2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing

Improvement of Packet Loss Concealment for MP3 Audio Based on Switching of Concealment Method and Estimation of MDCT Signs

Akinori Ito*, Kiyoshi Konno*, Masashi Ito† and Shozo Makino‡
*Graduate School of Engineering, Tohoku University
6-6-05 Aramaki aza Aoba, Aoba-ku, Sendai, 980-8579 Japan
E-mail: aito@fw.ipsj.or.jp
†Faculty of Engineering, Tohoku Institute of Technology
35-1, Yagiyama-Kasumicho, Taihaku-ku, Sendai, 982-8577 Japan
‡Faculty of Science and Technology, Tohoku Bunka Gakuen University
6-45-1 Kunimi, Aoba-ku, Sendai, 981-8551 Japan

***Abstract*—This paper describes packet loss concealment methods for MP3 audio. The proposed methods are based on estimation of modified discrete cosine transform (MDCT) coefficients of the lost packets. The estimation of MDCT coefficients of lower dimensions is performed by switching two concealment methods: the sign correction method and the correlation-based method. The concealment methods are switched based on redundant side information calculated subband-by-subband for reducing MDCT prediction errors. Next, a method for improving estimation of MDCT coefficients of higher dimensions was proposed. The method estimates the absolute value and sign of an MDCT coefficient independently. The subjective evaluation experiment proved that both of the improvement methods for lower and higher dimensions effectively improved the subjective audio quality.**

***Keywords*-packet loss concealment; MP3; MDCT;**

I. Introduction

Recently many Web services have provided audio and video streaming services over the Internet. While most of them employ the transport control protocol (TCP) that is a unicast protocol based on connection, a connectionless protocol such as the real-time transport protocol (RTP) [1] is beneficial because of its suitability to real-time application and broadcasting. On using the RTP as a transport protocol, we have to solve a problem of packet losses. Many techniques for recovery from packet losses have been proposed so far [2], [3]. Considering audio streaming, only simple recovery methods based on packet repetition is employed for commercial audio streaming service for speech signal such as VoIP [4]. However, this is insufficient for applying to high quality audio streaming such as MP3 [5].

As a previous work, we proposed packet loss concealment (PLC) techniques for MP3-coded audio signal [6]. In this work, we proposed PLC methods that used signs of modified discrete cosine transform (MDCT) coefficients as side information. In addition to the side information, correlation of MDCT coefficient was also exploited. The sign information and correlation provided good estimation of the lost signal [7]. As a result, we could improve signal-to-noise ratio (SNR) of the decoded signal using the PLC methods.

In this paper, we first conduct a subjective evaluation of the previously proposed PLC methods, and point out that results of the objective and subjective tests are inconsistent. Next, we investigate the reason of the inconsistency, and propose a method for further improvement of subjective audio quality. Finally, we propose a method to improve quality of high frequency part of the signal where no side information is given. Combining the proposed methods, we prove the effectiveness of the methods through subjective evaluation experiment.

II. Packet loss concealment of MP3-coded audio signal using side information

A. Overview of the method

In this section, we briefly describe the PLC method proposed in [6]. In MP3 framework, the input audio signal is split into 32 subbands using a polyphase filter bank, and the signal of each subband is converted into 18 frequency components using MDCT. As a result, 576 coefficients are obtained for one frame of one channel. In this section we only consider the single-channel case for simplicity, but the multi-channel signal can be treated in the same way.

Let i-th MDCT coefficients in t-th frame be $M_i(t)$. The t-th frame $\boldsymbol{M}(t)$ is defined as

$$\boldsymbol{M}(t) = (M_1(t), \ldots, M_{576}(t)). \tag{1}$$

Next, we define sign information $\boldsymbol{S}(t)$ as

$$\boldsymbol{S}(t) = (S_1(t), \ldots, S_K(t)) \tag{2}$$

where $1 \leq K \leq 576$ and

$$S_i(t) = \mathrm{sign}(M_i(t)) \tag{3}$$

$$\mathrm{sign}(x) = \begin{cases} 1 & \text{if} \quad x \geq 0 \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

978-0-7695-4222-5/10 $26.00 © 2010 IEEE
DOI 10.1109/IIHMSP.2010.132

In the previous work [6], K was set to 50 that gave sufficient audio quality. Then we transmit MDCT coefficients of a certain frame combined with sign information of the different frame. As one MP3 packet of a channel contains two frames, the k-th packet $\boldsymbol{P}(k)$ is composed as

$$\boldsymbol{P}(k) = (\boldsymbol{M}(2k-1), \boldsymbol{S}(2k-2), \boldsymbol{M}(2k), \boldsymbol{S}(2k+1)). \quad (5)$$

When k-th packet is lost, the sign information of frames in the k-th packet can be obtained from $(k-1)$-th and $(k+1)$-th packets. Then we estimate the MDCT coefficients of t-th frame as follows.

$$\tilde{M}_i(t) = \begin{cases} f(S_i(t), M_i(t+\Delta_t)) & \text{if} \quad i \leq K \\ M_i(t+\Delta_t) & \text{otherwise} \end{cases} \quad (6)$$

where

$$\Delta_t = \begin{cases} 1 & t = 2k \\ -1 & t = 2k-1. \end{cases} \quad (7)$$

In addition, the interleaving technique can be used for increasing robustness against burst packet losses.

Estimation of the lost MDCT coefficients is performed using the function $f(s, m)$. In the *sign correction method*, the following f_0 is used:

$$f(s, m) = f_0(s, m) = s|m| \quad (8)$$

The other method is the *correlation-based estimation method*, where the following f_1 is used:

$$f(s, m) = f_1(s, m) = s\left(w_i|m| + (1-w_i)\sqrt{\frac{2\sigma_i^2}{\pi}}\right) \quad (9)$$

$$w_i = \frac{|\rho_{i1}|(|\rho_{i2}|-1)}{2|\rho_{i1}\rho_{i2}| - |\rho_{i1}| - |\rho_{i2}|} \quad (10)$$

where σ^2 is variance of $M_i(t)$, ρ_{i1} is correlation coefficient between $|M_i(t)|$ and $|M_i(t-1)|$, and ρ_{i2} is that between $M_i(t)$ and $S_i(t)$.

As predicted in [7], the correlation-based method gave higher SNR than the sign correction method.

B. Objective and subjective evaluation of the conventional methods

Although the correlation-based method showed better SNR, it does not necessarily mean that subjective quality of the signal restored by the correlation-based method is better than that by the sign correction method. To investigate the relationship between objective and subjective quality, we carried out an evaluation experiment. The experimental conditions are shown in Table I.

SNR was used as an index of objective evaluation, and 5-grade DMOS (degradation mean opinion score) was used as that of subjective evaluation. The result is shown in Table II. As this result depicts, the correlation-based method improved the objective evaluation, but the subjective evaluation was not improved.

Table I
EXPERIMENTAL CONDITIONS

Testing materials	Three pieces from the RWC music DB [8] drawn from three genres (Classic, Pop, Jazz), 15s each
Signal format	44.1kHz 16bit sampling, 2 channels
Packet length	2 frames (2304 samples/packet/ch)
Loss rate	10%, 20% (random loss)
Encoding	MP3, 160kbps
Subjects	4 males

Table II
RESULT OF OBJECTIVE AND SUBJECTIVE EVALUATIONS

Evaluation	Sign correction	Correlation-based
SNR [dB]	11.7	12.7
DMOS	3.99	3.50

We then investigated the cause of this inconsistency. To this end, we focused on degradation that occur locally in time-frequency domain. To analyze the degradation, we calculated SNR subband by subband. One subband in this analysis consists of five dimensions of the MDCT coefficients. Examples of subband SNR are shown in Figure 1. As these examples show, the correlation-based method gives better SNR for most of subbands when the degradation is not severe, especially lower subbands that have larger power. However, when severe degradation occurs, subbands with very low SNR are observed in the correlation-based method, which seems to be a cause of degradation of subjective audio quality. These local degradations of SNR happen at which the subband power change rapidly, which causes large error of estimated values of correlation coefficients and variance of MDCT coefficients.

III. IMPROVEMENT OF MDCT ESTIMATION USING CONCEALMENT METHOD SWITCHING

As revealed in the previous experiment, the correlation-based method gives better SNR for most of subbands and frames, but the SNR occasionally takes very low value at several subbands and frames. Therefore, the basic idea of improving the correlation-based method is to avoid using the correlation-based method when estimation error of an MDCT coefficient is large. This idea can be achieved by examining estimation errors by the two estimation methods (sign correction and correlation-based method) at the encoder side and sending information of better method as another side information.

This improvement can be realized as follows. First, we consider sending the side information for switching estimation method for every 5 dimensions of the MDCT coefficients when sending packets. Assuming K be a multiple of 5, we calculate errors of the t-th subband $\varepsilon_x^2(t, j)$

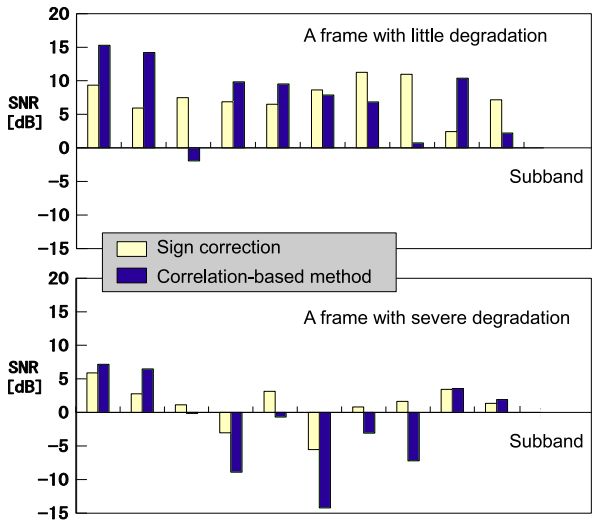

Figure 1. Examples of subband SNR for frames with little and severe degradation

$(j = 1, \ldots, K/5)$ as

$$\varepsilon_x^2(t,j) = \sum_{i=5j-4}^{5j} (M_i(t) - f_x(S_i(t), M_i(t+\Delta_t)))^2 \quad (11)$$

where x is either 0 or 1 where f_0 or f_1 is used as an estimation method, respectively. Then the information for method switching is calculated as follows:

$$Q_j(t) = \begin{cases} 1 & \text{if} \quad \varepsilon_1^2(t,j) < \varepsilon_0^2(t,j) \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

Using $Q_j(t)$, a packet is composed as

$$\begin{aligned} \boldsymbol{P}(k) \;=\; & (\boldsymbol{M}(2k-1), \boldsymbol{S}(2k-2), \boldsymbol{Q}(2k-2), \\ & \boldsymbol{M}(2k), \boldsymbol{S}(2k+1), \boldsymbol{Q}(2k+1)) \end{aligned} \quad (13)$$

where

$$\boldsymbol{Q}(t) = (Q_1(t), \ldots, Q_{K/5}(t)). \quad (14)$$

When t-th packet is lost, the MDCT coefficient is estimated as follows:

$$\tilde{M}_i(t) = \begin{cases} f_{Q_j(t)}(S_i(t), M_i(t+\Delta_t)) & \text{if} \quad i \le K \\ M_i(t+\Delta_t) & \text{otherwise} \end{cases} \quad (15)$$

where

$$j = \left\lfloor \frac{i-1}{5} \right\rfloor + 1. \quad (16)$$

Note that this method needs another side information $\boldsymbol{Q}(t)$, which means we need to use smaller K for using same amount of side information as the conventional methods.

IV. Estimating higher MDCT coefficients without side information

The method described in the previous section is to improve estimation of MDCT coefficients using side information. However, as shown in Eq. (6) and (15), the MDCT coefficients higher than the K-th dimension are estimated by just using the coefficients of the neighboring frames. Therefore, if we can improve estimation of these coefficients, we achieve further improvement of recovered audio quality.

Two ideas are introduced to achieve this: one is to improve estimation of absolute value of MDCT coefficients, and the other one is to improve sign estimation of MDCT coefficients.

In the previous method, the absolute values of $M_i(2k-1)$ and $M_i(2k)$ for $i > K$ are estimated as

$$|\tilde{M}_i(2k-1)| \;=\; |M_i(2k-2)| \quad (17)$$
$$|\tilde{M}_i(2k)| \;=\; |M_i(2k+1)|. \quad (18)$$

Here, we propose a method to estimate the absolute value using a linear interpolation:

$$D_i(k) = \frac{|M_i(2k+1)| - |M_i(2k-2)|}{3} \quad (19)$$

$$|\tilde{M}_i(2k-1)| \;=\; |M_i(2k-2)| + D_i(k) \quad (20)$$
$$|\tilde{M}_i(2k)| \;=\; |M_i(2k-2)| + 2D_i(k). \quad (21)$$

Next, we propose an estimation of sign of an MDCT coefficient. In the previous method, signs of MDCT coefficients $(i > K)$ are estimated as

$$\tilde{S}_i(2k-1) \;=\; S_i(2k-2) \quad (22)$$
$$\tilde{S}_i(2k) \;=\; S_i(2k+1). \quad (23)$$

Instead of just using signs of the previous or next frame, we propose a method of estimating signs based on bigram probability, as follows.

$$\tilde{S}_i(t) = \arg\max_s P(s|S_i(t-1), t, i) \quad (24)$$

where $P_t(s|s', t, i)$ is a conditional probability of s where the sign of the previous frame is s' $(s, s' \in \{-1, 1\})$. The probability is estimated from the previous N frames $S_i(t-N), \ldots, S_i(t-1)$. In the experiment of the next section, the previous 50 frames were used $(N = 50)$ according to the result of a preliminary experiment. The actural calculation of $\tilde{S}_i(t)$ is performed in simpler way, as follows.

$$T_i(t) = \sum_{\tau=t-N}^{t-2} S_i(\tau) S_i(\tau+1) \quad (25)$$

$$\tilde{S}_i(t) = \text{sign}(T_i(t)) S_i(t-1) \quad (26)$$

V. Experiment

A. Experimental conditions

We conducted subjective evaluation of the conventional and proposed methods. The sign correction method and the switching method were examined as the estimation method of MDCT coefficients of lower dimensions. For the higher dimensions, we compared conventional method (denoted as Copy) and the proposed method that estimates absolute values and signs independently (denoted as Interp and sign est.). 50 bit/frame were used as the side information of the conventional method ($K = 50$), and 48 bit/frame were used for the proposed method ($K = 40$, 40 bits for signs and additional 8 bits for method switching).

We employed 9 male subjects for evaluation, and the packet loss condition was limited to 20% random loss. The other conditions were same as shown in Table I.

The evaluation was conducted based on the XAB test [9]. The subjective difference grades (SDG) were calculated as scores of the test signals, which varies from -4 to 0, 0 to be the best (degradation is imperceptible).

B. Experimental results

Figure 2 shows the experimental result. This result showed that both improvement of lower dimensions (the concealment method switching) and that of higher dimensions (linear interpolation of absolute values and sign estimation) were effective for improving the subjective audio quality. From result of one-way layout ANOVA, statistically significant differences were found between concealment methods. Then we conducted multiple comparison test with Bonferroni correction, and significant differences were found between every two concealment methods, as shown in Figure 2.

As a result, we obtained SDG score of -1.78, which was 1.3 point higher than the conventional method. Moreover, absolute quality of the proposed method (-1.78) was reasonably high (better than "slightly annoying") considering that the packet loss rate was severe (20%).

VI. Conclusion

We proposed methods for improving subjective audio quality. First, we conducted subjective evaluation for conventional PLC methods and investigated the cause of quality degradation using the correlation-based method. Next, we proposed a method that switches the sign correction method and the correlation-based method based on subband-by-subband errors. Furthermore, an improvement method of MDCT coefficients of higher dimensions was proposed. The subjective evaluation experiment proved that both of the improvement methods for lower and higher dimensions were effective.

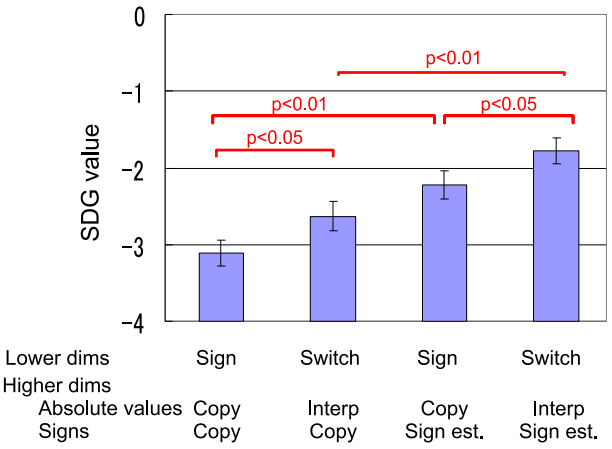

Figure 2. Result of the subjective evaluation experiment

References

[1] C. Perkins, "RTP: Audio and Video for the Internet," Addison-Wesley, 2003.

[2] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, 1998.

[3] B. W. Wah, X. Su, and D. Lin, "A survey of error-concealment schemes for real-time audio and video transmissions over the Internet," in *Proc. IEEE Int. Symp. on Multimedia Software Engineering*, 2000.

[4] ETSI, "Substitution and muting of lost frames for full rate speech channels." ETSI Recommendations GSM 6.11, 1992.

[5] ISO, "Information technology — coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s — part 3: audio." ISO/IEC 11172-3, 1993.

[6] A. Ito, T. Sakai, K. Konno, S. Makino and M. Suzuki, "Packet loss concealment for MDCT-based audio codec using correlation-based side information," Int. J. of Innovative Computing, Information and Control, vol. 6, pp. 1347–1362, 2010.

[7] A. Ito and S. Makino, "Designing side information for multiple description coding," J. of Information Hiding and Multimedia Signal Processing, vol. 1, pp. 10–19, 2010.

[8] M. Goto, "Development of the RWC music database," in *Proc. of 18th Int. Cong. on Acoustics*, pp. 553–556, 2004.

[9] International Telecommunication Union, "Methods for the subjective assessment of small impairments in audio systems, including multichannel sound systems." Recommendation ITU-R BS. 1116-1, 1997.