Parametric Quality-Estimation Model for Adaptive-Bitrate Streaming Services

Kazuhisa Yamagishi and Takanori Hayashi

***Abstract*—The use of adaptive-bitrate streaming services over networks has been increasing in recent years. The quality of adaptive-bitrate streaming services is primarily affected by the video resolution, the audio and video bitrate, bitrate adaptation, stalling due to a lack of playout buffer, and the content length. Therefore, service providers should monitor quality in real time to confirm the normality of their services. To accurately monitor quality, a model that can be used for quality estimation should be developed. To develop such a model, we first conducted extensive subjective quality assessment tests. We then developed a model using the subjective data obtained in the tests. Finally, we verified the performance of the proposed model by applying it to unknown data sets (different from the training data sets used to develop the model) and confirmed its high quality-estimation accuracy.**

***Index Terms*—Quality, adaptive-bitrate streaming, monitoring, compression, stalling.**

I. Introduction

SIGNIFICANT progress has recently been made in the development of technologies such as encoders and decoders (codecs) [1], [2], streaming protocols [3], [4], and networks. Consequently, streaming service providers can deliver high-resolution (e.g., high definition (HD), ultra HD-1 (UHD-1) and UHD-2) audiovisual content over Internet Protocol (IP) networks.

Streaming services can be classified as real-time transport protocol (RTP)-based streaming (e.g., Linear TV) or hypertext transfer protocol (HTTP)-based streaming (e.g., adaptive-bitrate streaming [HTTP Live Streaming (HLS)] [3] and moving picture experts group dynamic adaptive streaming over HTTP (MPEG-DASH) [4]).

RTP-based streaming provides customers with linear TV over networks (e.g., x-digital subscriber line (xDSL), fiber to the home (FTTH), fiber to the curb (FTTC), cable, and other techniques) using RTP packets. Quality degradation primarily occurs because of compression and IP packet loss. Regarding compression, the quality is affected by the audio bitrate, video resolution, framerate, and bitrate. IP packet loss leads to, for example, freezing of the audio and video frame or audio and video frame loss.

HTTP-based streaming provides customers with the best possible quality for a certain network condition using Transmission Control Protocol (TCP) packets because the client can adaptively select a media file with the suitable bitrate. Quality degradation occurs as a result of compression, network conditions (e.g., packet loss, insufficient bandwidth, delay, and jitter), and lack of a playout buffer [5]–[9]. Like RTP-based streaming, because of compression, the quality is also affected by the audio bitrate, video resolution, framerate, and bitrate. Network issues decrease the throughput and introduce delay at the application layer. Consequently, the playout buffer slowly fills or depletes. When the buffer is empty, the playback of the audiovisual content is interrupted until sufficient data for playback is received. This leads to initial loading and a stalling event. When throughput reduction occurs, the quality level that is most suitable under the current network conditions can be selected (i.e., adaptation) because there are several files, i.e., chunks/segments, corresponding to representations of different bitrates on the server.

The quality factors of both types of streaming services can be summarized as follows. The first is how the source audiovisual content is encoded before transmission. The second is how the IP packets are transmitted over networks. The third is how the encoded audiovisual stream is decoded and displayed at the client terminal, e.g., a set-top box, personal computer (PC), smartphone, or tablet.

Regarding the final perceived quality, the content length (in this paper, the duration of viewing without stalling) also affects the quality because of temporal effects, especially for longer content.

Therefore, to monitor the normality of streaming services, i.e., end-point quality, it is necessary to develop an objective quality-estimation model that can be used for quality estimation at the client.

Objective quality-estimation models can be categorized into two types: media-layer models [10]–[19], which take media signals as input to estimate quality, and parametric models [20]–[42], which take application-layer information, e.g., the bitrate, framerate, resolution, frame type (I-frame, B-frame, and P-frame), quantization parameter (QP), motion vector, and stalling information, as inputs to estimate the quality.

For quality monitoring at the client, a parametric model is suitable because the client is not allowed to access media signals due to the encrypted media-related bitstream and because a low computational cost is required to avoid consuming too much of the client resources.

Although there are parametric models that take bitstream information, e.g., frame type (I-, B-, and P-frames), QP, and motion vector, as input [20]–[26], high computational power is necessary to parse bitstream information. Therefore, a parametric quality-estimation model that does not use bitstream information as input should be developed [28]–[36].

The authors are with NTT Network Technology Laboratories, NTT Corporation, 3-9-11 Midori-cho, Musashino-shi, Tokyo 180-8585, Japan.

Manuscript received July 12, 2016; revised November 15, 2016 and January 10, 2017.

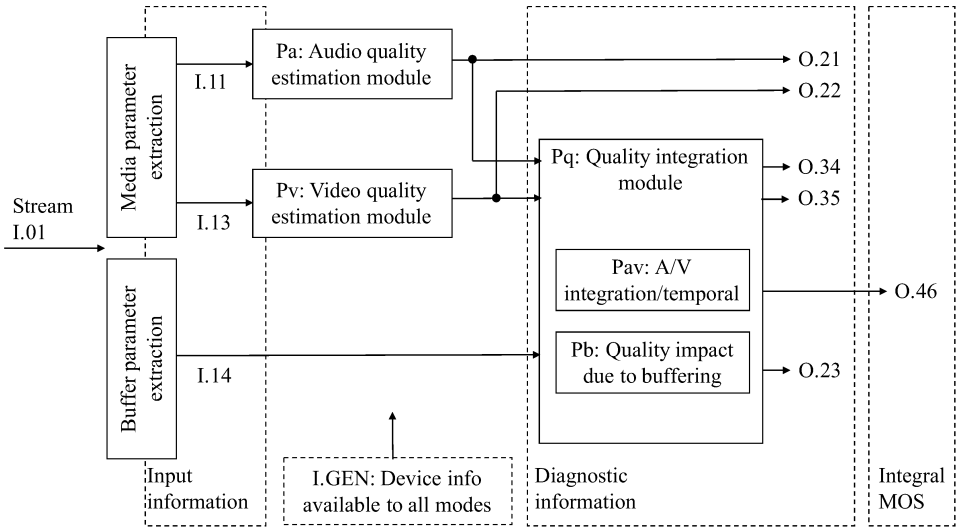


Fig. 1. Block diagram of PNATS model proposed by ITU-T SG12

The International Telecommunication Union Telecommunication Standardization Sector Study Group 12 (ITU-T SG12) studied and standardized the parametric non-intrusive assessment of audiovisual media streaming quality for RTP-based streaming services (so-called PNAMS) – lower resolution [LR: i.e., quarter common intermediate format (QCIF, 176 × 144 pixels), quarter video graphics array (QVGA, 320 × 240 pixels), or half VGA (HVGA, 320 × 480 pixels)], higher resolution [HR: i.e., standard definition (SD, 720 × 480 pixels and 720 × 576), and high definition (HD, 1280 × 720 or 1920 × 1080 pixels)] applications [28]–[32] because the quality estimation accuracy for these models was sufficiently high. There are still some issues regarding developing a model for higher-resolution video, i.e., 4K UHD and 8K UHD video.

ITU-T SG12 has also been developing a model for parametric non-intrusive assessment of TCP-based multimedia streaming quality (so-called PNATS). A block diagram proposed by ITU-T SG12 of the PNATS model is shown in Fig. 1. The model consists of a parameter extraction module, audio quality estimation module (Pa), video quality estimation module (Pv), and quality integration module (Pq). The input is defined as I.01: stream, I.11: audio-related parameters, e.g., audio bitrate, I.13: video-related parameters, e.g., video bitrate, I.14: stalling events, i.e., a tuple of start time and duration, both measured in seconds, I.GEN: the resolution of the image displayed to the user and device type on which the media is played (either PC or mobile device). The output is defined as: O.21: audio coding quality per output sampling interval, O.22: video coding quality per output sampling interval, O.23: perceptual buffering indication, O.34: audiovisual segment coding quality per output sampling interval, O.35: final audiovisual coding quality score, which includes aspects of temporal integration, O.46: final media session quality score. O.21, O.22, and O.34 are output per-one-second scores on a 1-5 quality scale. O.23, O.35, and O.46 are output single scores on a 1-5 quality scale for the session. However, the model has not yet been developed. ITU-T has plans to finalize the development of the model and standardize it in 2017. We address this issue in this paper.

For end-point quality monitoring purposes, a parametric quality-estimation model should estimate quality that is affected by bitrate adaptation (e.g., the audio and video bitrate, video resolution, and framerate per segment of adaptive-bitrate streaming), playout buffering (e.g., initial loading and stalling due to the lack of a playout buffer), and content length [5]. Although video framerate reduction due to video bitrate reduction is often used for smaller screens (e.g., smartphones), video framerate reduction due to video bitrate reduction is not generally performed for larger screens (e.g., TVs, personal computers, and tablets). Service providers often add advertisements to generate revenue and profits. These advertisements are provided randomly or personalized. They affect the impact of initial loading on quality because advertisements are immediately played after the initial loading. Therefore, the impact of video framerate reduction and initial loading on quality is out of the current paper's scope.

We developed a parametric quality-estimation model that can be used to estimate end-point quality for adaptive-bitrate streaming services. We applied the block diagram proposed by ITU-T SG12 to our model. Our model takes the audio and video bitrate, video resolution, stalling-related information, and content length parsed from received packets without media bitstreams at the client application as input because parsing bitstream information requires computational power, and this information is usually encrypted, as described above.

We first describe conventional models and issues that need to be addressed in Section II. We then present our proposed parametric quality-estimation model in Section III. We explain subjective quality assessment tests in Section IV. We verify in Section V that the proposed model has sufficient quality-estimation accuracy using training and validation data sets. We discuss some relevant considerations in Section VI. Finally, we conclude with a summary and mention potential future work in Section VII.

II. Related Work

As described in Section I, it is desirable to develop a parametric quality-estimation model that can be used to estimate quality that is affected by the audio and video bitrate, video resolution, bitrate adaptation, stalling events, and content length. In this section, we describe the conventional models and the issues that must be addressed.

A. Impact of Compression, Audiovisual Interaction, and Adaptation on Quality

Conventional parametric quality-estimation models [28]–[37] estimate quality on the basis of a constant bitrate because a constant bitrate for a certain video resolution (e.g., standard definition (SD), 720p, and 1080i/p) is used in RTP-based linear TV services. Although it is possible to estimate quality for a short-term length, i.e., a segment for adaptive-bitrate streaming, the conventional models cannot be used to estimate quality that is affected by bitrate adaptation because the bitrate and video resolution vary according to the network condition. Although video-frame-type estimation is incorporated into several models [29]–[32], [37], such estimation is unrealistic in adaptive-bitrate streaming because it is difficult to detect the first packet of a video frame (e.g., a marker bit in an RTP header) in adaptive-bitrate streaming. The audio and video bitrate and video resolution can be parsed using e.g., a media presentation description from metadata (MPD in MPEG-DASH [4]) in a client. Therefore, it is necessary to develop a model that takes audio and video bitrate and video resolution per segment as input.

Audio-quality-estimation models, which output short-term audio quality (approximately 10 to 30 sec in this paper) on the basis of a constant bitrate have been extensively studied and have reached high quality-estimation accuracy [28]–[32]. It is considerable that such audio-quality-estimation models can be used to estimate the quality of a segment of length e.g., 10 sec in adaptive-bitrate streaming.

Video-quality-estimation models, which output short-term video quality, have been proposed [33]–[36]. These models [33], [34] take only bitrate as input; thus, they cannot be used to evaluate the impact of video resolution on video quality. The T-V model [35] was developed for the SD and HD resolutions on the basis of the bitrate, and the enhanced G.1070 model [36] was developed for SD, VGA, CIF, and QCIF. However, these models need to be optimized for a certain video resolution because other video resolutions are used in adaptive-bitrate streaming, e.g., 180p, 240p, 360p, and 720p. Therefore, it is ideal to develop a model that takes video resolution (i.e., horizontal pixels and aspect ratio or number of pixels) and bitrate as input and for which it is not necessary to optimize the model for a certain video resolution.

An audiovisual-quality-estimation model that outputs short-term audiovisual quality has been proposed [29], [38]–[40], and it performs well in terms of quality-estimation accuracy. It is notable that such audiovisual-quality-estimation models can be used to estimate the quality of a segment, e.g., 10 sec, in adaptive-bitrate streaming.

Long-term, i.e., 1- to 3-min content in this paper, coding quality-estimation models have been proposed [41], [42]. Takanori et al. [41] modeled the impact of quality for short segments and the temporal effect on the final video quality in a 3-min video without audio. The model performed well in terms of the quality-estimation accuracy. However, the video resolution and audiovisual interaction were not taken into account. Yun et al. [42] modeled the impact of bitrate switching and, the primacy and the recency effect on the final quality. They took into account the amount of bitrate switching. However, the difference in quality before and after switching was not considered in the model. If the difference in quality before and after switching is small, it is not noticeable. Therefore, if the number of switches is incorporated into the model, the quality difference should be incorporated simultaneously. Video resolution and audiovisual interaction were also not taken into account. Yun et al. also demonstrated that their model, which estimates quality using the average quality per segment, did not perform well in terms of the quality-estimation accuracy. Therefore, the impact of the combination of video resolution, adaptation, and audiovisual interaction on audiovisual coding quality needs to be addressed.

B. Impact of Stalling on Quality

Conventional parametric quality-estimation models [29], [37] can be used to estimate quality that is affected by a stalling event. However, these models can only be applied to constant-bitrate content, i.e., content without bitrate adaptation. In the case of adaptive-bitrate streaming, it is necessary to develop a model that can be used to estimate quality that is affected by bitrate adaptation. Furthermore, even when the total stalling length is the same for both, for example, 30-sec and 3-min contents, the impact on the final quality is different for each of them. Therefore, the impact of content length on quality cannot be evaluated with such a model.

Kamal et al. proposed a bitstream-based model [43] that can be used to estimate the quality affected by stalling events of adaptive-bitrate streaming. Although this type of model is out of the current paper's scope because the QP is based on the bitstream, as described in Section I, a part of the model for stalling can be used to evaluate the impact of stalling on quality. However, the impact of the interval between stalling events and content length on quality cannot be evaluated because the number, average length, and maximum length of stalling events are taken as input.

Yao et al. proposed a media-based model [44] that takes the number and duration of stalling events and the output of the video quality metric (VQM) [10] as input. They also proposed a bitstream-based model [45] that takes the number and duration of stalling events and the motion vector as input. Although these models are out of the current paper's scope because the output of the VQM is based on the media signal and motion vectors are based on the bitstream, as described in Section I, a part of the model for stalling can be used to evaluate the impact of stalling on quality. However, the impact of the interval between stalling events and content length on quality is not taken into account in these models.

C. Summary of Issues That Must Be Addressed

As described above, there are issues with estimating the quality of a video of arbitrary resolution in short-term video-quality-estimation models. However, conventional audio and audiovisual-quality-estimation models can be used to estimate the quality of a segment in adaptive-bitrate streaming, as described above. In long-term audiovisual-coding-quality estimation, the impact of the combination of video resolution, adaptation, and audiovisual interaction on audiovisual coding quality needs to be taken into account. To evaluate the impact of stalling on quality, the content length needs to be taken into account. To estimate the final audiovisual quality, i.e., media session quality, with high quality-estimation accuracy, it is necessary to conduct a subjective test that varies the combination of audio and video bitrate, video resolution, bitrate adaptation, stalling events, and content length and to model their characteristics.

III. Parametric Quality-Estimation Model

We propose our parametric quality-estimation model in this section. The block diagram shown in Fig. 1 was proposed by ITU-T SG12, and we applied it to our model. In the model, we take into account the issues that must be addressed, as mentioned in Section II, where the impact of video framerate reduction and initial loading on quality is out of the current paper's scope, as mentioned in Section I. Note that, in the PNATS model, the framerate reduction is addressed because the mobile mode is considered in addition to the PC mode. Also note that the initial loading is addressed because the advertisement issue mentioned in Section I is not considered.

A. Scope of the Parametric Quality-Estimation Model

Our parametric quality estimation model takes audio- and video-related parameters, e.g., bitrate, and stalling event parameters, e.g., the number of stalling events, extracted from a client as input because bitstreams are often encrypted to protect content copyrights. The output of our model is the final audiovisual quality, i.e., media session quality, of adaptive-bitrate services for TV screens. To cover a wide range of quality, 426×240 to 1920×1080 pixels, 100 to 10000 kbps of video bitrate, and 64 to 196 kbps of audio bitrate can be treated using our model because low-resolution and low-bitrate content is provided to TV sets using Apple TV or Chromecast. Regarding audio and video codecs, the combination of AAC-LC and H.264/AVC (main profile and level 4.0) is supported because they are widely used in TV sets. Although the profile and level of H.264/AVC can be selected for the combination of resolution and framerate to improve quality, a single profile and level is assumed by the model presented in this paper because, by definition, the model cannot have access to the bitstream and because the main target of the model is TV sets.

B. Extraction of Application-Layer Information

To estimate the audiovisual quality, i.e., media session quality, for adaptive-bitrate streaming services, it is necessary to extract application-layer information from a client. To extract application-layer information, parameter extraction modules need to be implemented into the client application, and extraction modules must deliver parameters to the quality estimation modules, as shown in Fig. 1.

The media parameter extraction module extracts the audio and video bitrate and video resolution per segment from the client, e.g., the MPD in MPEG-DASH, because an application on the client first receives, for example, MPD from the server to select media chunks/segments. The quality level changes due to bitrate adaptation; thus, the audio bitrate [$abr(t)$], video resolution [$vrs(t)$, total number of pixels] and bitrate [$vbr(t)$] are calculated and stored per second with the module. The module delivers parameters to the audio and video quality estimation modules.

To correctly play, applications/decoders need to manage when stalling events occur and when playback is resumed. A stalling event log with timestamps can be used to calculate the stalling parameters. The buffer parameter extraction module extracts the number of stalling events (N), total length of stalling (L), and average interval between stalling events (A) from the stalling event log with timestamps. The audiovisual content length (T), which does not include the stalling length, is extracted from the client. The module delivers parameters to the quality integration module.

As one of the use cases, ExoPlayer provided by Google can be used to extract media and stalling parameters on the basis of the procedures described above.

C. Audio-Quality-Estimation Module

The audio-quality-estimation module outputs audio quality per 1-sec sampling interval, i.e., the same as O.21 in Fig. 1, by using the audio bitrate [$abr(t)$].

In general, audio quality is saturated to a maximum audio quality, i.e., approximately 5, with sufficient audio bitrate and decreases to 1 as the audio bitrate decreases. In addition, the decrease degree depends on the codec implementation. This characteristic is often modeled using a logistic function [29]. Therefore, audio quality is modeled using a logistic function, i.e., the audio quality saturates to the maximum audio quality (a_1), and the decrease degrees are determined by coefficients, a_2 and a_3, as follows:

$$AQ(t) = Max(1, Min(5, a_1 + \frac{1 - a_1}{1 + (abr(t)/a_2)^{a_3}})), \quad (1)$$

where t = 1, 2, ..., and T is in seconds; $abr(t)$ (kbps) is calculated from the audio bitrate for a segment, e.g., when the audio bitrate for the first 5-sec segment is 192 kbps, $abr(t)$ = 192 kbps (t = 1, 2, ..., and 5); $AQ(t)$ represents the audio quality per second; a_1, a_2, and a_3 are constants and positive values; the $Max()$ function calculates the maximum value; and the $Min()$ function calculates the minimum value.

D. Video-Quality-Estimation Module

The video-quality-estimation module outputs video quality per 1-sec sampling interval, i.e., the same as O.22 in Fig. 1, using the video bitrate [$vbr(t)$] and video resolution [$vrs(t)$].

Like audio quality, video quality is also saturated to a maximum quality with sufficient video bitrate according to the video resolution and decreases to 1 as the video bitrate decreases. The characteristics are also modeled using a logistic function, as observed in (2) [33]. The maximum video quality according to the video resolution is not always saturated to approximately 5, and the maximum video quality decreases as the video resolution decreases. Therefore, it is assumed that the characteristic can be modeled using a Michaelis-Menten function, as observed in (3). The impact of the bitrate reduction on video quality depends on the video resolution and decreases as the video resolution decreases, as observed in (4). Therefore, we model the video-quality-estimation module as follows:

$$VQ(t) = VQ_{MAX}(t) + \frac{1 - VQ_{MAX}(t)}{1 + (vbr(t)/\tau(t))^{v_1}}, \quad (2)$$

$$VQ_{MAX}(t) = Max(1, Min(5, 1 + 4 \cdot \frac{v_3 \cdot vrs(t)}{v_2 + vrs(t)})), \quad (3)$$

$$\tau(t) = \frac{v_4 \cdot vrs(t) + v_6}{1 - \exp(-v_5 \cdot vrs(t))}, \quad (4)$$

where $VQ(t)$ represents the video quality per second, $VQ_{MAX}(t)$ represents the maximum video quality per second for a certain video resolution, and v_1 through v_6 are constants with positive values. $vrs(t)$ and $vbr(t)$ are calculated on the basis of the video resolution and bitrate for a segment, as in $abr(t)$.

E. Audiovisual-Quality-Estimation Module

The audiovisual-quality-estimation module outputs the audiovisual quality per 1-sec sampling interval, i.e., the same as O.34 in Fig. 1, using the audio and video quality [$AQ(t)$ and $VQ(t)$]. This module is not depicted in the original block diagram proposed by ITU-T (Fig. 1). However, to better understand how the model works, it has been added to the Pq module in our block diagram.

Audiovisual quality is expressed using audio quality and video quality, and according to [29] and [38], the characteristics can be modeled using a multiple regression function. Therefore, the audiovisual-quality-estimation module can be expressed as follows:

$$AVQ(t) = Max(1, Min(5, av_1 + av_2 \cdot AQ(t) + av_3 \cdot VQ(t) + av_4 \cdot AQ(t) \cdot VQ(t))), \quad (5)$$

where $AVQ(t)$ is the audiovisual quality per second and av_1 through av_4 are constants with positive values.

F. Audiovisual-Integration/Temporal Module

The audiovisual-integration/temporal module outputs the final audiovisual coding quality, i.e., the same as O.35 in Fig. 1, using the audiovisual quality ($AVQ(t)$) and content length (T).

The final audiovisual coding quality is affected by the audiovisual quality per 1-sec sampling interval, the temporal effect, and the content length. The impact of the temporal effect on audiovisual coding quality increases as time increases; thus, it is conceivable that the characteristics are modeled by an exponential function, as observed in (7) and (8). The impact of bitrate switching on quality needs to be incorporated in the model. If the number of bitrate switches is taken as input, the model also needs to consider the difference between video qualities before and after switching. For example, there are 5 quality levels (i.e., QL0, QL1, QL2, QL3, and QL4). If the difference in the quality between QL1 and QL2 is very small (e.g., 0.1 in the 5-scale MOS), and if the difference in the quality between QL3 and QL4 is large (e.g., 0.5 in the 5-scale MOS), the difference between QL1 and QL2 is not noticeable, whereas the difference between QL3 and QL4 is noticeable. Therefore, if the number of switches is incorporated into a model, the difference in quality before and after a switch should be incorporated simultaneously. In addition, bad audiovisual quality per 1-sec sampling interval has a large impact on the final audiovisual coding quality. To take these characteristics into account, we introduce a function of $w_2(t)$, i.e., Eq. (9), and it is assumed that the characteristics can be modeled using a linear function, as observed in (9). When the quality is increased or decreased owing to bitrate switching, the values of $AVQ(t)$ and w_2 are also increased or decreased. It is conceivable that these features take into account the impact of switching on quality. Therefore, the number of switches and the difference between qualities before and after a switch are not used in our proposed model to avoid the overfitting issue. As a result, the audiovisual integration/temporal module can be expressed as follows:

$$AVCQ = \frac{\sum_{t=1}^{T} w_1(t) \cdot w_2(t) \cdot AVQ(t)}{\sum_{t=1}^{T} w_1(t) \cdot w_2(t)}, \quad (6)$$

$$w_1(t) = t_1 + t_2 \cdot \exp(\frac{u(t)}{t_3}), \quad (7)$$

$$u(t) = \frac{t}{T}, \quad (8)$$

$$w_2(t) = t_4 - t_5 \cdot AVQ(t), \quad (9)$$

where $AVCQ$ is the audiovisual coding quality for an audiovisual content length (T =, e.g., 60 and 180 sec) and t_1 to t_5 are constants with positive values.

G. Quality Integration Module

The quality integration module outputs the final media session quality, i.e., the same as O.46 in Fig. 1, by using the audiovisual coding quality, stalling events, and content length.

The media session quality exponentially decreases as the number of stalling events and the total length of stalling events increase. Because the interval between stalling events also affects the media session quality, we propose hypotheses for the impact of the interval between stalling events on the media session quality. When the interval between stalling events is short, it is conceivable that the impact on media session quality is small owing to temporal effects. If the interval is short, e.g., a few seconds, and the number of stalling events is two, users would feel that the stalling event occurred once in a long sequence, whereas if the interval is long, users would feel that the stalling event occurred twice. From these hypotheses, we introduced the average interval between stalling events. When the average interval is short, the impact on the media

session quality is small, and vice versa. The total length of stalling and the average interval between stalling events are normalized from the audiovisual content length because it is thought that the stalling impact on quality depends on the audiovisual content length. For example, the impact of a few-second stalling event on the final quality differs depending on whether the content length is short or long. In addition, there is the possibility of using the viewing duration instead of the content length. The improvement in estimation accuracy when using either viewing duration or content length is almost the same because the total stalling length normalized by either the viewing duration or content length is also divided by the coefficient in Eq. (11). To take these aspects into account, we suggest that the stalling length be normalized by the content length. These characteristics are modeled in (11).

The quality-integration module takes the number of stalling events N, the total length of stalling events L (sec), the average of the interval between stalling events A (sec), the audiovisual coding quality $AVCQ$, and the content length T as input. Therefore, the media session quality (MSQ) can be modeled as follows:

$$MSQ = 1 + (AVCQ - 1) \cdot S, \quad (10)$$

$$S = \exp(-\frac{N}{s_1}) \cdot \exp(-\frac{L/T}{s_2}) \cdot \exp(-\frac{A/T}{s_3}), \quad (11)$$

where s_1, s_2, and s_3 are constants with positive values.

H. Procedures for Calculating Coefficients

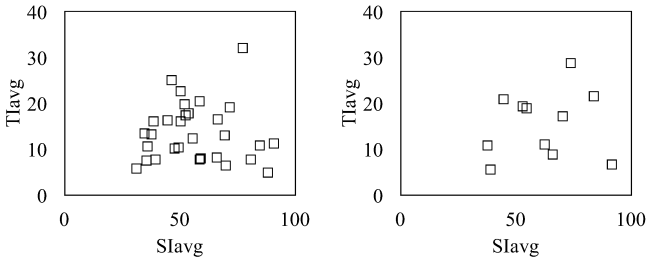
All the coefficients (a_1, a_2, a_3, v_1 to v_6, av_1 to av_4, t_1 to t_5, s_1, s_2, and s_3) of our proposed model are calculated on the basis of the least squares method using subjective data. The calculation of coefficients proceeds as follows. In step 1, initial values are set for all coefficients, i.e., a_1, a_2, a_3, v_1 to v_6, av_1 to av_4, t_1 to t_5, s_1, s_2, and s_3. In step 2, the coefficients (a_1, a_2, and a_3) of the audio quality estimation module are trained. In step 3, the coefficients (v_1 to v_6) of the video quality estimation module are trained. In step 4, the coefficients (av_1 to av_4) of the audiovisual quality estimation module are trained. In step 5, the coefficients (t_1 to t_5) of the audiovisual integration/temporal module are trained. Finally, in step 6, the coefficients (s_1, s_2, and s_3) of the quality integration module are trained. In each step, the target coefficients are trained using all the training data, and the others are not trained. Steps 2 to 6 are repeated until the error is minimized.

IV. Subjective Quality Assessment Tests

We conducted two subjective quality assessment tests [Experiments 1 (1-min content) and 2 (3-min content)] to train the model and verify its quality-estimation accuracy. All the subjective data were our own confidential information; thus, the data are only for our private use.

A. Audiovisual Content – Source Reference Circuits

To conduct a subjective test, high-quality content is required, and a large amount of long high-quality content is not available for research. Therefore, we asked two professional video production companies, Q-tec and NTT-IT, to shoot audiovisual content. This content was categorized into sports, dancing, music, TV drama and shopping, cooking, and scenery. A source reference circuit (SRC) that has various spatial-temporal characteristics, e.g., spatial detail and motion, needs to be selected for subjective tests, as described in ITU-T Rec. P.910 [46]. According to ITU-T Rec. P.910, spatial information (SI) is defined as the maximum standard deviation of the pixels in each Sobel-filtered video frame (SImax), and temporal information (TI) is defined as the maximum standard deviation of the motion difference feature (TImax). However, we also defined TIavg and SIavg using the average value to determine the characteristics of an entire scene.



(a) Experiment 1 (b) Experiment 2

Fig. 2. SIavg vs. TIavg in Experiments 1 and 2

We used 30 different types of audiovisual SRCs, each lasting 1 min, in Experiment 1 and 11 different types of audiovisual SRCs, each lasting 3 min, in Experiment 2. Scatterplots of the relationships between SIavg and TIavg are shown in Fig. 2. The average and maximum values of SI and TI are listed in Table I. As Fig. 2 and Table I indicated, these values were widely scatted.

The original video format was 1920 × 1080/30 fps.

All audio SRCs were normalized at the nominal level of -26 dBov, and the sampling rate was 48 kHz.

B. Experimental Settings – Hypothetical Reference Circuits

The video was encoded using H.264 (main profile and level 4.0), and the audio was encoded using AAC-LC. In Experiments 1 and 2, 11 quality levels (QLs) for compression were used, as listed in Table II. The segment length was 5 sec when the group of picture (GoP) length was 1 sec, and the segment length was 4 sec when the GoP length was 2 sec. In Experiment 1, 30 hypothetical reference circuits (HRCs) were used, as shown in Fig. 3. In Experiment 2, 11 HRCs were used, as shown in Fig. 4. Here, "QL = -1" in Figs. 3 and 4 means that a stalling event occurred.

C. Test Stimuli – Processed Audiovisual Sequences

We used 60 processed audiovisual sequences (PAVSs) in Experiment 1 (1-min test), as listed in Table III, and 22 PAVSs in Experiment 2 (3-min test), as listed in Table IV, where the number of each cell represents the PAVS number, the column number represents the SRC number, and the row number represents the HRC number in the matrix. Ideally, each SRC should be assigned to all the HRCs if the SRC is short, e.g., a

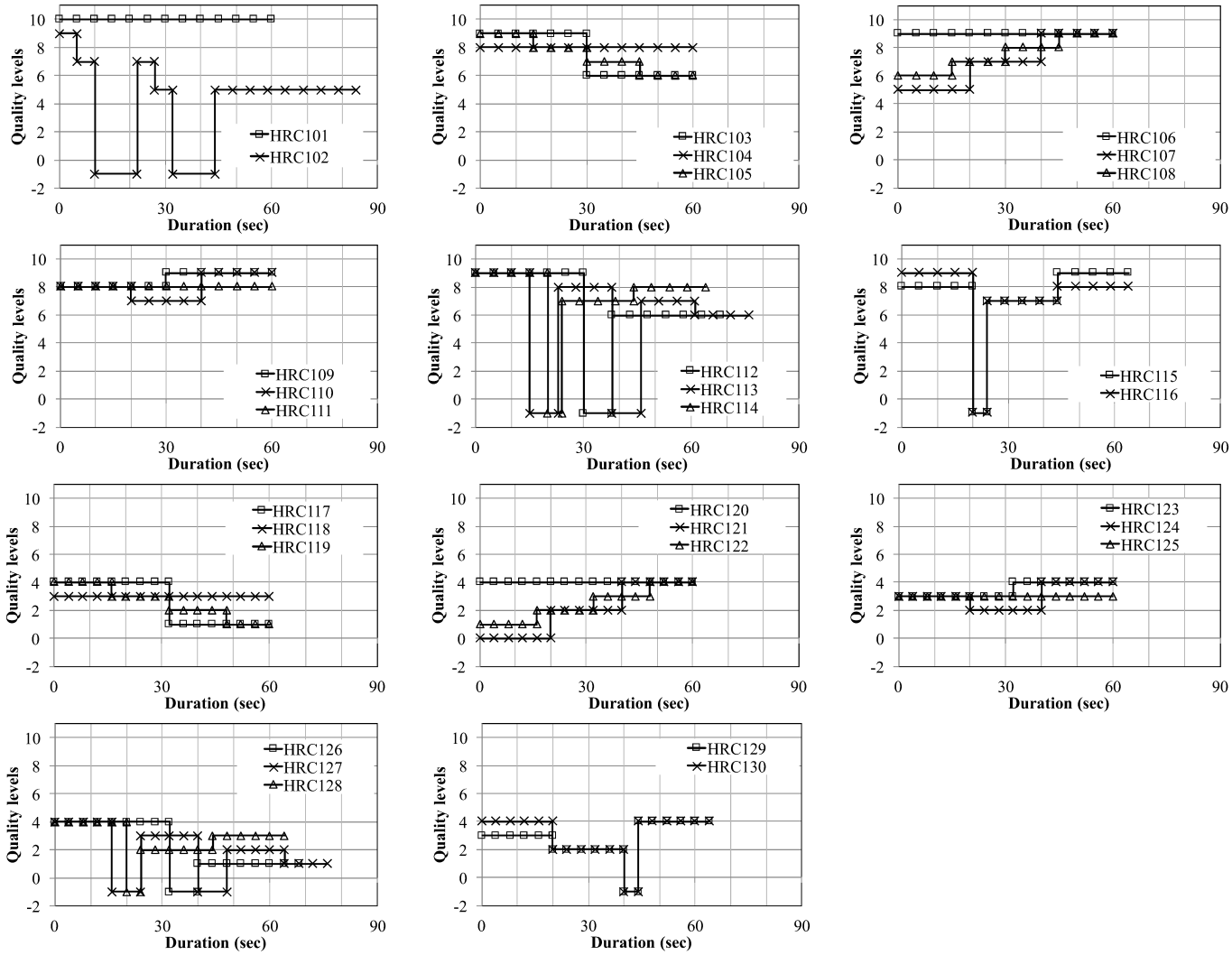


Fig. 3. HRCs in Experiment 1

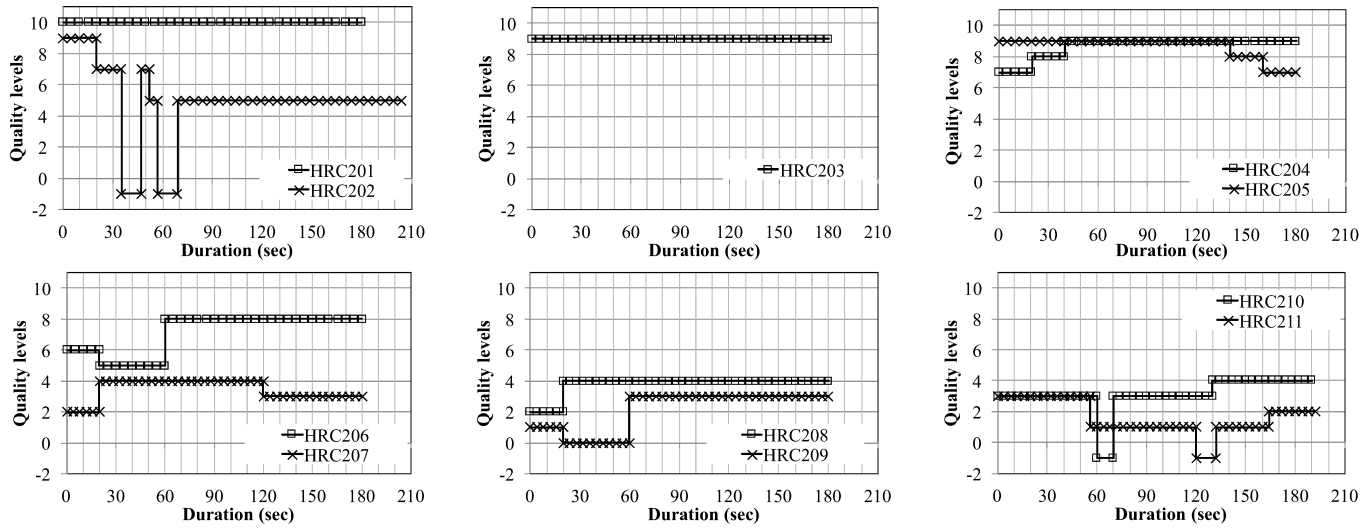


Fig. 4. HRCs in Experiment 2

10-sec sequence. However, in the case of a longer sequence, it is difficult to assign SRCs to all HRCs because the subjective test duration should be limited, i.e., a subjective test longer than 2 hrs should be avoided from the viewpoint of subject

TABLE I
SI AND TI IN EXPERIMENTS 1 AND 2

(a) Experiment 1

SRC	Title	SIavg	SImax	TIavg	TImax
101	Samba	59	95	20	96
102	Animal	70	147	13	61
103	Cooking	56	73	12	87
104	River	77	110	32	124
105	Piano	36	71	7	17
106	Athletics	47	79	25	105
107	Bicycle	59	72	8	50
108	Car	85	114	11	89
109	Classic	50	61	10	61
110	Dance	54	71	18	101
111	Family	36	74	10	79
112	Fish	38	72	13	30
113	Fried rice	70	102	6	78
114	Future world	45	78	16	45
115	Floor exercises and vault	39	69	16	67
116	Rings and pommel horse	35	51	13	64
117	Music clip	40	104	8	34
118	Jam	67	93	16	59
119	Japanese dance	66	72	8	19
120	Karate	52	91	20	85
121	Land	81	166	8	62
122	Marine life	51	77	16	42
123	Living	48	67	10	79
124	Makeup	31	42	6	68
125	Office	59	81	8	76
126	TV shopping	89	100	5	80
127	Grassland	53	80	17	64
128	Taxi	72	108	19	77
129	Train	91	147	11	90
130	Waterpolo	51	100	23	75

(b) Experiment 2

SRC	Title	SIavg	SImax	TIavg	TImax
201	Animal	53	112	19	60
202	Badminton	45	78	21	85
203	Cooking	63	104	11	102
204	Train	71	126	17	111
205	Music clip	38	105	11	35
206	Japanese dance	66	73	9	19
207	Karate	55	128	19	89
208	Makeup	40	72	5	77
209	River	84	143	21	124
210	Samba	74	121	29	99
211	TV shopping	92	107	6	88

TABLE II
QUALITY LEVELS IN EXPERIMENTS 1 AND 2

QL	Resolution (pixels)	Framerate (fps)	GoP (sec)	Video bitrate (kbps)	Audio bitrate (kbps)
0	426×240	30	2	100	64
1	640×360	30	2	250	96
2	852×480	30	2	400	96
3	1280×720	30	2	1200	128
4	1920×1080	30	2	2000	128
5	426×240	30	1	150	64
6	640×360	30	1	350	96
7	852×480	30	1	500	96
8	1280×720	30	1	1600	128
9	1920×1080	30	1	2500	128
10	1920×1080	30	1	10000	196

fatigue and test reliability. In addition, if an SRC is repeated many times, it becomes boring for subjects. Therefore, we designed each SRC to be assigned only to two specific HRCs. The drawback of this test design is that the impact of the SRC on quality may not be observed. This issue needs to be clarified in future work.

Each PAVS was encoded using the codec and an assigned HRC. The PAVSs were decoded into YUV420/8-bit files and WAV files. In the stalling HRC, we added loading to a frozen video frame in the YUV domain and silent audio to the WAV file.

The PAVSs were classified into two groups such that one experiment could be conducted to train the model and the other could be conducted to verify the quality-estimation accuracy of the model for unknown data. That is, when Experiment 1 was used for the training data, Experiment 2 was used for the validation data, and vice versa, i.e., cross-validation was performed.

D. Subjective Quality Assessment Method and Environment

In the subjective quality assessment, the subjective audiovisual quality, i.e., media session quality, was evaluated using an absolute category rating (ACR) method with a five-grade quality scale (5: Excellent, 4: Good, 3: Fair, 2: Poor, 1: Bad) [47]. The quality descriptions on the rating scale were given in Japanese. There was a training session and four sub-sessions for each test. In Experiment 1 (1-min test), there were 15 PAVSs per training session or sub-session. In Experiment 2 (3-min test), there were six PAVSs for the training session, six PAVs for each of the first two sub-sessions, and five PAVSs for each of the last two sessions. There were 3-min breaks between sub-sessions in both experiments. As a result, the total test duration was less than 2 hours. The presentation order of the PAVSs was randomized in these tests. We used a 42-inch LCD monitor and headphones. Participants viewed each PAVS at a distance of 3H (approximately 157.2 cm), where H indicates the ratio of viewing distance to picture height, and listened to each PAVS at a 73-dB(A) SPL.

E. Test Subjects

Twenty-four participants aged 20 - 39 participated in each experiment. They were non-experts who were not directly concerned with audiovisual quality as a part of their work; therefore, they were not experienced assessors.

V. PERFORMANCE EVALUATION

We first trained our proposed model using training data and then verified the quality-estimation accuracy of our proposed model for the training and validation data.

A. Minimum Performance Requirement

This section describes our target quality-estimation accuracy (minimum performance requirement). In ITU-T Rec. P.1201, the performance of a model for higher- and lower-resolution modes is described. The root mean square error (RMSE) and Pearson's correlation coefficient (PCC) were used to evaluate the performance. The RMSE ranged from 0.32 to 0.52 for the higher mode and from 0.36 to 0.65 for the lower mode. The

TABLE III
PAVSs IN EXPERIMENT 1

SRC ＼HRC	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130
101	101																													131
102		102																											132	
103			103																									133		
104				104																							134			
105					105																					135				
106						106																			136					
107							107																	137						
108								108															138							
109									109													139								
110										110											140									
111											111									141										
112												112							142											
113													113					143												
114														114			144													
115															115	145														
116															146	116														
117														147			117													
118													148					118												
119												149							119											
120											150									120										
121										151											121									
122									152													122								
123								153															123							
124							154																	124						
125						155																			125					
126					156																					126				
127				157																							127			
128			158																									128		
129		159																											129	
130	160																													130

TABLE IV
PAVSs IN EXPERIMENT 1

SRC ＼HRC	201	202	203	204	205	206	207	208	209	210	211
201	201	212									
202	213	202									
203			203	214							
204			215	204							
205					205	216					
206					217	206					
207							207	218			
208							219	208			
209									209	220	
210										210	221
211									222		211

PCC ranged from 0.86 to 0.94 for the higher mode and from 0.70 to 0.95 for the lower mode.

As described in Section I, by definition, our target model does not have a video-frame-type estimation module, although the P.1201 model does. Because the RMSE and PCC of our model seem worse than those of the P.1201 model, we used the worst values, i.e., "RMSE $\leq$ 0.65" and "PCC $\geq$ 0.70," as a minimum performance requirement to verify that the quality-estimation accuracy of the model is sufficient.

B. Cross-Validation (Training Data: Experiment 1)

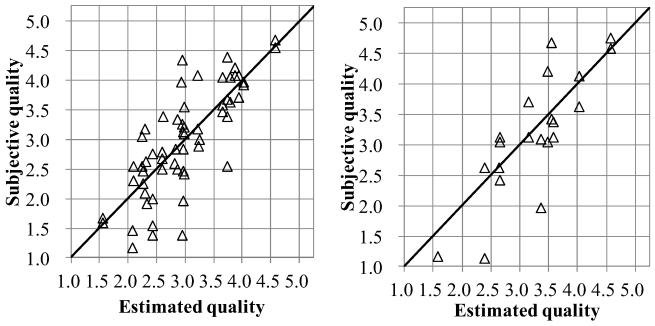
We calculated the coefficients (a_1, a_2, a_3, v_1 to v_6, av_1 to av_4, t_1 to t_5, s_1, s_2, and s_3) of the proposed model using 60 PAVSs for Experiment 1. We estimated the subjective qualities for the training data, i.e., Experiment 1, and validation data, i.e., Experiment 2. For the training data, the RMSE was 0.54, and the PCC was 0.79. For the validation data, the RMSE was 0.58, and the PCC was 0.85. The relationship between the estimated quality and subjective quality for the training and validation data is shown in Fig. 5.

C. Cross-Validation (Training Data: Experiment 2)

We conducted cross-validation by changing the training and validation data. We calculated the coefficients (a_1, a_2, a_3, v_1 to v_6, av_1 to av_4, t_1 to t_5, s_1, s_2, and s_3) of the proposed model using 22 PAVSs for Experiment 2. We estimated the subjective qualities for the training data, i.e., Experiment 2, and validation data, i.e., Experiment 1. For the training data, the RMSE was 0.50, and the PCC was 0.88. For the validation data, the RMSE was 0.57, and the PCC was 0.78. The relationship between the estimated quality and subjective quality for the training and validation data is shown in Fig. 6.

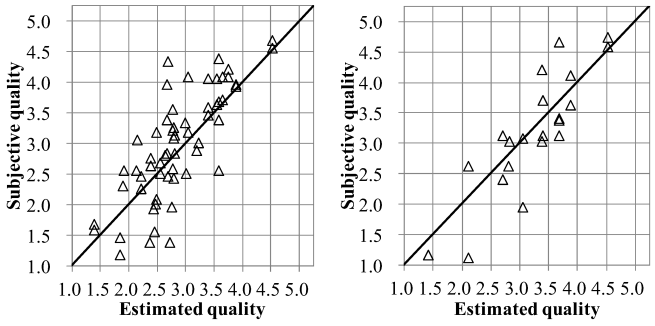
D. Coefficient Sets of Parametric Quality-Estimation Model

Our proposed parametric quality-estimation model was well trained because both the RMSEs and PCCs were almost the



(a) Training (b) Validation

Fig. 5. Quality-estimation accuracies for training (Experiment 1) and validation data (Experiment 2)



(a) Training (b) Validation

Fig. 6. Quality-estimation accuracies for training (Experiment 2) and validation data (Experiment 1)

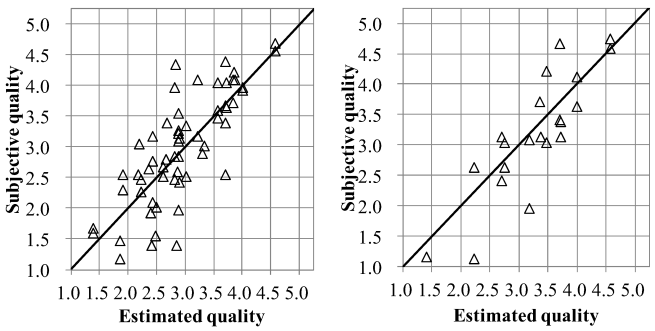
TABLE V
COEFFICIENTS OF PARAMETRIC QUALITY-ESTIMATION MODEL

Coefficients	Value
a_1	5.00000
a_2	9.85820
a_3	1.03576
v_1	1.24070
v_2	342504
v_3	1.17954
v_4	0.000312265
v_5	0.996968
v_6	48.3152
av_1	0.000000
av_2	0.000000
av_3	0.0100822
av_4	0.193344

Coefficients	Value
t_1	0.0477514
t_2	0.0000747564
t_3	0.196471
t_4	0.0336057
t_5	0.00728420
s_1	5.27470
s_2	36555.1
s_3	1.16663

same for the cross-validation results even when the training data were changed. Therefore, we optimized the coefficients (a_1, a_2, a_3, v_1 to v_6, av_1 to av_4, t_1 to t_5, s_1, s_2, and s_3) of our model using all 82 PAVSs for Experiments 1 and 2. The coefficient values are listed in Table V.

The relationship between estimated quality and subjective quality in Experiments 1 and 2 is shown in Fig. 7. As listed in Table VI, the results indicated that quality-estimation accuracy was sufficient because the RMSE and PCC satisfied the minimum performance requirement.

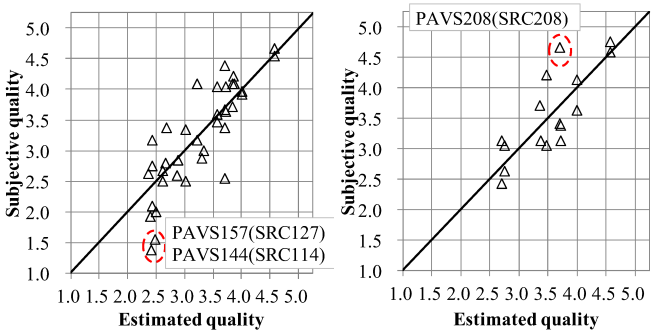


(a) Experiment 1 (b) Experiment 2

Fig. 7. Quality-estimation accuracies in Experiments 1 and 2

TABLE VI
RMSES AND PCCS IN EXPERIMENTS 1 AND 2

	All	Experiment 1	Experiment 2
RMSE	0.54	0.55	0.52
PCC	0.82	0.79	0.88



(a) Experiment 1 (b) Experiment 2

Fig. 8. Quality-estimation accuracies for compression and adaptation in Experiments 1 and 2

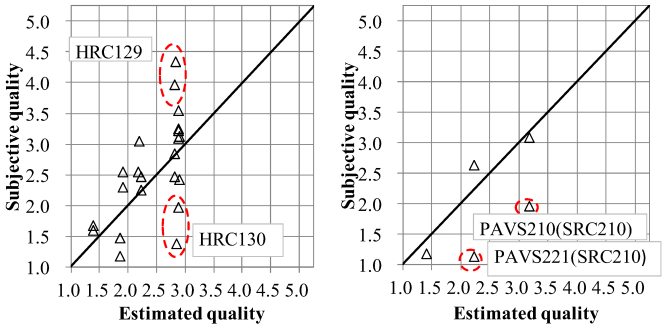
VI. CONSIDERATIONS

Some considerations should be noted to explain the results in detail.

As described in Section II, the following issues need to be addressed: a) the impact of the combination of video resolution, adaptation, and audiovisual interaction on audiovisual coding quality and b) the impact of content length on stalling quality. Therefore, we discuss the RMSEs for all PAVSs (also see Fig. 7), those without stalling, i.e., only compression conditions (Fig. 8), and those with stalling (Fig. 9). Table VII lists the RMSEs, and the number of PAVSs is indicated in parentheses.

These results demonstrated that quality-estimation accuracy was high for 54 PAVSs without the stalling event regardless of the content length and combination of video resolution, adaptation, and audiovisual interaction. However, the results indicated that the quality-estimation accuracy was not sufficiently high for 28 PAVSs with a stalling event.

We first explain that there were some plots with large estimation errors, i.e., estimation errors larger than 1.0, in PAVS147 (SRC114, HRC117), PAVS157 (SRC127, HRC104),



(a) Experiment 1 (b) Experiment 2

Fig. 9. Quality-estimation accuracies for stalling events in Experiments 1 and 2

TABLE VII
RMSEs FOR COMPRESSION AND STALLING

Type of PAVS	All	Experiment 1	Experiment 2
Compression and stalling	0.54(82)	0.55(60)	0.52(22)
Compression	0.45(54)	0.46(38)	0.43(16)
Stalling	0.68(28)	0.67(22)	0.71(6)
Stalling*	0.36(22)	0.39(18)	0.22(4)

*: PAVS129, PAVS132, PAVS130, PAVS131, PAVS210 and PAVS221 were not included in the calculation of RMSE.

and PAVS208 (SRC208, HRC208). As described in Section IV, two SRCs were assigned to an HRC. It is conceivable that our audiovisual integration/temporal module estimated the quality accurately because the quality estimation for one of the two SRCs was fine. However, it should be considered that our video quality estimation module could not evaluate the impact of SRC on the quality for some of the content. Although, by definition, our model cannot access the bitstream, if information about the bitstream (e.g., quantization parameter) is available, the quality estimation accuracy could be improved. We then explain why quality-estimation accuracy was low for PAVSs with a stalling event. As shown in Fig. 9, some scatter plots indicate low performance, i.e., estimation error larger than 1.0. Because the subjective quality was approximately 4.0 in PAVS129(HRC129) and PAVS132(HRC129) even when a stalling event occurred, we investigated their PAVSs. The video frame in which the stalling event occurred was not degraded owing to compression. Therefore, participants did not notice any degradation during the stalling event. However, the video frame where the stalling event occurred was degraded owing to compression in PAVS130(HRC130), PAVS131(HRC130), PAVS210(SRC210), and PAVS221(SRC210). Therefore, participants noticed a large amount of compression degradation, i.e., block noise, during the stalling event. From these investigations, it is conceivable that the impact of stalling events on quality can be almost ignored when one occurs with small compression degradation and that the impact of a stalling event on quality increases during an event with a large amount of compression degradation. However, by definition, such quality impacts cannot be evaluated with the proposed parametric quality-estimation model because it cannot take media signals and bitstreams as input. Incidentally, PAVS129, PAVS132, PAVS130, PAVS131, PAVS210, and PAVS221 were not included in the calculation of the RMSE; therefore, the RMSEs for the other 22 PAVSs are listed in Table VII. From the results, we found that the RMSE for the stalling event was not high when the extreme cases were removed.

From the evaluation of the quality estimation accuracy, it is conceivable that the media session quality can be modeled using Eqs. (1) to (11). Although the number of switches is not incorporated into the model, the quality estimation accuracy was sufficiently high. Therefore, it is conceivable that functions $AVQ(t)$ and $w_2(t)$ take into account the impact of switching on quality.

The impact of stalling on the media session quality is modeled using the exponential function; the total length of stalling events and the average interval between stalling events are linearly normalized by the content length. As described in Section IV, PVSs with 4- to 24-sec stalling lengths were used for both the 1-min and 3-min tests, and the quality estimation accuracy for the stalling events for both the 1-min and 3-min tests was high, with the exception of some extreme stalling cases, as indicated in Table VII. The media session quality was sometimes overestimated or underestimated. This issue depends on whether the video frame in which the stalling event occurred was degraded owing to compression. Therefore, it is conceivable that the media session quality can be modeled well by the exponential function; the total length of stalling events and the average interval between stalling events are linearly normalized by the content length.

VII. CONCLUSIONS

We have proposed a parametric quality-estimation model that can be applied to estimate the quality of adaptive-bitrate streaming services. We identified issues that need to be addressed from conventional studies regarding the impact of the combination of an arbitrary video resolution, bitrate adaptation, audiovisual interaction, stalling event, and content length on quality. We developed a parametric quality-estimation model for adaptive-bitrate streaming services to address these issues. We conducted subjective quality assessment tests to train our proposed model and verify its quality-estimation accuracy. The results demonstrated that the model can be used to estimate the quality of adaptive-bitrate streaming services with high quality-estimation accuracy; when a stalling event occurs and there is a slight or large amount of compression degradation, the quality-estimation error will be large.

The following issues require further study. The individual modules were not validated in terms of the quality estimation accuracy. Therefore, subjective tests are required to validate the individual modules. In addition, we assigned an SRC to a specific HRC because of the limitation of the subjective test duration; thus, the impact of SRC on the quality should be investigated. We used the Main profile of the H.264 codec in subjective tests, although the High profile is also used for TV sets. Therefore, we need to verify whether our model can be applied to quality estimation on the High profile used for TV sets. Although the number of switches and the difference in quality before and after the switches were not incorporated

into the model, it is possible to improve the quality estimation accuracy if both are incorporated. Therefore, we need to verify whether there is an improvement of the quality estimation accuracy for such a model. Our model was not validated for extremely short or long intervals between stalling events. Hence, subjective tests are required to investigate this issue. We developed a model for TV sets in this work. Thus, we need to verify whether our model can be applied to quality estimation on smartphones, i.e., smaller screens. Additionally, UHD video is becoming more popular. Therefore, our model needs to be extended to higher video resolution. As described in Section VI, there were some plots with a large error in terms of the quality-estimation accuracy. Media-based and/or bitstream-based quality-estimation models may be developed because these offer the possibility of improving the accuracy of quality estimation if media signal-based parameters (e.g., TI and SI) and/or bitstream information (e.g., quantization parameters and motion vectors) can be used as input for our model. The relationship between an advertisement and initial loading was out of our paper's scope. However, this relationship should be investigated in future work.

REFERENCES

[1] *Advanced Video Coding for Generic Audiovisual Services*, ITU-T Recommendation H.264, Feb. 2016.

[2] *High efficiency video coding*, ITU-T Recommendation H.265, Apr. 2015.

[3] Http live streaming. [Online]. Available: https://developer.apple.com/streaming/

[4] *Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats*, ISO/IEC 23009-1:2014, May 2014.

[5] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hossfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, Firstquarter 2015.

[6] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 20–27, Apr. 2012.

[7] W. Robitza, M. N. Garcia, and A. Raake, "At home in the lab: Assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, May 2015, pp. 1–6.

[8] M. N. Garcia, D. Dytko, and A. Raake, "Quality impact due to initial loading, stalling, and video bitrate in progressive download video services," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, Sept. 2014, pp. 129–134.

[9] N. Staelens, J. D. Meulenaere, M. Claeys, G. V. Wallendael, W. V. den Broeck, J. D. Cock, R. V. de Walle, P. Demeester, and F. D. Turck, "Subjective quality assessment of longer duration video sequences delivered over HTTP adaptive streaming to tablet devices," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 707–714, Dec. 2014.

[10] *Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference*, ITU-T Recommendation J.144, Mar. 2004.

[11] *Objective Perceptual Multimedia Video Quality Measurement in the Presence of a Full Reference*, ITU-T Recommendation J.247, Aug. 2008.

[12] *Objective Perceptual Multimedia Video Quality Measurement of HDTV for Digital Cable Television in the Presence of a Full Reference*, ITU-T Recommendation J.341, Jan. 2011.

[13] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sept. 2004.

[14] P. L. Callet, C. Viard-Gaudin, S. Pechard, and E. Caillault, "No reference and reduced reference video quality metrics for end to end QoS monitoring," *IEICE Trans. on Commun.*, vol. 89, no. 2, pp. 289–296, Feb. 2006.

[15] Y. Xue, B. Erkin, and Y. Wang, "A novel no-reference video quality metric for evaluating temporal jerkiness due to frame freezing," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 134–139, Jan. 2015.

[16] M. C. Q. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements," in *IEEE International Conference on Image Processing 2005*, vol. 3, Sept. 2005, pp. III–141–4.

[17] F. Yang, S. Wan, Y. Chang, and H. R. Wu, "A novel objective no-reference metric for digital video quality assessment," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 685–688, Oct. 2005.

[18] *Perceptual Audiovisual Quality Measurement Techniques for Multimedia Services over Digital Cable Television Networks in the Presence of a Reduced Bandwidth Reference*, ITU-T Recommendation J.246, Aug. 2008.

[19] T. Yamada, Y. Miyamoto, Y. Senda, and M. Serizawa, "Video-quality estimation based on reduced-reference model employing activity-difference," *IEICE Trans. Commun.*, vol. 92-A, no. 12, pp. 3284–3290, Dec. 2009.

[20] *Parametric non-intrusive bitstream assessment of video media streaming quality*, ITU-T Recommendation P.1202, Oct. 2012.

[21] *Parametric non-intrusive bitstream assessment of video media streaming quality – lower resolution application area*, ITU-T Recommendation P.1202.1, Oct. 2012.

[22] *Parametric non-intrusive bitstream assessment of video media streaming quality – higher resolution application area*, ITU-T Recommendation P.1202.2, May 2013.

[23] O. Verscheure, X. Garcia, G. Karlsson, and J. P. Hubaux, "User-oriented QoS in packet video delivery," *IEEE Network*, vol. 12, no. 6, pp. 12–21, Nov. 1998.

[24] K. Watanabe, K. Yamagishi, J. Okamoto, and A. Takahashi, "Proposal of new QoE assessment approach for quality management of IPTV services," in *2008 15th IEEE International Conference on Image Processing*, Oct. 2008, pp. 2060–2063.

[25] S. O. Lee, K. S. Jung, and D. G. Sim, "Real-time objective quality assessment based on coding parameters extracted from H.264/AVC bitstream," *IEEE Trans. Consum. Electron.*, vol. 56, no. 2, pp. 1071–1078, May 2010.

[26] A. P. C. da Silva, P. Rodriguez-Bocca, and G. Rubino, "Optimal quality-of-experience design for a P2P multi-source video streaming," in *2008 IEEE International Conference on Communications*, May 2008, pp. 22–26.

[27] L. Anegekuh, L. Sun, E. Jammeh, I. H. Mkwawa, and E. Ifeachor, "Content-based video quality prediction for hevc encoded videos streamed over packet networks," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1323–1334, Aug. 2015.

[28] *Parametric non-intrusive assessment of audiovisual media streaming quality*, ITU-T Recommendation P.1201, Oct. 2012.

[29] *Parametric non-intrusive assessment of audiovisual media streaming quality - lower resolution application area*, ITU-T Recommendation P.1201.1, Oct. 2012.

[30] K. Yamagishi and S. Gao, "Light-weight audiovisual quality assessment of mobile video: ITU-T Rec. P.1201.1," in *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, Sept. 2013, pp. 464–469.

[31] *Parametric non-intrusive assessment of audiovisual media streaming quality - higher resolution application area*, ITU-T Recommendation P.1201.2, Oct. 2012.

[32] M. N. Garcia, P. List, S. Argyropoulos, D. Lindegren, M. Pettersson, B. Feiten, J. Gustafsson, and A. Raake, "Parametric model for audiovisual quality assessment in IPTV: ITU-T Rec. P.1201.2," in *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, Sept. 2013, pp. 482–487.

[33] K. Yamagishi and T. Hayashi, "Non-intrusive packet-layer model for monitoring video quality of IPTV services," *IEICE Trans. Fundamentals*, vol. 92-A, no. 12, pp. 3297–3306, Dec. 2009.

[34] J. Gustafsson, G. Heikkila, and M. Pettersson, "Measuring multimedia quality in mobile networks with an objective parametric model," in *2008 15th IEEE International Conference on Image Processing*, Oct. 2008, pp. 405–408.

[35] A. Raake, M. N. Garcia, S. Moller, J. Berger, F. Kling, P. List, J. Johann, and C. Heidemann, "T-V-model: Parameter-based prediction of IPTV quality," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2008, pp. 1149–1152.

[36] J. Joskowicz and J. C. L. Ardao, "A general parametric model for perceptual video quality estimation," in *2010 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR 2010)*, June 2010, pp. 1–6.

[37] *Amendment 2: New Appendix III – Use of ITU-T P.1201 for non-adaptive, progressive download type media streaming*, ITU-T Recommendation P.1201 Amendment, Dec. 2013.

[38] *Requirements for an objective perceptual multimedia quality model*, ITU-T Recommendation J.148, May 2003.
[39] D. S. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, Dec. 2004.
[40] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 973–980, Oct. 2006.
[41] T. Hayashi, G. Kawaguti, J. Okamoto, and A. Takahashi, "Subjective quality estimation model for video streaming services with dynamic bit-rate control," *IEICE Trans. Commun.*, vol. 89, no. 2, pp. 297–303, Feb. 2006.
[42] Y. Shen, Y. Liu, Q. Liu, and D. Yang, "A method of QoE evaluation for adaptive streaming based on bitrate distribution," in *2014 IEEE International Conference on Communications Workshops (ICC)*, June 2014, pp. 551–556.
[43] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC," in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, Jan. 2012, pp. 127–131.
[44] Y. Liu, S. Dey, D. Gillies, F. Ulupinar, and M. Luby, "User experience modeling for dash video," in *2013 20th International Packet Video Workshop*, Dec. 2013, pp. 1–8.
[45] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, "Deriving and validating user experience model for dash video streaming," *IEEE Trans. Broadcast.*, vol. 61, no. 4, pp. 651–665, Dec. 2015.
[46] *Subjective Video Quality Assessment Methods for Multimedia Applications*, ITU-T Recommendation P.910, Apr. 2008.
[47] *Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in any Environment*, ITU-T Recommendation P.913, Mar. 2016.

Kazuhisa Yamagishi received his B.E. degree in Electrical Engineering from the Tokyo University of Science in 2001 and his M.E. and Ph.D. degrees in Electronics, Information, and Communication Engineering from Waseda University in Japan in 2003 and 2013, respectively. He joined NTT Laboratories in 2003. He has been engaged in the development of objective quality-estimation models for multimedia telecommunications. From 2010 to 2011, he was a Visiting Researcher at Arizona State University. He received the Young Investigators' Award (IEICE) in Japan in 2007 and the Telecommunication Advancement Foundation Award in Japan in 2008.

Takanori Hayashi received his B.E., M.E., and Ph.D. degrees in Engineering from the University of Tsukuba, Ibaraki in 1988, 1990, and 2007, respectively. He joined NTT Laboratories in 1990 and has been engaged in quality assessment of multimedia telecommunication and network performance measurement methods. Currently, he is the Manager of the Service Assessment Group at NTT Laboratories. He received the Telecommunication Advancement Foundation Award in Japan in 2008 and the Telecommunication Technology Committee Award in Japan in 2012.