Enhancement of QoE in Audio–Video IP Transmission by Utilizing Tradeoff between Spatial and Temporal Quality for Video Packet Loss

Shuji Tasaka and Hikaru Yoshimi
Department of Computer Science and Engineering,
Graduate School of Engineering, Nagoya Institute of Technology, Nagoya 466–8555, Japan
Email: tasaka@nitech.ac.jp

***Abstract*— This paper proposes a methodology of video–stream output at the receiver for enhancing QoE (Quality of Experience) in audio–video IP transmission. The methodology copes with video packet loss by error concealment and/or video frame skipping; it utilizes the tradeoff relation of QoE between spatial and temporal quality caused by the two techniques. As a simple example of the methodology, we adopt a scheme which switches between the two techniques according to the percentage of video slices error–concealed in a frame; the scheme is referred to as *SCS* (*Switching between error Concealment and frame Skipping*). We conducted experiments on the SCS with six contents. Taking into consideration the cross–modal interaction between audio and video, we then assessed QoE in terms of the psychological scale, which is more accurate than MOS (Mean Opinion Score). The experimental result shows that the SCS can improve QoE over the simple error concealment or frame skipping by selecting an appropriate threshold value of the error concealment ratio, which depends on the content type, video picture pattern and degree of video motion.**

I. Introduction

The ultimate goal of application services over the networks is to realize the overall acceptability as perceived subjectively by the end–user; this means guarantee of *QoE* (*Quality of Experience*) [1]. From a layered architectural point of view, QoE is identified as *user–level QoS* (*Quality of Service*). In the context of the IP networks, which are becoming increasingly important in practice and as the *Next Generation Network* (*NGN*) [2], the user–level QoS is on the top of *application–level* QoS[1].

One of the most important application services in the IP networks is audio–video transmission, which supports many popular services over the current Internet and is also considered an essential ingredient of multimedia application services in the NGN. In the IP networks, the quality of audio–video streams output at the receiver can deteriorate owing to packet loss, error and delay; this leads to the degradation of QoE.

It is often the case that the receiver resorts to some decoding technique of corrupted compressed streams to enhance the output audio–video quality; typical examples are *video error concealment* and *video frame skipping*.

The video error concealment intends to conceal the visual effects of packet loss and error by exploiting the temporal or spatial correlation with adjacent data [4], [5]. The temporal approach replaces a missing block with some appropriate block in a previously decoded frame; this concealment is usually applied to inter–coded frames. On the other hand, the spatial approach restores the missing blocks using the information in the current frame; for example, a missing block is interpolated from its four neighboring blocks. In either approach, however, the concealment is not necessarily perfect; therefore, it causes residual errors such as artifacts, and the errors can propagate to the succeeding frames. This degrades the spatial quality of the output video streams. At the expense of this degradation, the output frame rate is kept high; this implies high temporal quality.

The technique of video frame skipping at the receiver does not decode a frame unless all packets of the frame are correctly received. This results in skipping the frame, and the frame skipping continues until an intra-coded frame is decoded. Consequently, the spatial quality of the frames thus output is kept original, while the temporal quality degrades because of the decrease in the output frame rate.

The above observations suggest that there exists a tradeoff relation between temporal quality and spatial quality when error concealment and frame skipping are employed to cope with video packet loss and error; the former technique improves the temporal quality over the latter, while the latter keeps the spatial quality of output frames higher than the former. Thus, we can expect that an appropriate mixture of the two techniques enhances QoE compared to the adoption of either technique. To the best of the authors' knowledge, however, we can find no study on QoE from this point of view in the literature.

In addition to the QoE tradeoff issue above, we also notice three limitations in previous studies on video IP transmission in the context of QoE research.

First, the great majority of the previous studies treat no audio in spite of the fact that video is accompanied by audio in most applications. It has been recognized that audio and video interact with each other from a QoE point of view [6]; ITU-T has paid much attention to this cross–modal interaction and has established ITU–T Recommendation J.148 [7]. So, QoE assessment of video only is not sufficient for most multimedia applications, where we should consider audio and video together.

Secondly, many of the previous studies assess the output quality of video IP transmission only in terms of the *PSNR* (*Peak Signal to Noise Ratio*), which is not a QoE metric but an application–level QoS parameter representing the spatial quality of video; we need quality assessment with some QoE metric.

The third limitation is closely related to the second one; the PSNR–only approach implies no assessment of the temporal quality of video. That is, the delay jitter of received packets is not taken into consideration in the assessment. This is often validated when the receiver prepare the playout buffer to absorb the delay jitter. As a matter of fact, however, the delay jitter is not absorbed perfectly since the buffering time cannot be set to infinite. Packets arriving late are either discarded or output with jitter; in both cases, the temporal quality of the output video stream degrades. We must consider this degradation in the overall QoE assessment.

Thus, we should study the QoE tradeoff issue by taking account of the above three limitations of the previous studies. The purpose of this paper is to give a solution to this problem.

[1] In IP networks, six kinds of QoS are identified along the protocol stack: *physical–level*, *node(link)–level*, *network–level*, *end–to–end–level*, *application–level*, and *user–level* [3].

More precisely, in order to enhance the overall QoE of output audio–video streams, we propose a methodology of video-stream output utilizing the QoE tradeoff relation. As a first step toward the study on the methodology, this paper adopts a simple scheme of mingling error concealment and frame skipping, which switches between the two techniques according to the percentage of video slices error–concealed. We refer to this scheme as *SCS* (*Switching between error Concealment and frame Skipping*). Making an experiment on the scheme, we show the existence of the QoE tradeoff relation and the feasibility of QoE enhancement by the relation.

The authors have already addressed themselves to avoiding the three limitations mentioned earlier in a study on real–time estimation of QoE in audio–video IP transmission [8]. The study quantifies QoE of audio and video jointly in terms of *the psychological scale* [9], which is an *interval scale* in the psychometric method [10], [11]. The psychological scale can express human subjectivity more accurately than *MOS* (*Mean Opinion Score*), which is usually used as the QoE metric in ITU–T Recommendations[2]. This paper tackles the QoE problem by means of the psychological scale.

The remainder of the paper is organized as follows. Section II introduces the SCS scheme. Section III demonstrates an experimental network, contents to be assessed, and a method of measuring QoE with the psychological scale. Section IV presents experimental results and examines the QoE tradeoff relation and the effectiveness of the SCS. Section V concludes the paper.

II. The SCS Scheme

This section introduces the SCS, which is a simple scheme of mingling video error concealment and frame skipping techniques so that it can improve QoE over the employment of either technique. As mentioned earlier, the purpose of this paper is to show the feasibility of QoE enhancement with this methodology. Therefore, in this paper, we try to make the scheme of the methodology as simple as possible.

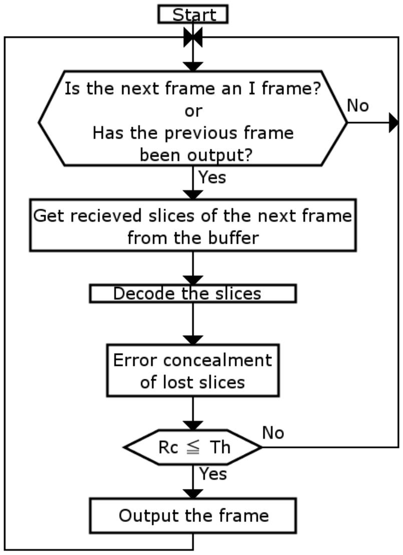

Fig. 1. Operation of the SCS scheme

Thus, as the SCS, we employ a video output scheme whose mode goes back and forth between error concealment and frame skipping. The initial mode is error concealment, and it then switches to frame skipping once the percentage of slices error–concealed in a frame, which we call the *error concealment ratio* and denote by R_c, exceeds a threshold value; the frame skipping continues until an intra–coded frame is decoded, at which time the mode switches back to the error concealment. Let T_h be the threshold value in units of %. Figure 1 describes how the SCS scheme operates.

For a video stream with a picture pattern of IPPPPP, we present examples of the output by the SCS when $T_h = 100, 40, 20$ and 0 in Fig. 2, where the large quadrilateral with bold lines represents an output frame, while the one with thin lines corresponds to a skipped one. A quadrilateral with a hatched area means an error–concealed slice. The figure in the parenthesis such as 10% is the value of R_c for the frame. Note that the case of $T_h = 100$ is equivalent to the pure error concealment technique, whereas $T_h = 0$ implies the simple frame skipping without error concealment.

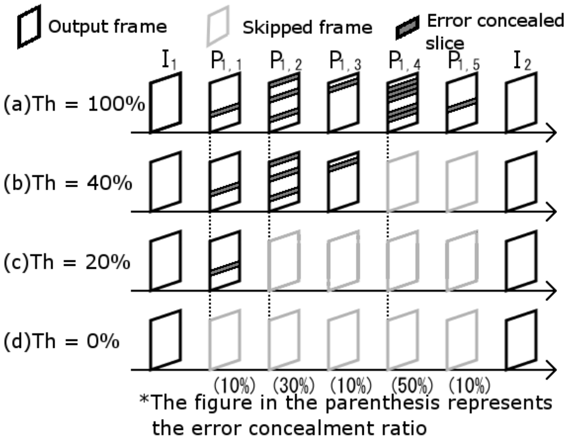

Fig. 2. Examples of the video stream output by the SCS scheme

III. Experimental Method

This section explains an experimental method to show the existence of the QoE tradeoff relation and the feasibility of QoE enhancement by the relation. Since the details of the relation are considered to depend on content types of the audio–video stream, we prepare three types: two contents for each type. We first present an experimental network along with the contents. We then describe how QoE is measured.

A. Experimental network and contents

1) Network configuration: Figure 3 shows the configuration of the experimental network; it consists of two routers (RiverStone's RS3000) and four PC's, which are used as a media sender (MS), a media recipient (MR), a Web server (WS), and a Web client (WC). The link between the routers and ones between a router and a PC are all Ethernet channels of 100 Mb/s.

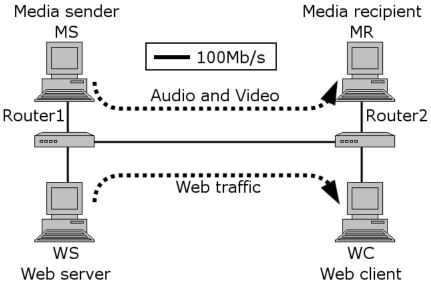

Fig. 3. Configuration of the experimental network

The MS transmits an audio stream and the corresponding video stream to the MR; the information unit for transfer between the application layers is referred to as the *MU* (*Media Unit*). A video MU is usually defined as a video frame and an audio MU as a constant number of audio samples. The video

[2] The accuracy of MOS as a QoE metric in comparison with the psychological scale is discussed in [9] and [12].

MU's and audio ones are transferred as two separate streams with RTP/UDP. The MR exerts playout buffering control of 1 second to absorb delay jitters of received MU's.

Table I gives specifications of audio and video used in the experiment. An audio MU composes a single UDP datagram, while a video MU is divided into 15 UDP datagrams each of which corresponds to a slice. We have set three kinds of picture patterns, which are I followed by $n-1$ P's (n=1, 5, and 15), to examine their effects on the QoE tradeoff relation.

TABLE I

SPECIFICATIONS OF AUDIO AND VIDEO

audio coding scheme	Linear PCM 24kHz 16bit 1ch
audio MU size [byte]	960
audio average bit rate [kb/s]	384 (=50MU/s)
video coding scheme	H.264 (JM11.0)
image size [pixel]	320×240
number of slices in a picture (interleave mode)	15 (20 macroblocks/slice)
video average MU rate [MU/s]	30
picture pattern	I IPPPP IPPPPPPPPPPPPPP
recording time [s]	10

As in [13], Web traffic is transferred from the Web server (WS) to the Web client (WC); it is generated according to the configuration of WebStone 2.5 [14]. WebStone generates Web client processes on the WC PC; those client processes retrieve specified files from the WS PC continuously. In the experiment, the number of the Web client processes were set to 20, 30, 40, 50, 75 and 100. As the number of the processes increases, the amount of the Web traffic becomes larger, and therefore packet loss occurs more frequently.

As the error concealment technique in this paper, we employ the one implemented in H.264/MPEG–4 AVC reference software JM11.0 [15]. For I–frames, the spatial approach is utilized. For P–frames, two techniques of the temporal approach are available: Frame Copy and Motion Copy. The former simply replaces the missing block with the spatially corresponding one of the previously output frame, while the latter utilizes the information of the motion vector in the replacement. This paper selects the Frame Copy scheme for simplicity.

In the experiment on the SCS, we set the threshold value T_h to 100%, 40%, 20% and 0%.

2) Contents: Referring to the VQEG multimedia test plan [16], we have selected three types of contents: *sport, animation* and *music video*. Sport has been selected as a video–dominant content type, where video plays a more important role than audio, while music video is considered audio–dominant. Animation has different features from sport and music video, especially in video with respect to the picture property and frame rate; the animation is usually made at a lower frame rate (say 24 fps or less) than the others.

For each content type, we have prepared two contents; thus, we have six contents totally. Outlines of them are as follows:

- *sport 1*: A group of people are doing aerobics, timing their movement to the instructor's voice and music. There is no scene change.
- *sport 2*: A racing car is running at a high speed on a narrow road (one scene change). The audio is composed of only the roar of the engine.
- *animation 1*: Two human characters are talking, while one is shouting, without background music (five scene changes).
- *animation 2*: A robot is trying to rescue a shouting boy from falling in the sky with fast background music (frequent scene changes).
- *music video 1*: A sitting young male is playing the ukulele without scene change. The audio is the ukulele's sound only.
- *music video 2*: A female singer is playing the piano and then singing while dancing. There are two scene changes.

Note that the second content of the same type (say sport 2) has higher motion video than the first one (say sport 1). Table II shows the value of *TI*(*Temporal perceptual Information*) for each content in addition to the video average bit rate. The TI measure is defined in ITU–T Rec. P.911 [17]; it indicates the amount of temporal changes of a video sequence. A higher value implies higher motion. The TI values in this table have been calculated by eliminating the effect of scene changes.

TABLE II

VIDEO AVERAGE BIT RATE AND TI VALUE FOR EACH CONTENT

Content	Average bit rate [kb/s]			TI value
	Picture pattern			
	I	IPPPP	I+14 P's	
sport 1	2592.713	1243.395	1040.902	16.652
sport 2	2241.426	1228.215	1076.502	55.773
animation 1	1482.044	521.113	356.598	36.552
animation 2	1440.342	641.110	518.262	38.455
music video 1	1890.361	788.026	618.478	12.924
music video 2	1658.756	790.520	690.626	48.597

B. Method of calculating QoE metric

As mentioned in Section I, this paper expresses QoE (user–level QoS) in terms of the interval scale, namely, the psychological scale [9]; we do not use MOS because of higher accuracy of the interval scale.

For the calculation of the interval scale, as in [9], this paper adopts the *method of successive categories*, which is composed of two steps: the *rating–scale method* and the *law of categorical judgment*. The rating–scale method specifies how the subjective measurement is made on *stimuli*, which are audio-video streams output at the receiver in our case; an assessor classifies the stimuli into a certain number of categories (e.g., five) each assigned an integer score (typically 5 through 1 in order of highly perceived quality). From the measurement results by the rating–scale method, the law of categorical judgment provides the interval scale [3].

Since the law of categorical judgment is based on several assumptions, we have to confirm the goodness of fit for the obtained scale. For a test of goodness of fit, we conduct *Mosteller's test* [10], [18]. Once the goodness of fit has been confirmed, we use the interval scale as the psychological scale.

C. Subjective measurement by the rating–scale method

We define an experimental run as the transmission of a content with a picture pattern at a constant level of the average Web traffic (i.e., when the number of Web client processes is kept constant). During each experimental run, we recorded the audio–video streams that the media recipient (MR) output; the recorded streams are regarded as stimuli for QoE measurement. Thus, we totally have 432 stimuli because of six contents, three picture patterns for each content, four values of T_h, and six levels of the average Web traffic.

In the rating–scale method, we utilized the following five categories of *impairment*: "imperceptible" assigned score 5, "perceptible, but not annoying" 4, "slightly annoying" 3, "annoying" 2, and "very annoying" 1, which are referred to as *Category 5* through *Category 1*, respectively.

[3] The MOS is calculated by simply taking an average of the scores for a stimulus over all assessors. However, it should be noted that this way makes an implicit assumption that the difference in score between any two successive categories means the same magnitude of the assessor's sensation (e.g., "$5-4$" has the same meaning as "$3-2$"). The assumption is not necessarily valid as shown in [9] and [12]. Thus, in the strict sense, MOS is an *ordinal scale*, which only has a greater–than–less–than relation between scores given by assessors. The law of categorical judgment does not make the above assumption.

We put the 432 stimuli in a random order and presented them to 32 assessors, using a PC with headphones and a 17–inch LCD display. The distance between the display and each assessor was set to that in the case where he/she usually uses a PC (i.e., approximately 50 cm through 1 m).

The assessors are Japanese males at twenties. They were non–experts in the sense that they were not directly concerned with audio and video quality as a part of their normal work. It took about 4.5 hours including break time for an assessor to assess all the stimuli.

IV. Experimental Results

This section first calculates the psychological scale and then presents the calculated result. Using the result, we examines the QoE tradeoff relation, in particular, how the contents, picture patterns and the degree of video motion affect the psychological scale value and how effective the SCS is.

A. Calculation of the psychological scale

In order to calculate the interval scale, we applied the law of categorical judgment to all the classification results of the 432 stimuli together. This way of the calculation has been selected so that we can compare the interval scales for the six contents each with three picture patterns on the same basis.

We carried out Mosteller's test for a test of the goodness of fit of the interval scale. We then found that the test with a significance level of 0.05 can reject the hypothesis that the observed value equals the calculated one. So, we checked stimuli which give large errors of Mosteller's test to find 40 ones. Removing the 40 stimuli, we saw that the hypothesis cannot be rejected. Consequently, for the 392 (= 432 − 40) stimuli, we can consider the interval scale as the psychological scale.

Since we can select an arbitrary origin in an interval scale, we set the minimum value of the psychological scales for the 392 stimuli to the origin. Under this condition, we also calculated the lower boundaries of the categories and got 4.649 for Category 5, 3.532 for Category 4, 2.371 for Category 3, and 1.010 for Category 2.

B. QoE tradeoff relation and effectiveness of SCS

We now examine the QoE tradeoff relation by the SCS for each of the three picture patterns in turn.

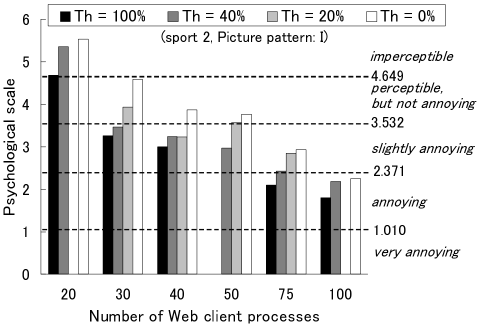

Fig. 4. Psychological scale versus number of Web client processes (sport 2, Picture pattern: I).

1) Picture pattern I: Figures 4, 5 and 6 plot the psychological scale at the four values of the SCS threshold T_h (100%, 40%, 20% and 0%) as a function of the number of Web client processes for sport 2, animation 2, and music video 2, respectively. In these figures, the lower boundaries of the categories are also plotted as straight dotted lines parallel to the abscissa. It should be noted that the results of the stimuli removed by Mosteller's test are not shown in the figures.

We then easily see that $T_h = 0\%$ (namely, the pure frame skipping) achieves the highest QoE among the four threshold values in almost all lossy environments[4]. We have confirmed that this is also the case with sport 1, animation 1, and music video 1, which have lower motion video than the three contents shown here.

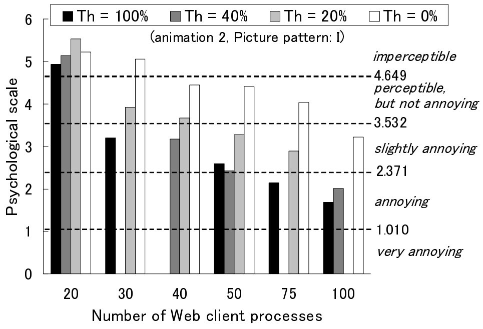

Fig. 5. Psychological scale versus number of Web client processes (animation2, Picture pattern: I).

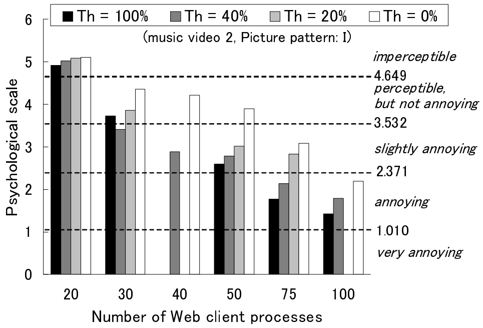

Fig. 6. Psychological scale versus number of Web client processes (music video 2, Picture pattern: I).

2) Picture pattern IPPPP: Figures 7 through 12 display the psychological scales for the six contents when the picture pattern is IPPPP. Note that skipping a frame in this picture pattern results in skipping all succeeding P frames.

Regarding sport 1 and sport 2, we observe that nonzero values of T_h provide higher QoE than $T_h = 0\%$. This is because the content type of sport is video–dominant and therefore the error concealment is effective.

Animation exhibits different results from sport. First, $T_h = 0\%$ is still better than the nonzero values. Second, the result is more remarkable in animation 1 than in animation 2. We can understand the first result by considering the fact that the animation is usually made at a frame rate lower than 30 (say 24 fps or less) and therefore the effect of the decrease in the output frame rate on human perception is small. The second one is due to higher motion of animation 2.

In music video 1, we find that $T_h = 0\%$ still achieves the highest QoE in lossy environments, namely, when the number of Web client processes is 40 and more. This is because music video 1 is an audio–dominant content with low motion. As for music video 2, on the other hand, nonzero threshold values give higher psychological scale values, since this content has high motion video and the video quality affects human subjectivity more than that of music video 1.

3) Picture pattern IPPPPPPPPPPPPPP: Figures 13, 14 and 15 present the results of sport 1, animation 2, and music video 1, respectively, for this picture pattern.

Comparing Fig. 13 and Fig. 7, we notice that the advantage of $T_h = 100$ over the other values in this picture pattern becomes clearer than that in the picture pattern of IPPPP; that is, the error concealment works more effectively. This is reasonable because of the video–dominant feature of sport.

Figure 14 indicates that nonzero T_h values achieve higher QoE than the zero value; this contrasts with the results in

[4] When the number of Web client processes is 20, packet loss scarcely occurred.

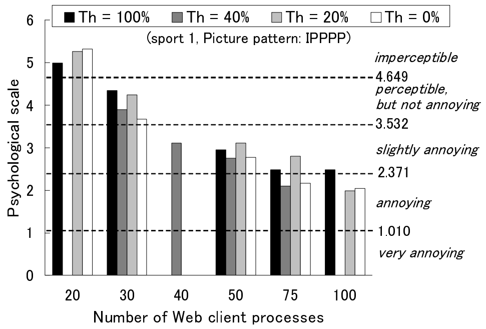

Fig. 7. Psychological scale versus number of Web client processes (sport 1, Picture pattern: IPPPP).

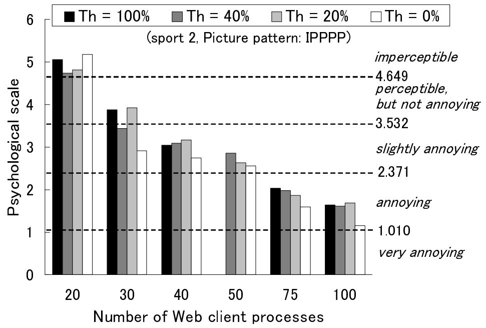

Fig. 8. Psychological scale versus number of Web client processes (sport 2, Picture pattern: IPPPP).

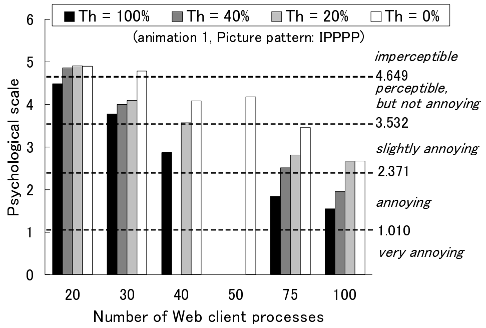

Fig. 9. Psychological scale versus number of Web client processes (animation 1, Picture pattern: IPPPP).

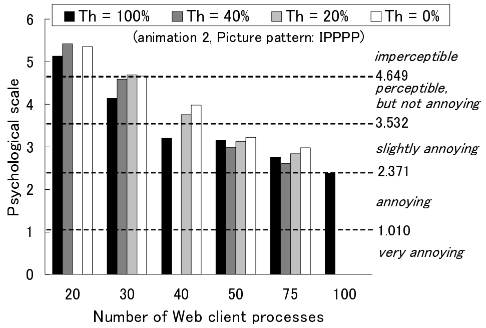

Fig. 10. Psychological scale versus number of Web client processes (animation 2, Picture pattern: IPPPP).

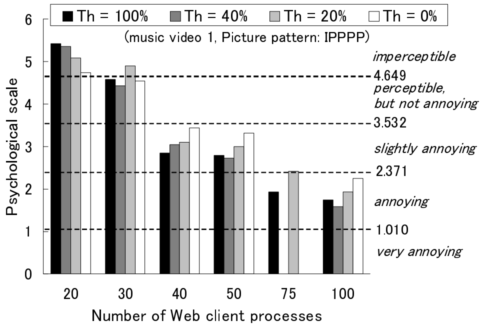

Fig. 11. Psychological scale versus number of Web client processes (music video 1, Picture pattern: IPPPP).

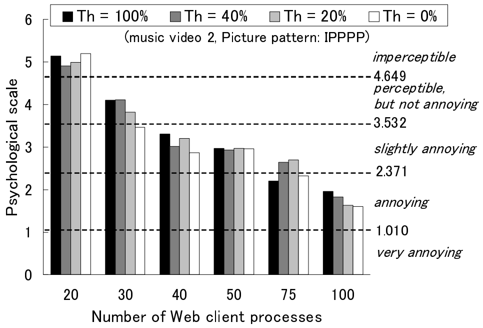

Fig. 12. Psychological scale versus number of Web client processes (music video 2, Picture pattern: IPPPP).

Figs. 5 and 10. Thus we know that even in the content type of animation, the temporal quality becomes important for picture patterns with many P's.

In Fig. 15, we find that a larger value of T_h provides a larger value of the psychological scale in lossy environments. This means that although music video is audio–dominant, the spatial quality should be preserved for this picture pattern.

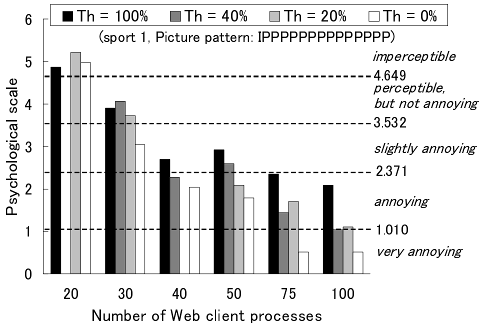

Fig. 13. Psychological scale versus number of Web client processes (sport 1, Picture pattern: IPPPPPPPPPPPPPP).

C. A way of setting T_h

We now consider how to set the threshold value T_h. In order to find a way of doing this, let us summarize the observation made in the previous subsection.

First, we have confirmed the existence of the QoE tradeoff relation and the effectiveness of the SCS, which depends on the content type, picture pattern, and degree of video motion.

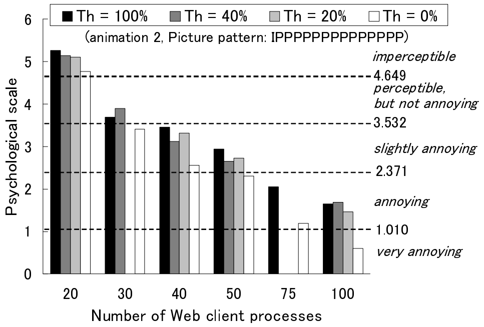

Fig. 14. Psychological scale versus number of Web client processes (animation2, Picture pattern: IPPPPPPPPPPPPPP).

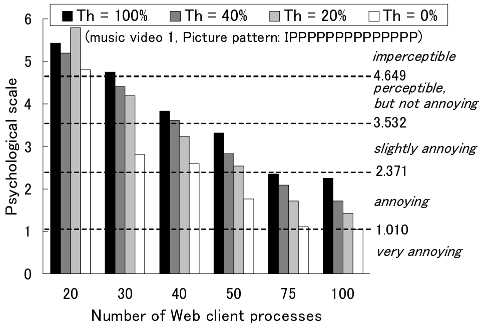

Fig. 15. Psychological scale versus number of Web client processes (music video 1, Picture pattern: IPPPPPPPPPPPPPP).

We then saw that T_h=0, namely, pure frame skipping, is the most effective for the I picture pattern of all the contents at least under the condition of this paper.

For video–dominant contents whose picture pattern includes P's, setting T_h to nonzero values works well; as the number of P's increases, a larger value of T_h achieves higher QoE.

Also, for audio–dominant contents and animation, nonzero values of T_h begin to become effective when the number of P's in the picture pattern becomes more than those of the video–dominant contents. The improvement with the nonzero value starts with less P's if the video motion is high. When there are many P's in the picture pattern and/or when the video motion is high, $T_h = 100\%$, which means the pure error concealment, is the best.

The above observations suggest several possible ways of setting T_h. Below, we propose a potential way of doing this.

If the picture pattern consists of only I, we set T_h to 0. Otherwise (i.e., in the case of including P's), we first set T_h to 100 temporarily as the initial working value and decode all received frames during a certain period of time, which is referred to as the *learning period*; meanwhile, we prepare a few values of T_h (e.g., 100, 40, 20 and 0) as the candidate of the formal value, which will be used after the period.

During the learning period, not all decoded video frames are output actually; although the SCS with $T_h = 100$ is applied to the output of video–dominant contents, that with $T_h = 0$ is used for audio–dominant contents and animation. While outputting the video and audio, we estimate the psychological scale for each of the candidate T_h values by means of a real–time estimation method such as that in [8], compare the estimates obtained at the end of the learning period and then select the T_h value that provides the maximum.

We can repeat the above process after the learning period in order to adapt to time–varying traffic.

The quantitative validation of the proposed way is now under investigation; an experimental result can be found in [19].

V. Conclusions

This paper proposed a methodology of video–stream output for QoE enhancement, which utilizes the QoE tradeoff relation between spatial and temporal quality caused by error concealment and frame skipping. In order to study the effectiveness of this methodology, we presented a simple scheme called SCS and made experiments on the scheme using six contents each with three picture patterns. The experimental result showed that the QoE tradeoff relation really exists and that the SCS achieves high QoE when the threshold value T_h is chosen appropriately. We also proposed a way of setting T_h on the basis of QoE real–time estimation.

Since this paper is a first step to the study on the methodology, we set the main purpose to showing the existence of the QoE tradeoff relation and the feasibility of QoE enhancement by the relation. Therefore, we chose a straightforward scheme of switching between a simple error concealment method and frame skipping, the SCS. Future work includes the quantitative validation of the proposed way of the threshold setting for the SCS and devising sophisticated schemes of mingling more advanced error concealment methods with frame skipping.

Acknowledgment

The authors thank Dr. Toshiro Nunome for his discussions.

References

[1] ITU-T Rec. G.100/P.10 Amendment 1, "Amendment 1: new appendix I definition of Quality of Experience (QoE)," Jan. 2007.

[2] C.–S. Lee and D. Knight, "Realization of the next–generation network," *IEEE Commun. Mag.*, vol.43 No.10 pp.34–41,Oct. 2005.

[3] S. Tasaka and Y. Ishibashi, "Mutually compensatory property of multimedia QoS," in *Conf. Rec. IEEE ICC2002*, pp. 1105–1111, Apr./May 2002.

[4] L. Atzori, F. G. B. De Natale, and C. Perra, "A spatio-temporal concealment technique using boundary matching algorithm and mesh–based warping (BMA–MBW)," *IEEE Trans. Multimedia*, vol.3,no.3,pp.326–338, Sep. 2001.

[5] S. Belfiore, M. Grangetto, E. Magli, and G. Olmo, "spatio-temporal video concealment with perceptually optimized mode selection," in *Proc. IEEE ICASSP*, Apr. 2003.

[6] S. Tasaka, J. Sako and Y. Ito, "Enhancement of user–level QoS in audio–video IP transmission by utilizing the mutually compensatory property," in *Conf. Rec. IEEE GLOBECOM2006*, Nov. 2006.

[7] ITU-T Rec. J.148, "Requirements for an objective perceptual multimedia quality model," May 2003.

[8] S. Tasaka and Y. Watanabe, "Real–time estimation of user–level QoS in audio video IP transmission by using temporal and spatial quality," in *Conf. Rec. IEEE GLOBECOM2007*, Nov. 2007.

[9] S. Tasaka and Y. Ito, "Psychometric analysis of the mutually compensatory property of multimedia QoS," in *Conf. Rec. IEEE ICC2003*, pp. 1880–1886, May 2003.

[10] J. P. Guilford, *Psychometric methods*, McGraw-Hill, N. Y., 1954.

[11] J. C. Nunnally and I. H. Bernstein, *Psychometric theory, Third edition*, McGraw-Hill, N. Y., 1994.

[12] Y. Ito and S. Tasaka, "Quantitative assessment of user–level QoS and its mapping", *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 572–584, June 2005.

[13] S. Tasaka, Y. Ito, H. Yamada and J. Sako, "A method of user–level QoS guarantee by session control in audio–video transmission over IP networks," in *Conf. Rec. IEEE GLOBECOM2006*, Nov. 2006.

[14] Mindcraft Inc., "WebStone benchmark information," http://www.mindcraft.com/webstone/.

[15] "H.264/MPEG-4 AVC reference software JM11.0," http://iphome.hhi.de/suehring/tml/index.htm.

[16] The Video Quality Experts Group, "Multimedia group test plan, draft version 1.19," http://www.its.bldrdoc.gov/vqeg/.

[17] ITU-T Rec. P.911, "Subjective audiovisual quality assessment methods for multimedia applications," Dec. 1998.

[18] F. Mosteller, "Remarks on the method of paired comparisons: III a test of significance for paired comparisons when equal standard deviations and equal correlations are assumed," *Psychometrika*, vol. 16, no. 2, pp. 207–218, June 1951.

[19] S. Tasaka, H. Yoshimi, A. Hirashima and T. Nunome, "The effectiveness of a QoE–based video output scheme for audio–video IP transmission," in *Proc. ACM Multimedia2008*, Oct. 2008.