The Effectiveness of a QoE–Based Video Output Scheme for Audio–Video IP Transmission

Shuji Tasaka
tasaka@nitech.ac.jp

Hikaru Yoshimi
pikkuru@inl.nitech.ac.jp

Akifumi Hirashima
tiery@inl.nitech.ac.jp

Toshiro Nunome
nunome@nitech.ac.jp

Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology, Nagoya 466-8555, Japan

ABSTRACT

This paper shows the effectiveness and feasibility of a video output scheme the authors proposed previously to enhance QoE (Quality of Experience) (i.e., perceptual QoS) of audio–video IP transmission. The scheme, SCS (Switching between error Concealment and frame Skipping), utilizes a trade-off relation between spatial and temporal quality caused by video error concealment and video frame skipping, both of which cope with video packet loss. The scheme switches from error concealment to frame skipping when the percentage of video slices error–concealed in a frame exceeds a threshold value, which is a key to achieving high QoE with SCS. We propose a way of selecting appropriate threshold values on the basis of real-time estimation of QoE. For that purpose, we performed two experiments for four values of the threshold. One experiment is the measurement of application–level QoS and QoE, from which we confirm the effectiveness of SCS and derive multiple regression lines that estimate QoE from the application–level QoS. The other is the measurement of the percentage of the selected threshold value by the proposed way with the multiple regression lines. Examining the QoE measured by the first experiment when we adopt the selected threshold values in the second experiment, we find that the way gives appropriate selections in many cases and therefore the SCS with the way of threshold selection is feasible.

Categories and Subject Descriptors

H.5 [Information Interfaces and Presentation]: Multimedia Information System—Evaluation/methodology, Audio, Video

General Terms

Measurement, Performance, Experimentation, Human Factors

Keywords

audio–video IP transmission, QoE, QoS, perceptual QoS, video error concealment, video frame skipping

1. INTRODUCTION

Audio–video IP transmission forms an indispensable basis of networked multimedia applications. One of the most important issues related to this technology is how to guarantee or enhance the output quality of audio–video streams, which suffer from impairment typified by packet loss, error and delay in the IP networks.

Various techniques to cope with the impairment problem have been developed; they include error–resilient coding, error–concealment [1], and video frame skipping as well as network QoS (Quality of Service) control such as IntServ [2] and DiffServ [3]. Among them, video error–concealment and frame skipping at the receiver are very popular ones, since these two techniques can be used at the discretion of the receiver independently of the network and are considered effective in improving the audio–video output quality. 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 technique of video frame skipping at the receiver does not decode a frame unless all packets of the frame are correctly received.

Now let us give a consideration to the improvement of quality (i.e., QoS) achieved by video error–concealment and frame skipping. First of all, we should realize that the two techniques improve application–level QoS1 and that its improvement leads to better human perception of the output streams, i.e., improvement of user–level QoS, which is also referred to as QoE (Quality of Experience) [7] in ITU–T.

Inspecting the video error–concealment and frame skipping from an application–level QoS point of view, we can easily notice an interesting property of the techniques: a

1In 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 [6].
2. RELATED WORK

Before introducing the SCS, we describe related work which is useful for understanding of the background of this study. We first point out four limitations in previous studies on video IP transmission. We next give a brief description of a method for measuring QoE and a method for the estimation.

2.1 Limitations in previous studies on IP video

In the context of QoE research on video IP transmission, we notice four limitations in the previous studies.

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 [9]; ITU-T has paid much attention to this cross-modal interaction and has established ITU-T Recommendation J.148 [10]. 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 prepares the play-out 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; in interactive applications, we are forced to restrict the buffering time to shorter one than that in the streaming services in order to achieve high QoE [11], [12]. 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.

The fourth limitation is their inapplicability to real-time estimation of QoE [13]. We can find many methods of assessing QoE of audio and/or video transmission in the literature. Typical examples are ITU-T and ITU-R recommendations, which are based on the measurement of subjective quality by using human assessors. In addition to these subjective methods, some objective methods are prepared. The great majority of them are Full Reference (FR) models, which estimate the QoE by comparing the stream to be assessed with the original stream, typically by calculating PSNR; however, this is impossible in real time since no original signal is available at the receiver in real time.

In [13], the authors have presented a methodology for avoiding the four limitations. This paper deals with the QoE problem by the methodology.

2.2 QoE measurement

In this paper, we express QoE in terms of the interval scale, which is referred to as the psychological scale [14]; we do not adopt MOS (Mean Opinion Score), which is the QoE metric mainly used in ITU-T/R recommendations and many of technical papers. This is because the psychological scale can represent the human subjectivity more accurately than MOS. The interval scale can be calculated by one of the psychometric methods [15], [16].

For the calculation of the interval scale, as in [14] and [17], 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\(^2\).

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 [15], [19]. Once the goodness of fit has been confirmed, we use the interval scale as the QoE parameter, which is therefore called the psychological scale.

2.3 QoE estimation

As in [14], [17] and [18], this paper estimates the psychological scale by means of QoS mapping between user–level and application–level. We perform the QoS mapping with multiple regression analysis [16] by defining the psychological scale as the dependent variable. As the independent variables, we employ application–level QoS parameters representing temporal and spatial quality, which can highly correlate with each other. This requires us to select appropriate independent variables with low cross–correlations from among the introduced variables.

Principal component analysis helps us find the correlations between the introduced independent variables. We first compute the principal component loadings of each variable up to the principal component that provides a large cumulative contribution rate (e.g., over 90%). On the basis of the principal component loadings, we classify the introduced variables into a certain number of classes.

We then pick up one variable from each class and calculate a multiple regression line for every combination of the variables picked up. From among the multiple regression lines thus calculated, we finally select one that achieves the largest value of the contribution rate adjusted for degrees of freedom or its square root (multiple correlation coefficient adjusted for degrees of freedom), which indicates goodness of fit of estimates to the corresponding measured values.

3. SCS AND ITS QOE

This section first gives a brief description of the SCS. It next presents an experimental network along with the contents to be assessed, and methods of measuring application–level QoS and QoE. We then show a few experimental results of QoE.

3.1 Principle

The SCS is a simple example of the methodology of video–stream output utilizing the QoE tradeoff relation, which the authors propose in [8]. It mingles video error concealment and frame skipping at the receiver so that it can improve QoE over the employment of either technique.

The mode of video output in SCS 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 \( T_h \); the frame skipping continues until an intra–coded frame is decoded, at which time the mode switches back to the error concealment.

3.2 Experimental network and contents

3.2.1 Network configuration

Figure 1 shows the configuration of the experimental network; it consists of two routers 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.

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 MUs 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 MUs.

Table 1 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 P’s, I followed by B’s, and B followed by B’s, to examine their effects on the QoE tradeoff relation.

As in [20], 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 [21]. WebStone generates Web client processes on the WC PC; those client processes retrieve specified files from the WS PC continuously. Table 2 shows the set of files to be retrieved. In the experiment, the

---

\(^{2}\)In the case of MOS, we simply take an average of the measured scores for a stimulus over all assessors. However, it should be noted that this method of the calculation 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 [14], [17] and [18]. 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.

---

**Table 1: Specifications of audio and video**

<table>
<thead>
<tr>
<th>audio MU size [byte]</th>
<th>video MU size [byte]</th>
<th>image size [pixel]</th>
<th>number of slices in a picture</th>
<th>video average MU rate [MU/s]</th>
<th>picture pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>audio average bit rate [kb/s]</td>
<td>video average bit rate [kb/s]</td>
<td>resolution [pixel]</td>
<td>(interleave mode)</td>
<td>(20 macroblocks/slice)</td>
<td>( R_C )</td>
</tr>
</tbody>
</table>
Table 2: The set of files to be retrieved from the web server

<table>
<thead>
<tr>
<th>file name</th>
<th>size [kbyte]</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>file500.html</td>
<td>0.5</td>
<td>0.350</td>
</tr>
<tr>
<td>file5k.html</td>
<td>5.0</td>
<td>0.500</td>
</tr>
<tr>
<td>file50k.html</td>
<td>50.0</td>
<td>0.140</td>
</tr>
<tr>
<td>file500k.html</td>
<td>500.0</td>
<td>0.009</td>
</tr>
<tr>
<td>file5m.html</td>
<td>5000.0</td>
<td>0.001</td>
</tr>
</tbody>
</table>

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 [22]. For I-frames, we utilize the spatial approach: A missing block is interpolated from its neighboring blocks in the current frame. 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 previous 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 \( T_b \) to 100%, 40%, 20% and 0%.

3.2.2 Contents

Referring to the VQEG multimedia test plan [23], 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 (S1):** A group of people are doing aerobics, timing their movement to the instructor’s voice and music. There is no scene change.
- **sport 2 (S2):** 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 (A1):** Two human characters are talking, while one is shouting, without background music (five scene changes).
- **animation 2 (A2):** A robot is trying to rescue a shouting boy from falling in the sky with fast background music (frequent scene changes).
- **music video 1 (M1):** A sitting young male is playing the ukelele without scene change. The audio is the ukelele’s sound only.
- **music video 2 (M2):** 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 3: Video average bit rate and TI value for each content

<table>
<thead>
<tr>
<th>content</th>
<th>average bit rate [kb/s]</th>
<th>TI value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>1440.342</td>
<td>356.598</td>
</tr>
<tr>
<td>I</td>
<td>1890.361</td>
<td>518.262</td>
</tr>
<tr>
<td>P</td>
<td>2241.426</td>
<td>618.478</td>
</tr>
<tr>
<td>14P</td>
<td>2592.713</td>
<td>864.747</td>
</tr>
<tr>
<td>14P’</td>
<td>3043.759</td>
<td>1121.959</td>
</tr>
</tbody>
</table>

Table 3 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 [24]; 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.

3.3 Measurement of application-level QoS

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 experiment run, the media recipient (MR) measured application–level QoS parameters which are listed in Table 4.

The nine parameters given in the first five rows mainly represent the temporal quality; in particular, the coefficient of variation of MU output interval is defined as the ratio of the standard deviation of the MU output interval to its average; therefore, it represents the smoothness of the output stream. The MU loss ratio is the ratio of the number of MUs not output at the recipient to the number of MUs transmitted by the sender. Also, the mean square error (MSE) \( E_{int} \) is an indicator of differential delay between audio and video, i.e., skew of lip-sync.

On the other hand, the parameter \( R_c \) is a video spatial quality metric; note that this corresponds to a No Reference (NR) model of perceptual video quality estimation, which needs no information on the original video stream.

We easily see that all the application–level QoS parameters above are automatically measurable; therefore, we can use them to estimate the psychological scale in real time.

3.4 Subjective measurement of QoE

By the rating–scale method, we collected data for calculating QoE.

We first made stimuli for subjective experiment as follows. During each experimental run, we recorded the audio–video
streams output by the media recipient (MR); the recorded streams are regarded as stimuli for QoE measurement. Thus, we totally had 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.

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.

4. QOE ASSESSMENT AND ESTIMATION

This section first assesses QoE by calculating the psychological scale from the measurement result in Subsection 3.3 and show how effective the SCS is. We then derive estimates of QoE by QoS mapping from application–level QoS measured in Subsection 3.3 and examine the accuracy of the estimation.

4.1 Measurement result of QoE

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.

Figure 2 plots the psychological scale as a function of the number of Web client processes and \( T_h \) (sport 2, Picture pattern: I).

![Psychological scale as a function of the number of Web client processes and \( T_h \) (sport 2, Picture pattern: I).](image)

Figure 3: Psychological scale as a function of the number of Web client processes and \( T_h \) (sport 2, Picture pattern: IPPPP).

![Psychological scale as a function of the number of Web client processes and \( T_h \) (sport 2, Picture pattern: IPPPP).](image)

Figure 4: Psychological scale as a function of the number of Web client processes and \( T_h \) (animation 2, Picture pattern: IPPPP).

Figure 5: Psychological scale as a function of the number of Web client processes and \( T_h \) (music 1, Picture pattern: IPPPP).

3When the number of Web client processes is 20, packet loss scarcely occurred.
QoE is given in [8].

1, Picture pattern: IPPPPP).

the condition that one parameter is selected from one class. We tried to select a combination for each content so that it can make the value of the multiple correlation coefficient adjusted for degrees of freedom as large as possible. As a result, we found the following multiple regression lines; in the equations, the estimate of the psychological scale is represented by \( \hat{U}_{CT} \), where the subscript CT means the content type, and \( R^* \) denotes the multiple correlation coefficient adjusted for degrees of freedom.

\[
\begin{align*}
\hat{U}_{S1}^1 &= 4.288 - 0.478 R_c - 0.066 L_v \quad (R^* = 0.879) \quad (1) \\
\hat{U}_{S2}^1 &= 4.038 - 0.470 R_c - 0.0734 L_v \quad (R^* = 0.878) \quad (2) \\
\hat{U}_{A1}^1 &= 4.211 - 0.613 R_c - 0.055 L_v \quad (R^* = 0.832) \quad (3) \\
\hat{U}_{A2}^1 &= 4.490 - 0.652 R_c - 0.0741 L_v \quad (R^* = 0.882) \quad (4) \\
\hat{U}_{M1}^1 &= 4.366 - 0.723 R_c - 0.0654 L_v \quad (R^* = 0.851) \quad (5) \\
\hat{U}_{M2}^1 &= 4.280 - 0.458 R_c - 0.058 L_v - 0.488 L_a \quad (R^* = 0.899) \quad (6)
\end{align*}
\]

We next performed nonlinear multiple regression analysis based on cubic polynomials [23]. Below, we show the obtained multiple regression lines, where \( \hat{U}_{CT}^3 \) represents the estimate of the psychological scale and the subscript CT means the content type as earlier.

\[
\begin{align*}
\hat{U}_{S1}^3 &= 4.732 - 1.517 R_c + 3.423 \times 10^{-1} R_c^2 \\
&\quad - 2.689 \times 10^{-2} R_c^3 - 1.111 \times 10^{-1} L_v \\
&\quad + 1.718 \times 10^{-3} L_v^2 - 1.519 \times 10^{-5} L_v^3 \quad (R^* = 0.943) \quad (7) \\
\hat{U}_{S2}^3 &= 4.492 - 1.363 R_c + 2.930 \times 10^{-1} R_c^2 \\
&\quad - 2.296 \times 10^{-2} R_c^3 - 1.317 \times 10^{-1} L_v \\
&\quad + 1.699 \times 10^{-3} L_v^2 - 8.607 \times 10^{-6} L_v^3 \quad (R^* = 0.940) \quad (8) \\
\hat{U}_{A1}^3 &= 4.564 - 1.883 R_c + 4.207 \times 10^{-1} R_c^2 \\
&\quad - 3.996 \times 10^{-2} R_c^3 - 4.752 \times 10^{-2} L_v \\
&\quad - 6.147 \times 10^{-4} L_v^2 + 9.779 \times 10^{-6} L_v^3 \quad (R^* = 0.927) \quad (9) \\
\hat{U}_{A2}^3 &= 4.905 - 2.232 R_c + 7.025 \times 10^{-1} R_c^2 \\
&\quad - 7.454 \times 10^{-2} R_c^3 - 1.243 \times 10^{-1} L_v \\
&\quad + 2.679 \times 10^{-3} L_v^2 - 3.364 \times 10^{-5} L_v^3 \quad (R^* = 0.951) \quad (10) \\
\hat{U}_{M1}^3 &= 4.826 - 2.372 R_c + 7.479 \times 10^{-1} R_c^2 \\
&\quad - 8.025 \times 10^{-2} R_c^3 - 1.074 \times 10^{-1} L_v \\
&\quad + 2.106 \times 10^{-3} L_v^2 - 2.499 \times 10^{-5} L_v^3 \quad (R^* = 0.943) \quad (11)
\end{align*}
\]

Therefore the effect of the decrease in the output frame rate on human perception is small, while music video 1 is an audio-dominant content with low video motion.

In Fig. 6, we show the psychological scale for music video 1 with picture pattern IPPPPP. Although this is a result for an audio-dominant content, a larger value of \( T_h \) namely more error concealment provides higher QoE in the lossy environments. This is because a frame skip incurs skipping all succeeding P frames, which degrades QoE.

More detailed consideration to the measurement result of QoE is given in [8].

### 4.2 Estimation of QoE

In order to estimate the psychological scale, we resort to multiple regression analysis according to the procedure described in Subsection 2.3. We perform both linear analysis and nonlinear one based on cubic polynomials. In the derivation for each content, we used measurement results of the application-level QoS parameters and the QoE parameter for the three picture patterns all together.

First, principal component analysis provided the principal component loadings of each variable, which classified the application-level QoS parameters (i.e., the independent variables) into five classes as shown in Table 5.

We first performed linear multiple regression analysis of all combinations of the application-level QoS parameters under...
This is for future study.

Therefore, deriving multiple regression lines for each picture and 6, the picture patterns affect the QoE tradeoff relation.

As we have found in Figs. 5 and 6, the picture patterns affect the QoE tradeoff relation. We now examine the accuracy of the estimate with the corresponding measured values. Because of space limitations, we show the results only for sport 2 and music video 1, which are video–dominant and audio–dominant, respectively.

Figures 7 through 10 plot the estimated values along with the measured ones as a function of the number of Web client processes in the case of sport 2 with picture pattern I for $T_h = 100\%$, 40\%, 20\% and 0\%, respectively. In these figures, we find that the psychological scale can be estimated with good accuracy by means of the nonlinear multiple regression line, i.e., Eq. (8).

Figures 11 through 16 display the case of picture pattern IPPPP in sport 2 and music video 1. In each figure, we also see that the estimated value approximately fits the measured one.

Figures 17 and 18 present the case of picture pattern IPPPPPPPPPPPPP in music video 1. In these figures, we notice some discrepancies between the measured value and the estimated one; the regression line, i.e., Eq. (11), is not so accurate in this case. As we have found in Figs. 5 and 6, the picture patterns affect the QoE tradeoff relation. Therefore, deriving multiple regression lines for each picture pattern separately can give higher accuracy of the estimation; this is for future study.

5. A WAY OF SETTING THE THRESHOLD

As seen from the observation so far, we have noticed that a key to achieving high QoE by SCS is how to select the threshold value $T_h$. The appropriate value depends on the content type, picture pattern, and degree of video motion.

\[
\hat{U}_{M2} = 4.638 - 1.630R_c + 3.910 \times 10^{-1}R_c^2
\]

\[-3.280 \times 10^{-2}R_c^3 - 1.245 \times 10^{-1}L_v
\]

\[+ 2.587 \times 10^{-3}L_v^2 - 2.823 \times 10^{-5}L_v^3
\]

\[-1.487 \times 10^{-1}L_a
\]

\[(R^2 = 0.940) (12)\]

Comparing the linear and nonlinear equations thus obtained, we see that the nonlinear equation provides a larger $R^2$ than the linear one for every content. Consequently, we utilize only the nonlinear equations in the following discussion.

4.3 Accuracy of estimation

We now examine the accuracy of the estimate with the nonlinear multiple regression lines by comparing them to the corresponding measured values. On the basis of the observations we have made in Subsection 4.1, the authors have proposed a way of setting $T_h$ by utilizing the estimate of QoE in [8].

In this paper, we slightly modify the way in [8] for simplicity of experiment. We give a description of the modified way in the following. 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

On the basis of the observations we have made in Subsection 4.1, the authors have proposed a way of setting $T_h$ by utilizing the estimate of QoE in [8].
period, while outputting the video and audio, we estimate the psychological scale for each of the candidate $T_h$ values in real time by means of the multiple regression lines given in the previous section. We compare the estimates obtained at the end of the learning period and then select the $T_h$ value that provides the maximum. If more than one $T_h$ takes the same maximum value of the estimate, we select the largest $T_h$ among them.

In [8], the quantitative validation of the proposed way was left as future work. Since we now have the multiple regression lines for real-time estimation in hand, we are ready to evaluate the proposed way, though it has been slightly modified.

We have conducted experiments on the way of selecting the threshold value under the same condition as that of Section 3 except for the duration (i.e., recording time) of the contents: it is 60 seconds instead of 10 seconds. The learning period was set to 10 seconds. For each content with a picture pattern at a constant level of the average Web traffic, we did the experiment 15 times and measured the percentage of the selected threshold value.
We show the measurement results in Figs. 19 through 22, which plot the percentage of the selected threshold value versus the number of Web client processes for sport 2 and music video 1.

Figure 19 indicates that $T_h = 0\%$ (namely, pure frame skipping) has always been selected when the number of Web client processes is more than 20 (namely, in lossy environments). From Fig. 2, we can confirm that this is the best selection. When the number of Web client processes is 20, over 90% of the selected value is $T_h = 100\%$. This is because packet loss scarcely occurred, and therefore all the $T_h$ values often have the same estimate, which leads to the selection of $T_h = 100\%$; this is also the case with Figs. 20 through 22.

In Figs. 20 and 22, we notice that nonzero values of $T_h$ were selected frequently; this is suitable selection as Figs. 3 and 6 reveal.

Also, in Fig. 21, we find that the great majority of the selection is $T_h = 0\%$, which is the best in lossy environments as seen from Fig. 5.

Thus, we can confirm that the proposed way of selecting the threshold value works well in many cases. However, a few issues are still left as future work. In order to get a better understanding of the effectiveness of the way, we have to measure QoE of the streams output in this way by subjective experiment. In addition, the preparation of multiple regression lines for the estimation is an important issue; the construction of the database of representative regression lines as described in [20] can be a promising approach to the issue. Also, some way of setting $T_h$ value without the estimation is another possibility; for example, we set a certain value according to some feature of the content such as the content type and picture pattern. This is for further study.

6. CONCLUSIONS

We examined the effectiveness of the SCS, an IP video output scheme utilizing the tradeoff relation between spatial and temporal quality caused by error concealment and frame skipping. We made an experiment of audio–video IP transmission with the SCS and measured application–level QoS and QoE.

Using the experimental result, we first confirmed the effectiveness of SCS by examining the measured QoE and noticed
that the selection of the threshold value for the switching is a key to successful implementation of SCS. We then estimated QoE from the measured application-level QoS by means of both linear and nonlinear multiple regression analysis. We found that the multiple regression lines obtained provide accurate estimates.

We further conducted an experiment on a way of selecting the threshold value. The experiment showed that the way gives appropriate selections in many cases.

In order to make the SCS more practical and effective, we have several problems to be solved. How to select the threshold value is still an important issue as commented at the end of the previous section. Also, we should try more advanced error concealment methods instead of “Frame Copy” in this paper.

7. ACKNOWLEDGMENTS
This work was supported by The Telecommunications Advancement Foundation.

8. REFERENCES