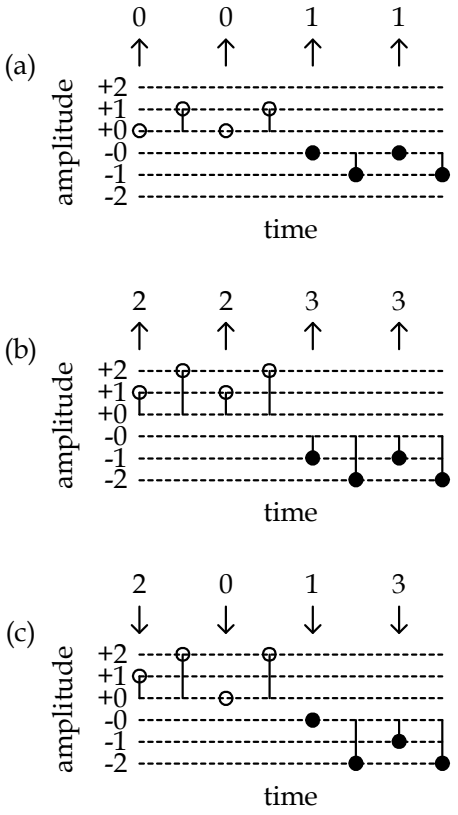

Fig. 15. Embedding procedure of the semi-lossless steganography technique:
(a) compression data before embedding, (b) amplitude modification, and (c) compression data after embedding.

Most of the practical cases meet this condition when the amplitude modification level is small enough. This is based on the fact that the amplitude of speech data statistically shows the exponential distribution (Rabiner, 1978). In general, the maximum magnitude of the 8 bit compression data is less than 127-*j*.

In this sense, this technique can be categorized as a reversible steganography technique, if this condition is satisfied. However, if this condition is not satisfied, this technique cannot recover the original speech data any more. Therefore, this technique is named semi-lossless steganography technique in this study.

The embedding procedure of the semi-lossless steganography technique is defined as follows.

$$c' = \begin{cases} +j & (|c'| = j, b = 2^j) \\ \vdots & \\ +1 & (|c'| = j, b = 2) \\ +0 & (|c'| = j, b = 0) \\ -0 & (|c'| = j, b = 1) \\ -1 & (|c'| = j, b = 3) \\ \vdots & \\ -j & (|c'| = j, b = 2^j + 1) \end{cases} \tag{4}$$

The capacity of the lossless steganography technique is defined as N bit, where N represents the number of the compression data in which the secret message can be embedded. On the other hand, the capacity of the semi-lossless steganography technique is defined as $N(\log_2(j+1)+1)$ bit. The capacity of the semi-lossless steganography technique increases according to the amplitude modification level. However, undesirable clipping may occur more frequently in such a situation.

After the extracting procedure of the secret message, the semi-lossless steganography technique recovers the 8 bit compression data as follows.

$$c = \begin{cases} c' - j & (+j < c' \le +127) \\ 0 & (-j \le c' \le +j) \\ c' + j & (-127 \le c' < -j) \end{cases} \tag{5}$$

Of course, the recovery procedure is necessary for decoding the original speech data. However, this procedure is omitted in the conventional telephony systems that do not implement the semi-lossless steganography technique. In such a situation, there is no way to remove the degradation from the speech data.

In order to evaluate such degradation, this study investigated the quality of the modified speech data by using PESQ (Perceptual Evaluation of Speech Quality) (ITU-T, 2001). PESQ is widely employed as an objective evaluation measure of the speech quality in telephony communications. Taking account of the characteristics of human auditory perception, PESQ positively correlates with a subjective evaluation measure such as MOS (Mean Opinion Score). PESQ score ranges from 4.5 to -0.5. The higher the PESQ score, the better the speech quality.

Figure 16 shows the average PESQ scores with 95 % confidence intervals. These were calculated from the 16 speech data employed in the evaluation for the capacity of the proposed technique.

As shown in this figure, it is indicated that the amplitude modification causes some degradation. However, it is almost imperceptible when the amplitude modification level is small enough. This result may potentially assure the compatibility of the semi-lossless steganography technique with the conventional telephony systems. This means that normal playback of the speech data modified with the proposed technique is still acceptable in the conventional telephony systems that do not implement the semi-lossless steganography technique.

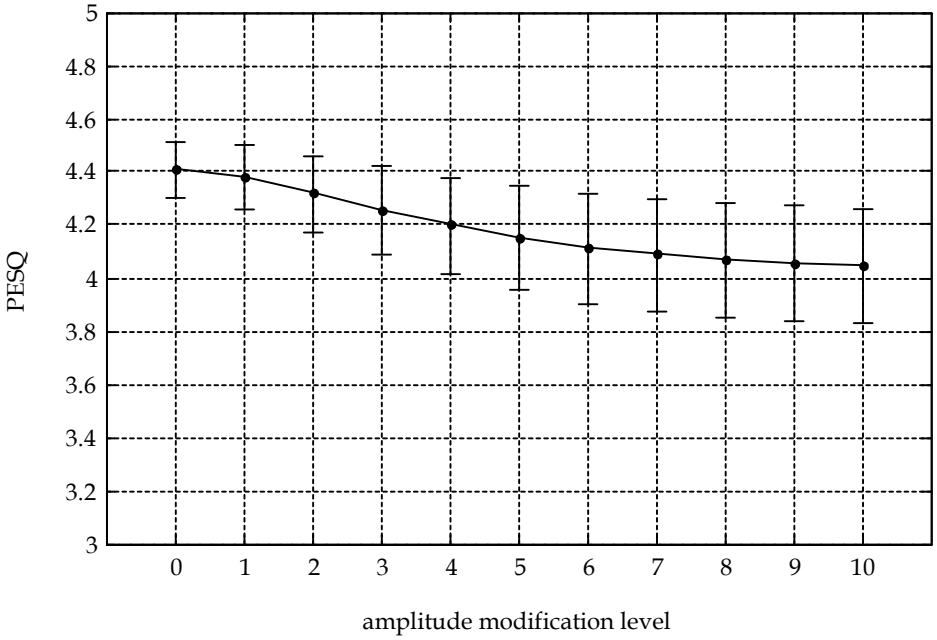

Fig. 16. Degradation of speech data by amplitude modification.

10. Conclusion

This article described an idea for the lossless steganography technique based on the characteristic of the folded binary code employed in several speech codecs, such as G.711 and DVI-ADPCM. In addition, an idea for the semi-lossless steganography technique is also described.

The proposed techniques take advantage of the redundancy of the speech codecs. It is a sort of the loophole of the speech codecs that can be employed as a container of secret message. Such a loophole plays an important role for embedding secret message without any degradation.

The concept of the proposed technique may potentially be applicable to other codecs that also employ the folded binary code. Besides G.711 and DVI-ADPCM, it is of interest to find out the codecs in which the proposed technique is available. In addition, it is also of interest to develop some practical applications that employ the proposed technique for transmitting secret message. Both of these topics are the future works of this study.

11. Acknowledgment

The author would like to express the gratitude to the Ministry of Education, Culture, Sports, Science and Technology of Japan for providing a grant (no.21760270) toward this study.

12. References

Aoki, N. (2003). A packet loss concealment technique for VoIP using steganography based on pitch waveform replication, IEICE Transactions on Communications, vol.J86-B, no.12, pp. 2551-2560

Aoki, N. (2006). A band extension technique for G.711 speech using steganography, IEICE Transactions on Communications, vol.E89-B, no.6, pp. 1896-1898

Aoki, N. (2007). A band extension technique for G.711 speech using steganography based on full wave rectification, IEICE Transactions on Communications, vol.J90-B, no.7, pp.697-704

Aoki, N. (2007). A technique of lossless steganography for G.711, IEICE Transactions on Communications, vol.E90-B, no.11, pp. 3271-3273

Aoki, N. (2008). A technique of lossless steganography for G.711 telephony speech, 2008 Fourth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP2008), Harbin, China, pp.608-611

Aoki, N. (2009). Lossless steganography techniques for IP telephony speech taking account of the redundancy of folded binary code, AICIT 2009 Fifth International Joint Conference on INC, IMS and IDC (NCM2009), Seoul, Korea, pp.1689-1692

Aoki, N. (2009). A lossless steganography technique for G.711 telephony speech, 2009 APSIPA Annual Summit and Conference (APSIPA ASC 2009), Sapporo, Japan, pp. 274-277

Aoki, N. (2010). A lossless steganography technique for DVI-ADPCM Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.J93-A, no.2, pp. 104-106

Aoki, N. (2010). A semi-lossless steganography technique for G.711 telephony speech, 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP2010), Darmstadt, Germany, pp. 534-537

Aoki, N. (2012). Enhancement of speech quality in telephony communications by steganography, Multimedia Information Hiding Technologies and Methodologies for Controlling Data, IGI Global (To be published)

ATR (1997). Speech Dialogue Database for Spontaneous Speech Recognition

Cox, I., Miller, M., Bloom, J., Fridrich, J., and Kalker, T. (2008). Digital Watermarking and Steganography, Second Edition, Morgan Kaufmann Publishers

ITU-T (1988). G.711, Pulse code modulation (PCM) of voice frequencies

ITU-T (2001). P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs

ITU-T (2005). G.191, Software tools for speech and audio coding standardization

Microsoft (1994). Multimedia Data Standards Update

Rabiner, L.R. and Schafer, R.W. (1978). Digital Processing of Speech Signals, Prentice-Hall.

RFC (1996). RFC1890, RTP profile for audio and video conferences with minimal control

Wright, D.J. (2001). Voice over Packet Networks, John Wiley & Sons