Real–Time Estimation of User–Level QoS in Audio–Video IP Transmission by Using Temporal and Spatial Quality

Shuji Tasaka and Yusuke Watanabe
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 deals with audio–video transmission over IP networks and proposes a method of estimating perceptual QoS (i.e., user–level QoS) in real time in terms of the interval scale, which is referred to as the psychological scale. The proposed scheme utilizes both temporal and No Reference spatial (picture's) quality, which are application–level QoS, and estimates the user–level QoS from the application–level QoS. To evaluate the effectiveness of the proposed scheme, we compare the proposed scheme with a scheme which uses only temporal quality for the estimation and a scheme with both temporal and Full Reference spatial quality. We made a simple experiment of audio–video IP transmission; we measured the application–level QoS directly and assessed the user–level QoS by the method of successive categories. Applying the principal component analysis and multiple regression analysis to the experimental results, we obtained multiple regression lines (i.e., equations for the estimation). As a result, we see that the user–level QoS can be estimated with high accuracy by using both temporal quality and spatial quality.**

I. Introduction

The audio–video transmission is an essential ingredient of multimedia application services in the IP networks, which are becoming an information infrastructure in the form of the *Next Generation Network* (*NGN*) [1] in addition to the current Internet.

In networked multimedia applications, the ultimate goal of the service is to provide perceived quality that satisfies the users [2]. In order to realize it, we have to perform some control and management of *QoS* (*Quality–of–Service*). This needs real–time assessment (monitoring) of QoS, in particular, *perceptual QoS*, which corresponds to *user-level QoS* in the context of the layered network architecture [1]. However, note that real-time measurement of user-level QoS is practically impossible, since the network operator cannot ask the users to report their perceptual quality in real time. This leads to an increasing demand for methods of estimating user-level QoS by using automatically measurable lower-level QoS parameters such as packet loss ratio and delay jitter.

In the literature, we can find various methods of assessing (i.e., measuring and/or estimating) subjective quality (i.e., user–level QoS) in audio and/or video transmission. Among them, the ones recommended by ITU–T and ITU–R are well–known. In these methods, however, we notice two main limitations in applying them to real-time estimation of user-level QoS in audiovisual transmission. One is that the great majority of the methods deals with only a single medium; either video only or audio only. The other is their inapplicability to the real–time estimation.

[1]In IP networks, six kinds of QoS are identified along the protocol stack: *physical–level*, *node–level*, *network–level*, *end–to–end–level*, *application–level*, and *user–level* [3]. QoS at a level is quantified by its *QoS parameters*.

Let us examine the two limitations in more detail below. First, we focus on video quality assessment. ITU–R Recommendation BT.500 [4] and ITU–T Rec. P. 910 [5] are typical examples of the method; the former deals with television pictures, and the latter is for digital video images. Both recommendations specify methods of *measuring* subjective quality by using human observers; therefore, they are not applicable to the *estimation* we need.

According to ITU–T Rec. J. 143 [6], estimation methods of perceptual video quality are classified into three models: *Full Reference* (*FR*), *Reduced Reference* (*RR*), and *No Reference* (*NR*). The FR model estimates the perceptual video quality by comparing the video stream to be assessed with the original stream; therefore, this model cannot be utilized for real–time estimation of the perceptual quality of received video streams, since no original signal is available at the receiver in real time. The RR model also uses some information on features of the original stream in the estimation, while the NR model needs no information on the original stream. Thus, the NR model is the most suitable for our purpose.

The *VQEG* (*Video Quality Experts Group*) [7] has been actively studying methods of predicting (i.e., estimating) perceptual video quality in addition to measuring it; it assesses the accuracy of proposed prediction methods and makes inputs to the creation of the recommendations of video quality assessment methods. The VQEG has provided the final report on the FR models in [8]. The RR and NR models are now under investigation.

We next turn our attention to audio quality assessment, for which ITU–T Rec. P.862 [9] and G. 107 [10] are often utilized. P.862 provides the *PESQ* score; this produces *MOS* (*Mean Opinion Score*), which is a user–level QoS parameter. The calculation of the PESQ score needs both received signal and original signal; this implies that PESQ is a kind of FR model. Also, G.107 defines the *E–model* for assessment of the subjective quality of VoIP; it gives the *R–factor*, which is convertible to MOS. However, the E–model is a network planning tool and is not used for the real–time estimation during the network operation.

The methods mentioned so far are intended for only a single medium, namely, either video only or audio only. Regarding this limitation, it should be noted that audiovisual transmission is featured by cross–modal influences of audio and video as studied in [11], [12], [13] and [14]. We need some methods which take the cross–modal influences into consideration. The approaches in the above papers, however, are not directly applicable to the real–time estimation.

There also exist ITU–T Recommendations for subjective quality assessment of both audio and video together, including P.911 [15] and J.148 [16]. However, P.911 cannot be used in the real–time estimation since it presents subjective quality measurement methods similar to those of P.910. Rec. J. 148 details the requirements for the development of an objective auditory–visual perceptual quality model taking into consideration the cross–modal influences; this is the first ITU–T Recommendation that clearly states the importance of the treatment of both audio and video in perceptual quality as-

sessment. However, J.148 has established only a basic model and does not provide the details of the assessment procedure.

In order to give a solution to the problem of real–time estimation of user–level QoS in audio–video IP transmission, Tasaka and Ito proposed a method in [17]. The method is based on QoS parameter mapping between application–level and user–level; the estimation scheme utilizes multiple regression lines that predict user–level QoS parameter values from application–level QoS parameter values. The reason why application–level QoS parameters has been selected as the independent variables for the estimation is that the application–level QoS can represent the temporal structures of audio and video as well as the video spatial (i.e., picture's) structure [2].

The application–level QoS parameters used in [17] are ones concerning the quality of *media synchronization* [3], which represents *temporal quality* of received audio–video streams in units of *MU* (*Media Unit*) [3]. Those QoS parameters are automatically measurable in real time. For simplicity of implementation, the method in [17] does not utilize any *spatial quality* (i.e., picture quality) of video for the estimation. That is, the receiver does not output a video MU unless all packets of the MU are received correctly; this implies no degradation of output picture quality, though it incurs the degradation of the temporal quality as video MU skipping. Quality estimation in the case of incomplete video MU output was left as future work.

The current paper improves the method in [17] by taking into consideration spatial quality of video in addition to the temporal quality. Spatial quality of video at the application–level is often assessed by means of some FR model, typically in terms of the *PSNR* (*Peak Signal to Noise Ratio*) or equivalently *MSE* (*Mean Square Error*) of a picture signal; note that its calculation needs not only a received picture but also the corresponding original picture. The method in the current paper utilizes a metric of video spatial quality based on an NR model proposed in [18]

The remainder of the paper is organized as follows. Section II introduces methods for the video spatial quality evaluation. Section III gives a brief description of a method for measuring user–level QoS and a method for the estimation. Section IV demonstrates an experimental methodology for measuring the application–level and user–level QoS. Section V presents experimental results and examines the accuracy of the proposed method. Section VI concludes the paper.

II. Methods for Video Spatial Quality Evaluation

In this paper, we utilize an NR method proposed in [18] to evaluate the degradation of video spatial quality due to *packet loss*. We have decided to employ this method since it has low computational complexity and therefore can be used for real–time monitoring of streaming video; this meets our requirement for the real–time estimation.

In addition, we calculate MSE of video luminance, though it is a metric of the FR model; this is just for comparison purposes in order to examine how effective the NR method is.

Let us give an outline of the NR method. The method defines a metric that expresses the quality degradation due to packet loss. The video decoder is supposed to use a simple replacement algorithm for error concealment; a lost macroblock is replaced by the corresponding macroblock from the previous frame. Therefore, when packet loss occurs, the edges of the replaced macroblock often become different from the ones of the adjacent macroblocks. Thus, the method exploits the structure of the artifact across the macroblock boundaries. The basic idea of the metric is to measure the difference in the edge strength between the one across the two adjacent macroblocks and the one within a macroblock. The method produces the metric by summing up the differences for all macroblocks of a video frame. In this paper, we denote this metric by S_{NR}.

III. Methods for User–Level QoS Measurement and Estimation

A. User–level QoS measurement

In this paper, we express user–level QoS in terms of a QoS parameter of the *interval scale*, which is referred to as the *psychological scale* [19]; we do not adopt MOS, which is the user–level QoS parameter 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 [20], [21].

For the calculation of the interval scale, as in [17] and [19], 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 (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 [4].

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* [20], [23]. Once the goodness of fit has been confirmed, we use the interval scale as the user–level QoS parameter, which is therefore called the psychological scale.

B. User–level QoS estimation

As in [17], [19] and [22], 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* [21] 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*, which indicates goodness of fit of estimates to the corresponding measured values.

[2]Note that QoS parameters at the end–to–end and lower levels do not reflect the media types treated because of the principle of the layered network architecture.

[3]Media synchronization can be classified into *intra–stream synchronization* and *inter–stream synchronization*. The former keeps the continuity of a single stream (audio or video), while the latter is synchronization between audio and video streams. A video MU is usually defined as a video frame and an audio MU as a constant number of audio samples. An MU is usually divided into two or more packets.

[4]In the case of MOS, we simply take an average of the measured integers 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 integer 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 [17], [19] and [22]. 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.

IV. Experimental Methodology

This section explains an experimental methodology to examine the effectiveness of the proposed estimation method. We first present an experimental network along with contents to be assessed. We then introduce application–level QoS parameters and explain how these parameters are employed in deriving multiple regression lines. Furthermore, we describe conditions under which user–level QoS was measured.

A. Experimental network and contents

1) Network configuration: Figure 1 shows the configuration of the experimental network; it consists of two routers (say Routers 1 and 2) 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). Routers 1 and 2 are RiverStone's RS3000. The link between the routers and ones between a router and a PC are all Ethernet channels. The link speed between a PC and a router is 100 Mb/s, while the one between the two routers is 10 Mb/s, which is therefore a bottleneck.

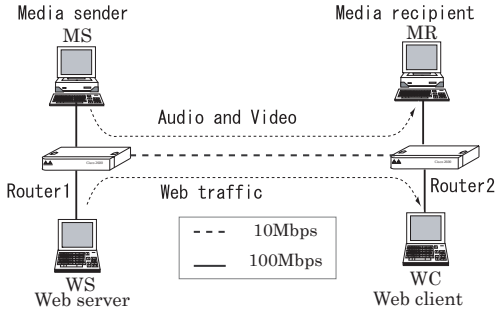

Fig. 1. Configuration of the experimental network

The MS transmits a pair of audio and video streams to the MR with UDP. Table I gives specifications of audio and video used in the experiment. We have set three kinds of picture patterns to examine its effect on the accuracy of the estimation. Since we have kept the average bit rates of audio and video constant regardless of the picture pattern, the spatial (i.e., picture) quality varies slightly from picture pattern to picture pattern. However, the difference in the quality is scarcely noticeable.

TABLE I

Specifications of audio and video

	audio	video
coding scheme	Linear PCM 16bit 48kHz 2ch	MPEG1
image size [pixel]	–	320 × 240
picture pattern	–	I
	–	IPPPPPPPPPPPPPP
	–	IBBPBBPBBPBB--PBBPBB
average MU size [byte]	9600	10400
MU rate [MU/s]	20	30
average bit rate [Mb/s]	1.536	2.5
recording time [s]	15	15

As in [2], the network transfers Web traffic as interference to the audio–video streams according to the configuration of WebStone 2.5, which is a Web server evaluation tool [24]. WebStone generates Web client processes on the WC PC; those client processes retrieve specified files from the WS PC continuously. Table II shows the set of files to be retrieved in our experiment.

The Web traffic competes with the audio-video stream for the link capacity; this causes packet loss of audio and video. When a video packet is lost, we resort to the same method for error concealment as that of [18]: the macroblock containing the lost video packet is replaced by the corresponding macroblock from the previous frame.

TABLE II

The set of files to be retrieved from the Web server

file name	size [kbyte]	probability
`file500.html`	0.5	0.350
`file5k.html`	5.0	0.500
`file50k.html`	50.0	0.140
`file500k.html`	500.0	0.009
`file5m.html`	5000.0	0.001

2) Contents: Referring to the VQEG test plan [7] as in [2], we have selected four types of contents: *sport, music video, movie* and *animation*. As a video–dominant content type, where video plays a more important role than audio, we have adopted *sport*, while we consider *music video* audio–dominant. As for *movie*, we regard both audio and video as important; *animation* belongs to the same type, but its video property is different from that of movie.

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

- *sport 1* and *sport 2*: Scenes of ice skating by a male skater with background music, and a clip of a soccer game, respectively. The audio includes a commentator's voice and spectators' cheers.
- *music video 1* and *music video 2*: Scenes of a male and a female singing and dancing, and scenes of male duo singing and dancing, respectively.
- *movie 1* and *movie 2*: In the former, a couple are talking at an airport check–in counter. In the latter, a couple on a date are talking at a restaurant.
- *animation 1* and *animation 2*: Scenes of a conversation between two male characters, and scenes of a young girl speaking to an old male character, respectively.

B. Measurement of application–level QoS

In each experiment, we changed the amount of the Web traffic by increasing the number of the Web client processes from 10 through 28 with an increment of two; this leads to ten levels of the average Web traffic. During each experiment, the media recipient (MR) measured values of application–level QoS parameters which are listed in Table III.

TABLE III

Application–level QoS parameters

application–level QoS parameters	notation for audio	notation for video
average MU rate [MU/s]	R_a	R_v
coefficient of variation of MU output interval	C_a	C_v
average MU delay [ms]	D_a	D_v
MSE of intra–stream synchronization [ms^2]	E_a	E_v
MSE of inter–stream synchronization [ms^2]	E_{int}	
NR metric of video	–	S_{NR}
MSE of video luminance	–	S_{FR}

The nine parameters given in the first five rows are the same as those in [17], which mainly represent the temporal quality[5]; we call these nine ones the *temporal parameters*. The last two parameters, S_{NR} and S_{FR}, are video spatial quality metrics, which correspond to the NR model and FR model, respectively.

In this paper, we consider three methods for the estimation: (1) a method using only the temporal parameters, (2) one with both temporal parameters and S_{NR}, and (3) one with both temporal parameters and S_{FR}, which are referred to as the *T method*, *TN method*, and *TF method*, respectively. Note that the TF method is a kind of FR model and therefore has been adopted just for comparison purposes.

[5]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. Also, E_{int} is an indicator of differential delay between audio and video, i.e., skew of lip–sync.

C. Measurement of user–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 experimental run, we recorded the audio–video streams that the media recipient (MR) output; the recorded streams are regarded as *stimuli* for user–level QoS measurement. Thus, we totally have 240 stimuli because of eight contents, three picture patterns for each content, and ten levels of the average Web traffic.

In the rating–scale method, we utilized the following five categories of impairment: "imperceptible" assigned integer 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 240 stimuli in a random order and presented them to 25 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 90 minutes for an assessor to assess all the stimuli.

V. Experimental Results

This section first presents the measurement result of the psychological scale and then derives its estimate by QoS mapping. We further compare the accuracy of the three estimation methods introduced in Subsection IV-B.

A. Measurement result of user–level QoS

We utilized the method of successive categories to calculate the interval scale from the results obtained in Subsection IV-C. For the comparison of the interval scales for the eight contents on the same basis, we applied the law of categorical judgment to all the classification results of the eight contents together, i.e., the 240 stimuli.

In order to test the goodness of fit of the interval scale, we carried out the Mosteller's test. As a result, we have found that the test with a significance level of 0.05 can reject the hypothesis that the observed value equals the calculated one. We then checked stimuli which give a large error of Mosteller's test to find nine ones. Removing the nine stimuli, we saw that the hypothesis cannot be rejected. Consequently, for the 231 ($= 240 - 9$) 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 231 stimuli to unity (i.e., 1). Under this condition, we also calculated the lower boundaries of the categories and got 6.347 for Category 5, 5.213 for Category 4, 4.082 for Category 3, and 2.751 for Category 2.

B. Estimation of user–level QoS

For the three estimation methods introduced in Subsection IV-B, we have derived a multiple regression line for each content according to the procedure described in Subsection III-B. In the derivation for each content, we used measurement results of the application–level QoS parameters and the user–level QoS 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 eight classes as shown in Table IV

In each method, we then performed multiple regression analysis of all combinations of the application–level QoS parameters under the condition that one parameter is selected from one class. As a result, we found that there is no single combination which achieves very large values of the contribution rate adjusted for degrees of freedom for all the contents. However, it is desirable to adopt a single combination for all the contents in order to facilitate comparison of their multiple regression lines. Thus, we have taken this approach so that the adopted combination can make values of the contribution rate adjusted for degrees of freedom as large as possible.

TABLE IV

Classification of application–level QoS parameters

	(1) T method	(2) TN method	(3) TF method
class	parameters	parameters	parameters
A	R_a,R_v	R_a,R_v	R_a,R_v
B	D_a,D_v	D_a,D_v	D_a,D_v
C	C_a	C_a	C_a
D	C_v	C_v	C_v
E	E_a	E_a	E_a
F	E_v	E_v	E_v
G	E_{int}	E_{int}	E_{int}
H	-	S_{NR}	S_{FR}

The application–level QoS parameters of the adopted combination have been statistically tested whether they make significant contributions to the multiple regression line. We then removed the parameters which do not make any significant contributions and again performed multiple regression analysis.

Below, we show the multiple regression lines thus obtained. Let us represent the estimate of the psychological scale by $\hat{U}^T$, $\hat{U}^{TN}$, or $\hat{U}^{TF}$; the superscripts T, TN and TF imply the T, TN, and TF methods, respectively. Also, let R^{*2} denote the contribution rate adjusted for degrees of freedom. Then, the regression lines for sport 1, movie 1, music video 1 and animation 1, for instance, are given as follows:

- sport 1

$$\hat{U}^T = 18.615 - 0.048D_v - 17.862C_a - 0.026E_v \quad (R^{*2} = 0.891) \tag{1}$$

$$\hat{U}^{TN} = 18.205 - 0.048D_v - 13.334C_a - 0.015E_v - 8.724S_{NR} \quad (R^{*2} = 0.950) \tag{2}$$

$$\hat{U}^{TF} = 15.898 - 0.042D_v - 13.275C_a - 0.016E_v - 0.003S_{FR} \quad (R^{*2} = 0.940) \tag{3}$$

- movie 1

$$\hat{U}^T = 15.646 - 0.031D_v - 13.645C_a - 0.028E_v \quad (R^{*2} = 0.875) \tag{4}$$

$$\hat{U}^{TN} = 18.204 - 0.031D_v - 8.729C_a - 0.015E_v - 13.005S_{NR} \quad (R^{*2} = 0.913) \tag{5}$$

$$\hat{U}^{TF} = 13.538 - 0.029D_v - 9.813C_a - 0.016E_v - 0.009S_{FR} \quad (R^{*2} = 0.914) \tag{6}$$

- music video 1

$$\hat{U}^T = 16.995 - 0.044D_v - 18.588C_a - 0.014E_v \quad (R^{*2} = 0.876) \tag{7}$$

$$\hat{U}^{TN} = 18.935 - 0.035D_v - 19.000C_a - 0.014E_v - 4.485S_{NR} \quad (R^{*2} = 0.897) \tag{8}$$

$$\hat{U}^{TF} = 15.214 - 0.040D_v - 15.289C_a - 0.009E_v - 0.001S_{FR} \quad (R^{*2} = 0.898) \tag{9}$$

- animation 1

$$\hat{U}^T = 15.146 - 0.024D_v - 12.767C_a - 0.037E_v \quad (R^{*2} = 0.849) \tag{10}$$

$$\hat{U}^{TN} = 20.924 - 0.023D_v - 10.790C_a - 0.050E_v - 44.514S_{NR} \quad (R^{*2} = 0.865) \tag{11}$$

$$\hat{U}^{TF} = 13.172 - 0.018D_v - 11.161C_a \quad (12)$$
$$-0.023E_v - 0.013S_{FR} \quad (R^{*2} = 0.904)$$

C. Comparison of the estimation methods

We now examine how accurately the three estimation methods predict the user–level QoS, namely, the psychological scale. Figures 2 through 4 plot the three kinds of estimated values along with the measured ones as a function of the number of Web client processes in the case of sport 1. Figures 5 through 7 present the case of music video 1. In these figures, the lower boundaries of the categories are also plotted as straight dotted lines parallel to the abscissa.

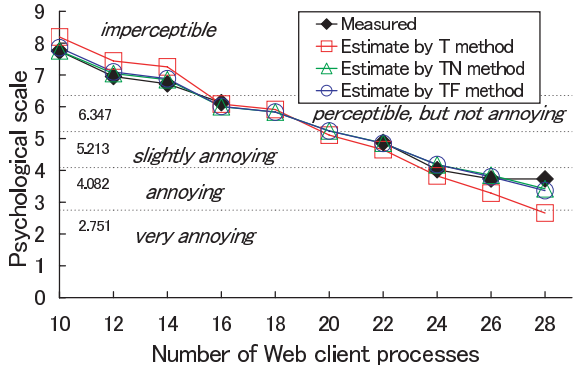

Fig. 2. Psychological scale versus number of Web client processes (sport 1, Picture pattern: I).

Furthermore, Table V displays correlation coefficients between estimates and measured values in the three estimation methods for the four contents.

TABLE V

CORRELATION COEFFICIENT BETWEEN ESTIMATES AND MEASURED VALUES FOR THE THREE ESTIMATION METHODS

	T method	TN method	TF method
sport 1	0.951	0.979	0.974
movie 1	0.942	0.962	0.962
music video 1	0.943	0.955	0.955
animation 1	0.930	0.940	0.958

Figures 2 through 7 and Table V indicate that the TN and TF methods, which utilize both temporal quality parameters and spatial one, provide more accurate estimates than the T method for all the four contents; also, the accuracy of the TN method is comparable to that of the TF method.

So far we have estimated user–level QoS for a content by the multiple regression lines obtained for itself; for example, Eq. (1), (2) and (3) were employed for sport 1.

Next, utilizing the multiple regression lines obtained in the previous subsection, we estimate the user–level QoS for the other content of the same type; namely, sport 2, movie 2, music video 2, and animation 2. For instance, Eq. (1), (2) and (3) are applied to sport 2.

TABLE VI

CORRELATION COEFFICIENT BETWEEN ESTIMATES BY THE MULTIPLE REGRESSION LINES FOR THE OTHER CONTENT AND MEASURED VALUES

	T method	TN method	TF method
sport 2	0.938	0.917	0.948
movie 2	0.895	0.904	0.919
music video 2	0.972	0.978	0.973
animation 2	0.846	0.893	0.888

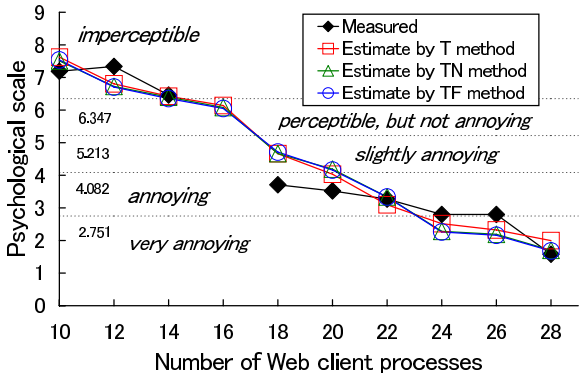

Fig. 3. Psychological scale versus number of Web client processes (sport 1, Picture pattern: IPPPPPPPPPPPPP).

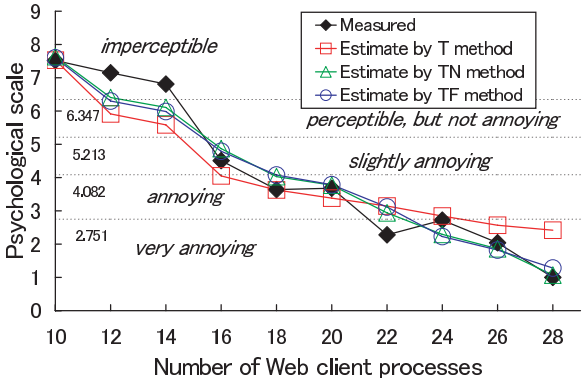

Fig. 4. Psychological scale versus number of Web client processes (sport 1, Picture pattern: IBBPBBPBBPBBPBBPBB).

Table VI shows correlation coefficients between the estimates and measured values. In this table, we observe that the TF method achieves higher values of the correlation coefficient than the T method for all the contents; on the other hand, the TN method does not provide more accurate estimates than the T method for all the contents. This is because the NR metric adopted in this paper is not effective for all content types. Even in the current case, however, note that the accuracy of the TN method is comparable to that of the TF method.

From the above observations, we have learned that appropriate metrics of video spatial quality can improve the accuracy of the estimation. However, the metrics should be based on some NR model from a practical point of view. The finding of such metrics is an important subject in the next step of this research.

Finally, we should mention that the T method can provide estimates comparable to those by the TN and TF methods in many cases. This suggests that the utilization of only temporal quality parameters can be a practical approach to the estimation in some situations. This is for further study.

VI. CONCLUSIONS

This paper proposed a real–time estimation method for user–level QoS of audio–video IP transmission by utilizing both temporal and No Reference spatial quality, namely, the TN method. In terms of the estimation accuracy, we compared it with a method using only the temporal parameters (the T method) for four content types; in addition, just for comparison purposes, we also treated a method with both temporal parameters and Full Reference spatial quality parameter (the TF method), though it cannot be used for real–time estimation.

As a result of experiment, we found that when their own multiple regression lines are used for the estimation, the TN

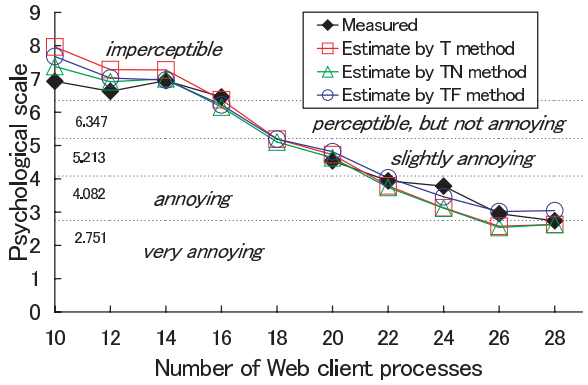

Fig. 5. Psychological scale versus number of Web client processes (music video 1, Picture pattern: I).

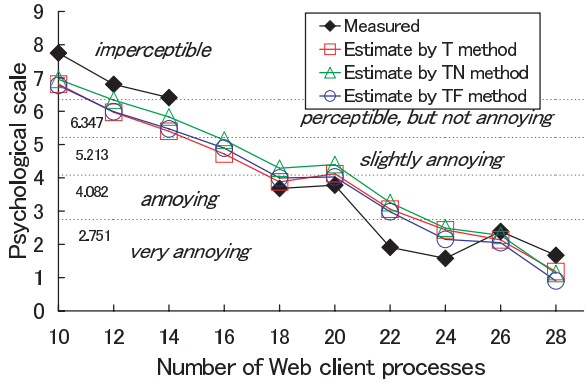

Fig. 6. Psychological scale versus number of Web client processes (music video 1, Picture pattern: IPPPPPPPPPPPPP).

and TF methods provide more accurate estimates than the T method for all the four contents; also, the accuracy of the TN method is comparable to that of the TF method. When we apply the multiple regression lines obtained for the other content of the same type, the TF method achieves higher accuracy than the T method. In this case, however, the TN method does not necessarily provide more accurate estimates than the T method. We also noticed that the T method can provide estimates comparable to those by the TN and TF methods in many cases. From a practical point of view, the T method can be an effective tool for the real–time estimation in some situations; this should be for further study.

Future work include the finding of No Reference metrics of video spatial quality and evaluation of the three estimation methods for a variety of content types.

Acknowledgment

This work was supported by the Grant-In-Aid for Scientific Research of Japan Society for the Promotion of Science under Grant 17360179.

References

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

[2] 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.

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

[4] ITU-R BT.500-11, "Method for the subjective assessment of the quality of television pictures," June 2002.

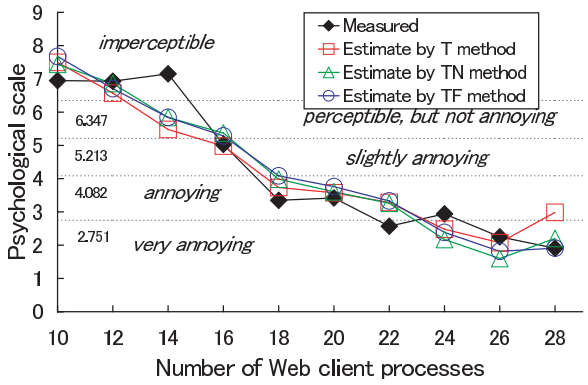

Fig. 7. Psychological scale versus number of Web client processes (music video 1, Picture pattern: IBBPBBPBBPBBPBBPBB).

[5] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," Sept. 1999.

[6] ITU-T J.143, "User requirements for objective perceptual video quality measurements in digital cable television," May 2000.

[7] The Video Quality Experts Group, "The Video Quality Experts Group Web Site," http://www.its.bldrdoc.gov/vqeg/.

[8] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase ", Aug. 2003.

[9] ITU-T Rec. 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," Feb. 2001.

[10] ITU-T Rec. G.107, "The E–model, a computational model for use in transmission planning," March 2003.

[11] C. Jones and D. J. Atkinson, "Development of opinion–based audiovisual quality models for desktop video–teleconferencing", in *Proc. IWQoS*, pp.196–203, Napa, CA, May 1998.

[12] J. G. Beerends and F. E. de Caluwe, "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc.*, vol. 47, no. 5, pp. 355–362, May 1999.

[13] D. S. Hands, "A basic multimedia quality model", *IEEE Trans. Multimedia*, vol. 6, no.6, pp. 806–816, Dec. 2004.

[14] S. Winkler and C. Faller, "Audiovisual quality evaluation of low-bitrate video", in *Proc. SPIE*, vol. 5666, San Jose, CA, 2005.

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

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

[17] S. Tasaka and Y. Ito, "Real–time estimation of user–level QoS of audio–video transmission over IP networks," in *Conf. Rec. IEEE ICC2006*, June 2006.

[18] R. V. Babu, A. S. Bopardikar, A. Perkis and O. I. Hillestad "No–reference metrics for video streaming applications", in *Proc. 14–th International Packet Video Workshop*, Irvine, CA, Dec. 2004.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] Mindcraft Inc. , "WebStone Benchmark Information," http://www.mindcraft.com/webstone/.