PAPER Special Section on Quality of Communication Networks and Services

Parametric Packet-Layer Model for Evaluation Audio Quality in Multimedia Streaming Services

Noritsugu EGI[†a)], **Takanori HAYASHI**[†], ***and*** **Akira TAKAHASHI**[†], ***Members***

SUMMARY We propose a parametric packet-layer model for monitoring audio quality in multimedia streaming services such as Internet protocol television (IPTV). This model estimates audio quality of experience (QoE) on the basis of quality degradation due to coding and packet loss of an audio sequence. The input parameters of this model are audio bit rate, sampling rate, frame length, packet-loss frequency, and average burst length. Audio bit rate, packet-loss frequency, and average burst length are calculated from header information in received IP packets. For sampling rate, frame length, and audio codec type, the values or the names used in monitored services are input into this model directly. We performed a subjective listening test to examine the relationships between these input parameters and perceived audio quality. The codec used in this test was the Advanced Audio Codec-Low Complexity (AAC-LC), which is one of the international standards for audio coding. On the basis of the test results, we developed an audio quality evaluation model. The verification results indicate that audio quality estimated by the proposed model has a high correlation with perceived audio quality.

key words: *audio, quality evaluation, parametric packet-layer model, multimedia streaming, IPTV, AAC*

1. Introduction

Multimedia streaming services have recently become increasingly important in the global market. One popular multimedia streaming service is Internet protocol television (IPTV), which delivers streaming multimedia such as audio, video, text, and graphics data over IP-based networks providing the required level of quality of experience (QoE). To provide high audio and video qualities for users, QoE monitoring for accurately managing QoE is indispensable. The most accurate method for evaluating QoE is subjective testing; however, it is time-consuming and expensive. For this reason, service providers and network providers demand an objective quality estimation method. The QoE of multimedia streaming services mainly consists of the quality of the audio and video components of the multimedia streaming. Audio quality and video quality have different effects on QoE. Therefore, audio and video qualities should be estimated separately for accurate estimation of QoE. We propose an objective QoE evaluation model for monitoring only the audio quality of multimedia streaming services.

Objective audio quality estimation methods can be categorized into three types [1]. The first type is the media-layer model, which estimates QoE by using speech/audio signals. The most widely used media-layer model for speech is ITU-T Recommendation P.862 "Perceptual Evaluation of Speech Quality (PESQ)" [2]. This model is used for estimating perceived speech quality of signals with a 3.4-kHz bandwidth without any prior knowledge about the configuration of terminals/networks, e.g., codec and packet-loss type. ITU-T Recommendation P.862.2 "Wideband Extension to PESQ" [3] is intended for assessing the speech quality of signals with a 7-kHz bandwidth in wideband telephone services. Furthermore, there are models that are applicable to audio signals with an audio bandwidth of up to 20 kHz [4], [5]. However, these models are often inconvenient for monitoring QoE because they need to directly obtain input and output speech signals and they require a large amount of computational load.

The second type, the parametric planning model, takes quality parameters as its inputs and formulates the relationships between each quality parameter and the QoE; therefore, this model is convenient for designing QoE. The most widely used model for 3.4-kHz speech is ITU-T Recommendation G.107 "E-model" [6]. Furthermore, there are models that are applicable to speech signals with a bandwidth of up to 7 kHz [7]. However, these models are only made for designing QoE, not for QoE monitoring.

The third type, the parametric packet-layer model is used to estimate QoE or calculate parameters that correspond to QoE using IP packet information that excludes audio-related payload information [8], [9]. This model can measure in-service QoE from quality parameters based on packet-header information and requires a low computational load. Therefore, this model is convenient for QoE monitoring. ITU-T standardized Recommendation P.564 [10] determines performance criteria for parametric packet-layer models in 3.1-kHz or 7-kHz telephony applications. However, there is no appropriate parametric packet-layer model for a multimedia streaming service that includes 20-kHz audio.

From the viewpoint of monitoring QoE, we present a parametric packet-layer model for audio quality of multimedia streaming services. Some audio parameters that are fluctuating when a service is provided are measured from information in the headers of received IP packets. We formulated relationships between the above audio quality parameters and perceived audio quality based on the results of subjective listening tests.

The remainder of this paper is organized as follows. In Sect. 2, we propose a framework of a parametric packet-

Manuscript received October 23, 2009.
Manuscript revised February 1, 2010.

[†]The authors are with NTT Service Integration Laboratories, NTT Corporation, Musashino-shi, 180-8585 Japan.
a) E-mail: egi.noritsugu@lab.ntt.co.jp

DOI: 10.1587/transcom.E93.B.1359

layer audio-quality estimation model. In Sect. 3, we explain the conditions of the subjective listening test performed to examine the relationships between the quality parameters and perceived audio quality, and we present a quality-estimation model derived from the subjective listening test results. In Sect. 4, we present a verification of the accuracy of our quality-estimation model. We discuss another subjective quality experiment to investigate the consistency between subjective quality and its objective estimation using the proposed model. In Sect. 5, we present our conclusions.

2. Framework of Parametric Packet-Layer Model for Evaluating Audio Quality

In multimedia streaming, there are IP packets that do not contain audio data. The loss of these IP packets does not affect the output of audio data; therefore, we only need to use IP packets that contain audio data. The judgment of whether IP packets contain audio data is based on the information in each packet header, e.g., the payload type in an RTP header or the packet identification in a TS header. These show what is located in the payload of each packet. Henceforth, we call IP packets that contain audio data "IP packets."

A framework of our audio quality evaluation model is shown in Fig. 1. A quality evaluation model contributing to QoE monitoring needs to take into account parameters that may fluctuate during multimedia streaming, such as packet loss and coding rate (e.g., by automatic bitrate adjustment depending on network condition). Therefore, these parameters are measured from packet header information in received IP packets. We use the pre-determined sampling rate, frame length, and audio codec type for each multimedia streaming service. For example, one audio bit rate usually corresponds to one sampling rate, and the audio codec type and frame length, which is the time corresponding to one IP packet, used in one multimedia streaming service are determined. In general, delay and jitter are not problematic due to the dejitter buffer, for example in STB. Therefore, this model does not consider degradation due to delay or jitter.

Quality degradation due to coding is obtained from a database of quality based on the audio bit rates and the sampling rates. We obtained quality degradation due to packet loss as follows. Packet-loss rate is a common index of quality degradation due to packet loss. However, we divided the packet-loss rate into a packet-loss frequency and average burst length of the packet-loss block, where one packet-loss block is one group of consecutive lost IP packets, packet-loss frequency is the number of packet-loss blocks in a multimedia sequence, and the average burst length is the average number of IP packets for each packet loss block. The purpose of this division is to investigate quality degradation caused by loss frequency and loss length. Audio quality is calculated on the basis of the relationships between these quality degradations and perceived audio quality.

We provide a model of a quality evaluation module for packet-loss. As a quality estimation module for coding, we use a database derived from the results of subjective listening tests. In the next section, we explain the subjective listening test that was performed to examine the relationships between these audio quality parameters and perceived audio quality.

3. Audio Quality Estimation Model

We performed a subjective listening test necessary for constructing the audio quality evaluation model. In this test, we assume that transmission is over RTP/UDP/IP and low bit rate coding is used such as mobile streaming services. From the test results, we developed an audio-quality estimation model.

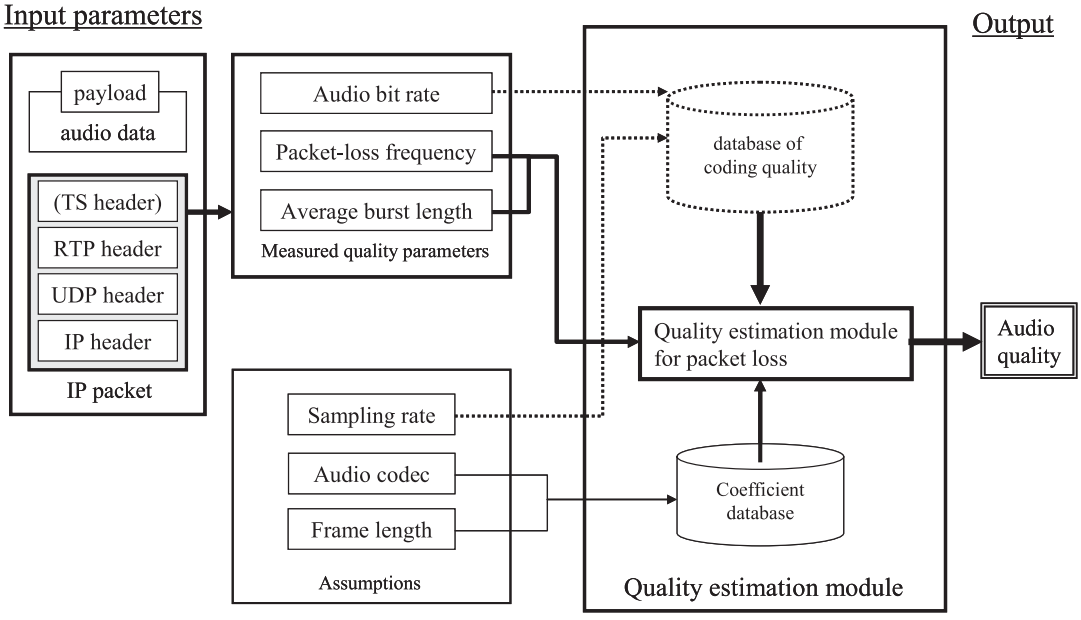


Fig. 1 Framework of parametric packet-layer model for evaluating audio quality.

Table 1 Conditions of subjective listening test.

No.	Audio bit rate [kbps]	Sampling rate [kHz]	Frame Length [ms]	Packet-loss frequency	Burst packet-loss length
1	16	8	128	0	-
2	32	8	128	0	-
3	32	11.025	92.88	0	-
4				1	1
5					4
6					8
7				2	1
8					4
9				4	1
10				8	1
11	32	16	64	0	-
12	32	22.05	92.88	0	-
13				1	1
14					4
15					8
16				2	1
17					4
18				4	1
19				8	1
20	48	16	64	0	-
21	48	22.05	92.88	0	-
22				1	1
23					4
24					8
25				2	1
26					4
27				4	1
28				8	1
29	48	32	96	0	-
30	64	16	64	0	-
31				1	1
32					4
33					8
34				2	1
35					4
36				4	1
37				8	1
38	64	22.05	92.88	0	-
39				1	1
40					4
41					8
42				2	1
43					4
44				4	1
45				8	1
46	64	32	96	0	-
47				1	1
48					4
49					8
50				2	1
51					4
52				4	1
53				8	1
54	64	44.1	92.88	0	-
55				1	1
56					4
57					8
58				2	1
59					4
60				4	1
61				8	1
62	16 bit Linear PCM	48	-	-	-

3.1 Test Conditions of Subjective Listening Test

The experimental conditions are listed in Table 1. First, we downsampled and encoded multimedia sequences at each sampling rate and audio bit rate. The combinations of audio bit rate and sampling rate depend on the multimedia streaming service. Therefore, we prepared various combinations. The audio codec was AAC-LC [11], which is used for audio streaming, audio downloading, and broadcasting systems, such as, Japanese broadcasting systems. Next, these multimedia sequences were transmitted through a network emulator over RTP/UDP/IP. Then, we controlled the IP packet loss and its burst length for the multimedia sequences at each bit rate. The packet-loss rate was set as the ratio of (packet-loss frequency) x (burst packet-loss length) to the total number of IP packets. Finally, we captured the decoded audio at a client terminal and adjusted the root mean square (RMS) level of the captured audio to −26 dBov. In this test, eight different audio sequences at a 48-kHz sampling rate with a 16-bit resolution (10 s each) were used, as listed in Table 2.

Table 2 Types of audio content.

Content	Speech	Audio
Evening news program	1 male	none
Weather news program	1 female	background music
Talk show	2 females	none
Cooking program	1 male and 1 female	background music + sizzling sound
TV commercial	2 males and 1 female	background music
Modern music program	1 male	pop music
Opera program	1 female	orchestra
Classical music program	none	violin

In the subjective quality assessment, the subjects listened to a reference sample (10 s) and a test sample (10 s) for the same content through a pair of headphones (Senheiser HD 30). The separation between the two samples was 1 s. After that, subjects compared the test sample for the reference sample using a degradation category rating (DCR) method [12]. The subjects were instructed to rate the conditions according to the five-point degradation category scale, as listed in Table 3. The quality descriptions for individual rating categories were given in Japanese. The subjects were allowed to listen to each sample only once, and the order of test samples was random. The listening level was adjusted to −18 dB Pa (76 dB SPL) at each ear reference point for 1-kHz sine wave at a level of −26 dBov. The listening room was soundproof. Before starting the evaluation of test samples, the subjects evaluated samples that were of good, bad, and average quality. The purpose of this session was to ensure that subjects understood the tasks and the quality range of the test samples. Forty non-expert subjects aged 19–49 participated in the experiment. The number of test samples was 496. The subjects repeated a 20-minute evaluation and a 20-minute break. The total duration was about 240 minutes in two days. This test was executed under audiovisual monitoring by two operators. For each condition, audio quality was represented as a Degradation Mean Opinion Score (DMOS). DMOS was simply averaged over 320 scores, which was the product of forty subjects and eight

Table 3 Five-point degradation-category scale.

5	Degradation is inaudible.
4	Degradation is audible but not annoying.
3	Degradation is slightly annoying.
2	Degradation is annoying.
1	Degradation is very annoying.

Table 4 DMOS values for test samples without packet loss.

Audio bit rate (kbps)	Sampling rate (kHz)	DMOS	Audio bit rate (kbps)	Sampling rate (kHz)	DMOS
16	8.000	**2.042**	64	16.000	**4.047**
32	8.000	**2.573**	64	22.050	**4.490**
32	11.025	**2.984**	64	32.000	**4.734**
32	16.000	**3.479**	64	44.100	**4.660**
32	22.050	**3.708**	16bit Linear PCM	44.100	**4.792**
48	16.000	**3.964**			
48	22.050	**4.323**			
48	32.000	**4.646**			

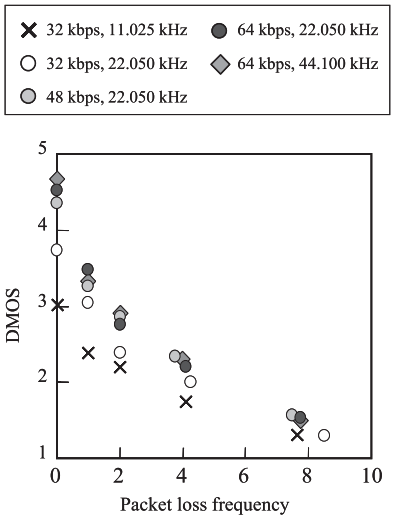


Fig. 2 Relationship between packet-loss frequency and DMOS.

audio sequences.

3.2 Quality Estimation Model

The DMOS values for the test samples without packet loss are listed in Table 4.

We investigated the degree of the quality degradation due to packet loss. First, we focused on test samples that had the same average burst length and frame length of audio sequence, thus excluding the effect of the difference in average burst length and frame length on the DMOS. Here, we use test samples with average burst length $ABL = 1$ and frame length $FL = 92.88$ (ms). The DMOS values of these test samples decreased exponentially with increasing packet-loss frequency, as shown in Fig. 2. When encoded quality without packet loss is higher, the degradation due to packet loss frequency has more effect on the DMOS. On the basis of this finding, we determined Eq. (1), which estimates the DMOS, where *Cq* (enCoded Quality without packet loss) is the DMOS for an audio sequence without packet loss (based on Table 4), *PLF* is packet-loss frequency, and a_1, a_2, and a_3 are constants that are determined by nonlinear least squares fitting (NLSF) to minimize the error between the estimated and measured DMOSs with respect to *PLF*.

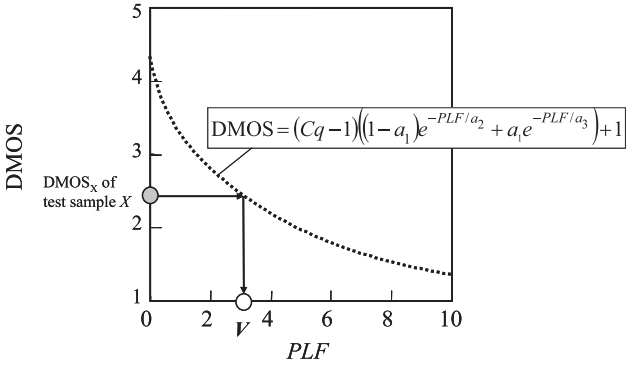


Fig. 3 Measurement of the value *V* based on Eq. (1).

$$\mathrm{DMOS} = (Cq-1)\left((1-a_1)\,e^{-PLF/a_2} + a_1 e^{-PLF/a_3}\right)+1 \quad (1)$$

Next, we consider an estimated equation focused on test samples with an average burst length $ABL = Y$ (> 1). For test samples that have the same conditions excluding *ABL*, when *ABL* is larger, DMOS is smaller. To represent quantitively a degree of degradation by one IP packet loss block with $ABL = Y$, we defined a virtual packet loss block *V* for each value of ABL. We set a value of *V* for $ABL = Y$, where the audio quality with $PLF = V$ and $ABL = 1$ and audio quality with $PLF = 1$ and $ABL = Y$ are the same. To calculate a value of *V* for $ABL = Y$, we used Eq. (1). For example, if DMOS of test sample X with $PLF = 1$ and $ABL = Y$ was $\mathrm{DMOS_X}$, we calculated *V* based on Eq. (1), as illustrated in Fig. 3, where the audio quality with $PLF = V$ and $ABL = 1$ is the same as $\mathrm{DMOS_X}$. We assumed that a degree of degradation by *p* IP packet-loss blocks with $ABL = Y$ is the same as that by $V \times p$ IP packet-loss blocks with $ABL = 1$ if a virtual packet loss block for $ABL = Y$ is *V*. On the basis of this supposition, when DMOS of test sample X' with $PLF = p$ (> 1) and $ABL = Y$' was $\mathrm{DMOS_{X'}}$, we calculated *V*' for $ABL = Y$' as follows. We calculated *V*, as illustrated in Fig. 3, where the audio quality with $PLF = V$ and $ABL = 1$ is the same as $\mathrm{DMOS_{X'}}$. Then, *V* is the same value as $V' \times p$ from the first supposition, and we calculated $V' (= V/p)$ for $ABL = Y$'. We calculated *V* for each condition like this. The relationship between *V* and the average burst length *ABL* is shown in Fig. 4. From this figure, we approximated $V = f(ABL)$ by linear approximation as Eq. (2), where a_4 is a constant determined by NLSF to minimize the error between the estimated and measured *V*s with respect to *ABL*.

$$V = a_4(ABL - 1) + 1 \quad (2)$$

From the supposition that a degree of degradation by *p* IP packet-loss blocks with $ABL = Y$ is the same as that by $V \times p$ IP packet-loss blocks with $ABL = 1$ if a virtual packet loss block for $ABL = Y$ is *V*, we calculated the DMOS of

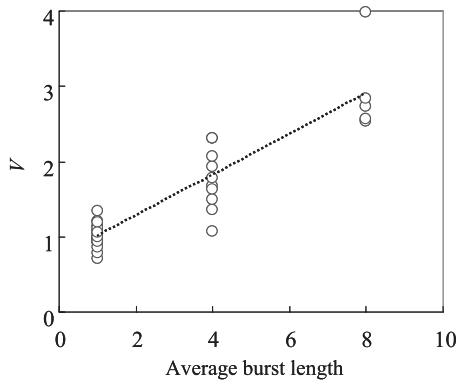


Fig. 4 Relationship between averaged burst length and V.

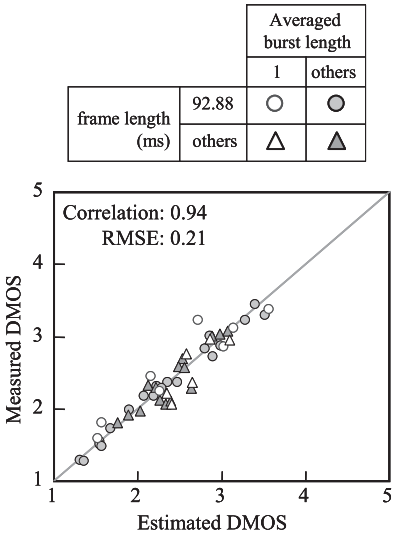


Fig. 5 Relationship between estimated DOMS and measured DMOS.

test samples with $ABL > 1$by Eq. (3), where we calculated V from ABL by Eq. (2). When ABL is 1, V is 1, that is, Eqs. (1) and (3) are the same. Therefore, we use only Eq. (3) regardless of ABL.

$$\text{DMOS} = (Cq-1)\left((1-a_1)\,e^{-V\cdot PLF/a_2}+a_1 e^{-V\cdot PLF/a_3}\right)+1 \quad (3)$$

Furthermore, we consider an estimated equation focused on test samples in which frame length FL' is not FL. The frame length is the time corresponding to one IP packet. We assumed that the frame length and average burst length have the same characteristic that the degraded time of audio data will be k times if the value of the frame length or average burst length is k times. On the basis of this supposition, when we consider test samples, in which frame length FL' is not FL, we regard that ABL will be FL'/FL times for the condition that the frame length is FL. Then, we calculated V of test samples with FL' by Eq. (4). When FL' is the same as FL, Eqs. (2) and (4) are the same. Therefore, we use only Eq. (4) regardless of FL.

$$V = a_4\left(\left(\frac{FL'}{FL}\right)ABL - 1\right) + 1 \quad (4)$$

With our model, we evaluated audio quality in multimedia streaming services using Eqs. (3) and (4). The calculated value of each constant in this test is as follows, a_1 = 0.8392, a_2 = 0.3976, a_3 = 4.6126, and a_4 = 0.5080.

3.3 Estimation Accuracy

The relationship between the measured DMOS and its objective estimation by Eqs. (3) and (4) is shown in Fig. 5. The correlation between the measured DMOS and estimated DMOS was high (0.94), regardless of the frame length and average burst length. Therefore, we concluded that this estimation model works for a wide range of audio bit rates, sampling rates, frame lengths, packet-loss rates, and average burst lengths.

4. Verification

In this section, we verify the accuracy of our quality estimation model using subjective listening tests. In verification test A, we verified our model as being able to evaluate unknown data with different conditions to those discussed in Sect. 3. In this test, test samples were constructed using the method described in Sect. 3.1. In verification test B, we verified our model as being able to evaluate audio data with a multimedia streaming system using a different protocol stack MPEG2-TS/RTP/UDP/IP. Encoded audio bitstream is stored in a TS packet that is smaller than an RTP packet. When a TS packet is smaller than a bitstream for one block size of encoding process, this bitstream is divided into more than one TS packet. In protocol stack RTP/UDP/IP, an RTP packet can contain a bitstream for one or more block sizes of the encoding process. This is a different characteristic from that in Sect. 3.1. In test B, we verify the validity of our model for this point.

4.1 Verification Test A

In this test, the experimental conditions and audio sequences were different from those of the test in Sect. 3. The experimental conditions are listed in Table 5.

Eight different audio sequences at a 48-kHz sampling rate with a 16-bit resolution (10 s each) were used (Table 6). We constructed test samples from multimedia sequences using the packet-loss simulation described in Sect. 3.1. Twenty-four non-expert subjects (aged 21–48) participated. The number of test samples was 248, and the total duration was about 120 minutes. Other experimental conditions of this test were the same as those described in Sect. 3.1.

The relationships between the subjective and objective DMOS calculated using Eqs. (3) and (4) are shown in Fig. 6. For quality estimation, Cq was selected from the corresponding values in Table 4, and the constants a_1, a_2, a_3, and a_4 were the same values estimated in Sect. 3.2. High correlation (0.98) was achieved between the subjective DMOS

Table 5 Conditions of verification test A.

No.	Audio bit rate [kbps]	Sampling rate [kHz]	Frame length [ms]	Packet-loss rate [%]	Burst packet-loss length
1	16	8	128	0.0	-
2				2.0	1
3				4.0	1
4				8.0	1
5				10.0	8
6					1-8
7	32	8	128	0.0	-
8				2.0	1
9				4.0	1
10				8.0	1
11				10.0	8
12					1-8
13	32	16	64	0.0	-
14				1.0	1
15				2.0	1
16				4.0	1
17				10.0	8
18					1-8
19	48	16	64	0.0	-
20				1.0	1
21				2.0	1
22				4.0	1
23				10.0	8
24					1-8
25	64	32	96	0.0	-
26				1.5	1
27				3.0	1
28				6.0	1
29				10.0	8
30					1-8
31	16 bit Linear PCM	48	-	-	-

Table 6 Types of audio content.

Content	Speech	Audio
Evening news program	1 male	blowing wind
Talk show	2 males	background music + car
Documentary	1 male	background music + train
Sports program	1 male	background music + karate
Shopping program	1 male	background music
Modern music program	1 male	pop music
Classical music program	none	orchestra
Instrumental music program	none	piano

and objective DMOS. The measured DMOS was slightly larger than the estimated DMOS. The reason is that there were many samples with a high packet loss rate in this test. Therefore, subjects were relatively tolerant of bad quality samples compared with the subjects discussed in Sect. 3. We consider that this difference can be revised by including the same sample criteria, such as a modulated noise reference unit (MNRU) [13], in all subjective tests.

4.2 Verification Test B

In this test, we focused on high bit rate coding such as High Definition Television (HDTV) services. We verified whether our proposed model can be used for accurately evaluating audio quality by executing two subjective listening tests B1 and B2. The purpose of test B1 is to compile a new set of Cq, a_1, a_2, a_3, and a_4. The purpose of test B2 is to verify its validity for a different protocol stack with unknown data by the using the same values of Cq, a_1, a_2, a_3, and a_4 as those of B1.

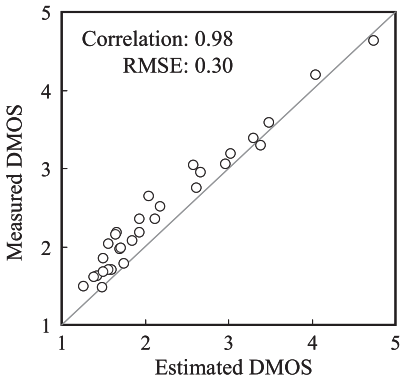


Fig. 6 Relationship between estimated DOMS and measured DMOS for verification test A.

We now explain subjective listening test B1. We encoded multimedia sequences at each audio bit rate. The audio codec was AAC-LC. Next, we passed these multimedia sequences over MPEG2- TS/RTP/UDP/IP. Then, we controlled the IP packet loss and its burst length for the multimedia sequences at each bit rate by using a network emulator. The packet-loss rate was set as the ratio of (packet-loss frequency) x (burst packet-loss length) to the total number of IP packets. Finally, we captured the decoded audio. The experimental conditions are listed in Table 7. In this case, a bitstream for one block size of the encoding process is divided into more than one TS packet. We fix the frame length to 21.3 ms, which is the time duration for one codec block size. In these tests, eight different audio sequences at a 48-kHz sampling rate with a 16-bit resolution (10 s each) were used, as listed in Table 2. Forty non-expert subjects aged 19–49 participated in the experiment. The number of test samples was 216, and the total duration was about 100 minutes. Other experimental conditions of this test were the same as those described in Sect. 3.1.

Next, we executed subjective listening test B2 to verify whether this model can be used for estimating quality accurately for unknown data. The experimental conditions are listed in Table 8, and eight different audio sequences were used, as listed in Table 6. The frame length of all conditions was 21.33 ms. We made test samples and obtained the DMOS of each test sample similar to subjective listening test B1. Twenty-four non-expert subjects (aged 21–48) participated. The number of test samples was 200, and the total duration was about 100 minutes. Other experimental conditions of this test were the same as those described in Sect. 3.1

After that, we determined values in the quality database and constants of Eqs. (3) and (4) based on the results of subjective listening test B1. The DMOS values of test samples with more than 80 kbps had no significant difference

Table 7 Conditions of subjective listening test B1.

No.	Audio bit rate [kbps]	Sampling rate [kHz]	Packet-loss Frequency	Burst packet-loss length
1	64	48	0	-
2			1	1
3			2	
4			4	
5			8	
6	80	48	0	-
7			1	1
8			2	
9			4	
10			8	
11	96	48	0	-
12			1	1
13			2	
14			4	
15			8	
16	192	48	0	-
17			1	1
18			2	
19			4	
20			8	
21			1	20
22			2	
23			4	
24			1	30
25			2	
26			4	
27	16 bit Linear PCM	48	-	-

Table 8 Conditions of subjective listening test B2.

No.	Audio bit rate [kbps]	Sampling rate [kHz]	Packet-loss rate [%]	Burst packet-loss length
1	80	48	0.00	-
2			0.15	1
3			0.30	
4			0.60	
5			1.20	
6			0.10	15
7			0.20	
8			0.40	
9	128	48	0.00	-
10			0.10	1
11			0.20	
12			0.40	
13			0.80	
14			0.10	15
15			0.20	
16			0.40	
17	256	48	0.00	-
18			0.05	1
19			0.10	
20			0.20	
21			0.40	
22			0.10	15
23			0.20	
24			0.40	
25	16 bit Linear PCM	48	-	-

with 95% confidence compared to the DMOS value of "16-bit Linear PCM" (No 45). Therefore, we used the DMOS value of "16-bit Linear PCM" as a quality database for more than 80 kbps, $Cq = 4.794$ in Eq. (3). The other constants of Eqs. (3) and (4) were determined based on the relationship among the DMOS and the measured parameters of "Packet-loss frequency" and an "Average burst length," as described in Sect. 3.2, where $a_1 = 0.5145$, $a_2 = 1.6658$, $a_3 = 5.0921$, and $a_4 = 0.0560$.

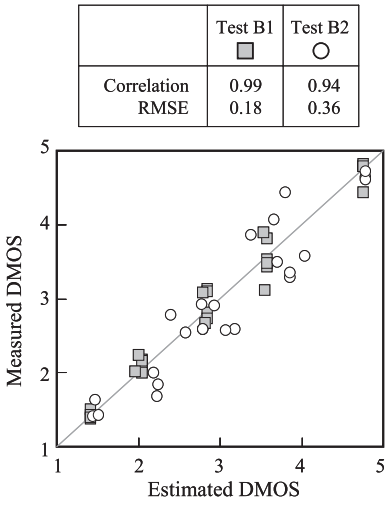


Fig. 7 Relationship between estimated DOMS and measured DMOS for verification test B.

Finally, by using these values and Eqs. (3) and (4), we calculated the objective DMOS for subjective listening tests B1 and B2. For quality estimation in test B2, Cq and the constants a_1, a_2, a_3, and a_4 were the same values calculated in test B1. The relationship between the subjective DMOS and the objective DMOS is shown in Fig. 7. High correlation (0.99 and 0.94) was achieved between the subjective DMOS and objective DMOS for both tests. However, the root mean square error (RMSE) for test B2 was relatively large. This occurred by the setting of the frame length. In this test, the bitstream for one block size of the encoding process is divided into several TS packets. Since the TS packet size is fixed, a ratio of the bitstream in one TS packet to the bitstream for one block size of the encoding process is small when the audio bit rate is higher. Therefore, degradation by loss of one IP packet is higher when the audio bit rate is smaller. However, our model does not take account of the audio bit rate for packet loss degradation. This seems to cause the error. Future work is to improve the accuracy of quality evaluation through the appropriate setting of the frame length using the bit rate and information of the IP packet structure.

5. Conclusion

To monitor audio quality for multimedia streaming services, we proposed an objective parametric packet-layer model for estimating perceived audio quality from audio bit rate, sampling rate, frame length, packet-loss frequency, and average burst length.

First, we proposed the framework of an audio quality estimation model for audio quality monitoring. We determined this framework on the basis of the IP packet protocol structure, measurable information during transmission

of multimedia data, and low complexity of measurement. Furthermore, this model is easy to be implemented because we need to measure only audio bit rate, packet-loss frequency, and average burst length as input parameters during multimedia streaming.

Next, we applied this model for quality monitoring in a multimedia streaming system using AAC-LC. From the results, we determined equations for estimating the DMOS of an audio sequence. These equations are based on the relationships among these parameters and perceived quality, and we showed that the estimated DMOS has a high correlation to the measured DMOS. We verified the accuracy of the proposed model for unknown data using the same multimedia streaming system. High correlation (0.98) between the subjective DMOS and objective DMOS shows that the proposed model can estimate audio quality accurately by obtaining the values of the constants of the equations in this model.

Finally, we verified the accuracy of the proposed model for multimedia streaming system using different protocol stack. We executed two listening test to obtain the value of the constants and examined the accuracy of the estimation for unknown data, respectively. These results also showed a high correlation (0.94) between the subjective DMOS and objective DMOS. Hence, we conclude that the proposed model can be used for monitoring audio quality from a viewpoint of estimation accuracy.

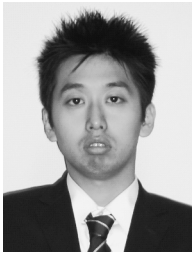
In the future, we would like to validate the proposed model for audio sequences coded by different audio codecs.

Acknowledgments

This study was partly supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications, Japan under project 073103002.

References

[1] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," IEEE Commun. Mag., vol.42, no.7, pp.28–34, July 2004.
[2] ITU-T Recommendation 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.
[3] ITU-T Recommendation P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Nov. 2007.
[4] ITU-R Recommendation BS.1387, "Method for objective measurement of perceived audio quality," Dec. 1998.
[5] T. Kanamori, Y. Higurashi, and N. Kitawaki, "Objective quality assessment of audio coding, taking account of coding and packet loss," IEICE Technical Report, CQ2006-91, Jan. 2007.
[6] ITU-T Recommendation G.107, "The E-model, a computational model use in transmission planning," Aug. 2008.
[7] ITU-T Recommendation G.107 Appendix II, "Provisional impairment factor framework for wideband speech transmission," Aug. 2008.
[8] A. Clerk, "Modeling the effects of burst packet loss and recency on subjective voice quality," Proc. IPTEL, pp.123–127, April 2001.
[9] S.R. Broom, "VoIP quality assessment: Taking account of the edge-device," IEEE Trans. Audio Speech and Language Processing, vol.14, no.6, pp.1977–1983, Nov. 2006.
[10] ITU-T Recommendation P.564, "Conformance testing for narrow-band voice over IP transmission quality assessment models," July 2006.
[11] ISO/IEC 13818-7, "Generic coding of moving pictures and associated audio information—Part 7: Advanced audio coding (AAC)," Jan. 2006.
[12] ITU-T Recommendation P.800, "Method for subjective determination of transmission quality," Aug. 1996.
[13] ITU-T Recommendation P.810, "Modulated noise reference unit (MNRU)," Feb. 1996.



Noritsugu Egi received his B.E. and M.E. degrees in Electrical Communication Engineering from Tohoku University in 2003 and 2005. He joined NTT Service Integration Laboratories of the NTT R&D Center, Tokyo, Japan, in 2005. Currently, he is researching speech and audio quality assessment.

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. Since joining NTT Laboratories in 1990, he has been engaged in the subjective quality assessment of multimedia telecommunication and network performance measurement methods. He is currently working on a multi-modal quality assessment method over IP networks. He received the Telecommunication Advancement Foundation Award in Japan in 2007.

Akira Takahashi received a B.S. degree in mathematics from Hokkaido University in Japan in 1988, M.S. degree in electrical engineering from the California Institute of Technology in the U.S. in 1993, and Ph.D. degree in engineering from the University of Tsukuba in Japan in 2007. He joined NTT Laboratories in 1988 and has been engaged in the quality assessment of audio and visual communications. Currently, he is the Manager of the Service Assessment Group in NTT Service Integration Laboratories. He has been contributing to ITU-T Study Group 12 (SG12) on QoS, QoE, and Performance since 1994. He is a Vice-Chairman of ITU-T SG12, a Vice-Chairman of Working Party 3 in SG12, and a Co-Rapporteur of Question 13/12 for 2009–2012 Study Period. He received the Telecommunication Technology Committee Award in Japan in 2004 and the ITU-AJ Award in Japan in 2005. He also received the Best Tutorial Paper Award from IEICE in Japan in 2006, the Telecommunications Advancement Foundation Awards in Japan in 2007 and 2008.