INVITED PAPER *Special Section on Multimedia QoS Evaluation and Management Technologies*

Multimedia Quality Prediction Methodologies for Advanced Mobile and IP-Based Telephony

Nobuhiko KITAWAKI[†a)], ***Fellow***

SUMMARY This paper describes the author's perspective on multimedia quality prediction methodologies for multimedia communications in advanced mobile and internet protocol (IP)-based telephony, and reports related experiments and trials. First, the paper describes the need for perceptual QoS (Quality of Service) assessment in which various quality factors in multimedia communications for advanced mobile and IP-based telephony are analyzed. Then an objective quality prediction scheme is proposed from the viewpoints of quality measurement tools for each quality factor and an opinion model for compound quality factors in mobile and IP-based communications networks. Finally, the author's current trials of measurement tools and opinion models are described.

key words: *perceptual QoS, quality prediction, opinion model, wideband speech, multimedia, hands-free, speech recognition and synthesis*

1. Introduction

The perceptual QoS (Quality of Service) for multimedia communications in advanced mobile and IP (Internet Protocol)-based telephony is very important today from the viewpoints of both customer satisfaction and provider management operations. Quality assessment for telephone-band communication has become well established, as evidenced by the number of publications and recommendations available [1]–[3]. However, multimedia communications in advanced mobile and IP-based telephony using PCs are entirely different from conventional telephone-band communication due to the requirement for hands-free and multimodal communication, which lead in turn to the need for wideband speech and multimedia including speech, audio and video.

Mobile communication technology has made significant progress, and is expected to be applied to vehicle communication, which requires hands-free operation. IP-based telephony via the Internet using PCs also requires hands-free communication. In both cases, wideband speech has become increasingly necessary, since hands-free communication using separate microphones and loudspeakers is indispensable, and in a hands-free situation, wideband speech is particularly helpful in enhancing the naturalness of communication.

It is very important to be able to accurately assess perceptual speech, audio, and video quality in multimedia services to ensure optimum communications system design, as well as effective transmission planning and management to satisfy customer requirements. This paper describes the author's perspective on multimedia quality prediction methodologies for multimedia communications in advanced mobile and IP-based telephony as well as current work on trials of these methodologies. This work is derived from proposals made to the International Telecommunications Union-Telecommunications Sector (ITU-T) Study Group 12 for telephone-band digital communications. The author first submitted proposals to this Group in 1981.

First, the paper describes the need for perceptual QoS assessment, in which various quality factors in multimedia communications of advanced mobile and IP-based telephony are analyzed. In the subjective QoS assessment method, perceptual QoS can usually be assessed by an opinion rating method. However, subjective assessment by opinion rating is time consuming and expensive. Therefore, the author aims to establish an objective QoS assessment methodology that correlates well with subjective QoS.

An objective quality prediction scheme is then proposed from the viewpoints of quality measurement tools for each quality factor and an opinion model for compound quality factors in multimedia communications of mobile and IP-based telephony. In particular, this paper describes objective QoS assessment for wideband coded speech distortion including discontinuous impairments such as IP packet loss, from the viewpoint of diagnostics, and it also describes an opinion model for transmission planning and monitoring from the viewpoint of various compound quality factors affecting perceptual QoS including wideband speech and video.

Finally, the author's current trials of measurement tools and opinion models are described. These include a wideband speech quality measure and artificial voice for 7 kHz wideband speech, an objective audio quality measure for CD (compact disk) music at various bandwidths, objective quality assessment methodologies for noise reduction algorithms in speech recognition-synthesis systems, echo canceller algorithms for hands-free communication, and comprehensive opinion models for audio-visual communications taking audio and video quality interaction into account.

2. View of Perceptual QoS Assessment for Optimum QoS Design and Reasonable QoS Management

This section describes the need for and approach to perceptual QoS assessment research in which various quality factors in multimedia communications of advanced mobile and

Manuscript received April 27, 2005.
Manuscript revised August 23, 2005.

[†]The author is with University of Tsukuba, Tsukuba-shi, 305-8573 Japan.
a) E-mail: kitawaki@cs.tsukuba.ac.jp

DOI: 10.1093/ietcom/e89–b.2.262

IP-based telephony are analyzed, especially a) wideband speech, audio, and video for multimedia, b) noise reduction and speech recognition synthesis for hands-free communication, and c) IP packet loss, acoustic echo, and pure delay for IP-based communications.

2.1 Necessity of Perceptual QoS Assessment

Reasonable and appropriate QoS assessment methodologies are required for QoS design and QoS management in multimedia communications of advanced mobile and IP-based communications systems from the viewpoints of both customer satisfaction and provider management operations. Perceptual QoS as evaluated by the customer is a very important input for service providers when designing and managing multimedia communications services.

Figure 1 shows the concept of perceptual QoS from the customer and provider viewpoints. From the customer's side, perceptual QoS can be measured by a subjective assessment method based on the customer's evaluation. From the provider's side, an objective QoS assessment using physical metrics that correlate closely with subjective QoS is desirable in order to ascertain the customer's evaluation of service quality. This information can then be used in the design and management of the communications terminals and networks. Of course, it is desirable for the objectively estimated QoS to be equivalent to the subjective evaluation of QoS by customers.

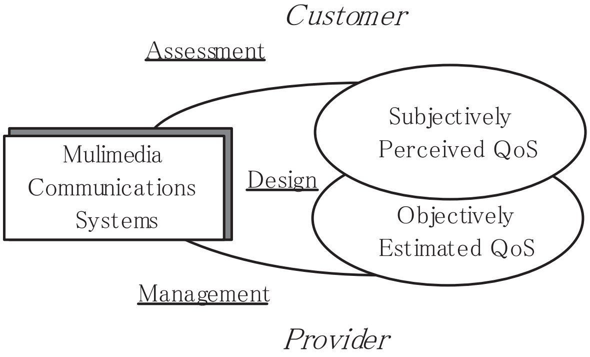

Customer
Assessment
Mulimedia Communications Systems
Design
Subjectively Perceived QoS
Objectively Estimated QoS
Management
Provider

Fig. 1 Concept of perceptual QoS assessment from customer and provider viewpoints.

Objective QoS assessment methodologies may be classified into diagnostic assessment tools for terminals and networks, transmission planning tools for communications system design, and in-service and non-intrusive monitoring tools for communication service management.

Perceptual QoS is affected by many factors in telecommunications systems, and is assessed using psychological evaluation of various quality factors. Figure 2 shows the approach to research on objective QoS measurement that the author adopted since 1975. First, objective QoS measurement methodologies for each quality category consisting of individual QoS factors having similar characteristics such as loudness, cording distortion and IP packet loss and delay, are studied. Then an opinion model for compound QoS factors in the communications systems is established.

2.2 QoS Factors in Multimedia Communications

Figure 3 shows QoS factors affecting perceptual QoS in multimedia communications of advanced mobile and IP-based telephony. These factors include telephone-band speech, wideband speech, audio and video, as well as hands-free quality factors such as noise reduction for speech recognition-synthesis systems and acoustic echo, and new quality factors from IP-based networks such as IP packet loss and circuit echo. In addition to these quality factors, display characteristics and viewing conditions for the assessors should be considered in multimedia communications.

The development of advanced mobile and IP-based telephony has been made possible by the advent of new coding technologies for telephone-band speech, wideband speech, audio, and video media, as well as new transmission technologies based on mobile and IP-based networks. QoS factors affecting perceptual QoS for multimedia services under modern communications networks include coding distortion, delay and media synchronization, acoustic echo, and discontinuous distortion such as IP packet loss,

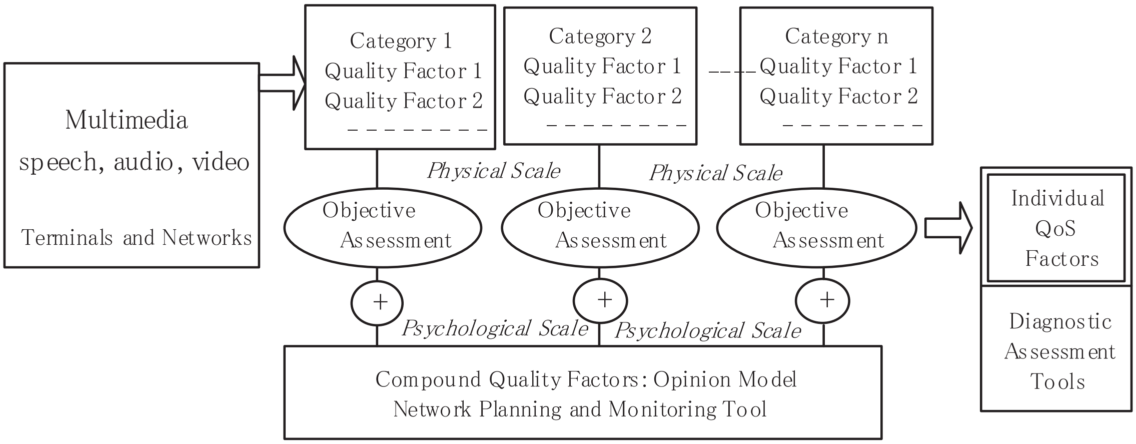
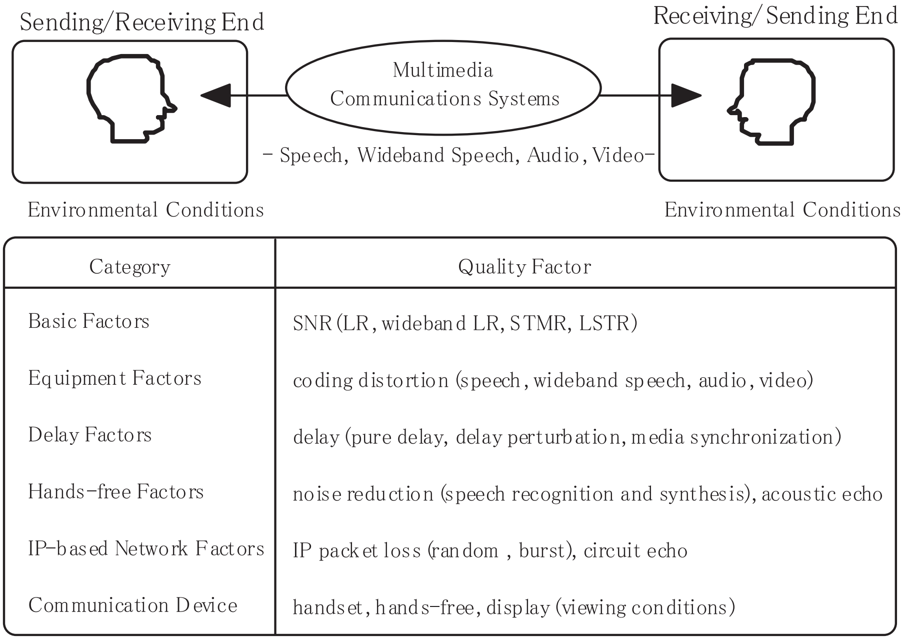


Fig. 2 Approach to objective QoS measurement research in multimedia communications.



Category	Quality Factor
Basic Factors	SNR (LR, wideband LR, STMR, LSTR)
Equipment Factors	coding distortion (speech, wideband speech, audio, video)
Delay Factors	delay (pure delay, delay perturbation, media synchronization)
Hands-free Factors	noise reduction (speech recognition and synthesis), acoustic echo
IP-based Network Factors	IP packet loss (random , burst), circuit echo
Communication Device	handset, hands-free, display (viewing conditions)

Fig. 3 Perceptual QoS factors affecting the perceived quality of multimedia communication.

bit error, and frame erasure.

Recently, wideband speech communication using 7 kHz wideband speech coding, as described in ITU-T recommendations G.722 (Sub-band ADPCM: SB-ADPCM), G.722.1 (Modulated Lapped Transform: MLT), and G.722.2 (Adaptive Multirate Wideband: AMR-WB), has become increasingly necessary in advanced IP telephony when using PCs and mobile communications in vehicles, since, for this, hands-free communication using separate microphones and loudspeakers is indispensable, and in this situation wideband speech is particularly helpful in enhancing the naturalness of communication.

A number of mobile videophone and PC-based applications have been developed. It is likely that such applications will become the major telecommunications services in the near future since the access network is evolving rapidly. For speech communications, ITU-T Recommendation G.107 “The E-model, a computational model for use in transmission planning” has been widely used as a network planning tool for IP-telephony services, as well as for conventional PSTN services [4]. In Japan, a method for assessing the speech quality of IP telephony was standardized as TTC Standard JJ-201.01 conforming to the G.107 in 2003 [5].

Its extension is now being studied by ITU-T Study Groups 12 and 9, and VQEG (Video Quality Expert Group) from the viewpoints of wideband speech, video, and multimedia. In order to provide a means for designing, evaluating, and monitoring multimedia services such as teleconferencing, videophone, and CSCW (Computer-Supported Cooperative Work) applications, it is important to develop an opinion model which can be employed in the evaluation of such multimedia communication services.

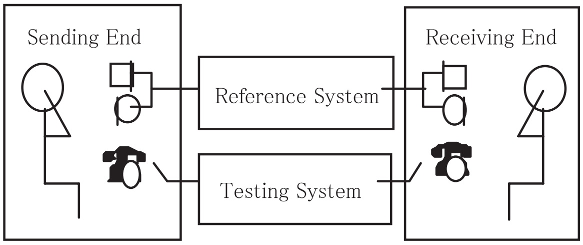


Fig. 4 Example of subjective QoS assessment methods (telephone speech quality assessment).

2.3 Subjective Assessment Methodologies

2.3.1 Subjective and Objective Assessment

Perceptual QoS has been principally assessed by the subjective methods. Figure 4 shows an example of telephone speech quality assessment. The performance of the testing system is directly rated by the speaker and listener (ACR: Absolute Category Rating), or relatively rated by being compared to a reference system (DCR: Degradation Category Rating).

The following opinion scale for an ACR test is the most commonly used for ITU-T telephone applications, and equivalent wording should be used depending on language: excellent, good, fair, poor, and bad. The experimenter allocates the following values to the scores: from 5 to 1. The arithmetic mean of any collection of these opinion scores is called the mean opinion score (MOS).

For the DCR test, the subjects should be instructed to rate the conditions according to the five-point degradation scale as follows: degradation is inaudible (5), degradation is audible but not annoying (4), degradation is slightly annoying (3), degradation is annoying (2), and degradation is very annoying (1). The quantity evaluated from the scores (degradation mean opinion score) is represented by the symbol DMOS.

In picture quality evaluation as well as in the case of telephone speech quality assessment shown in Fig. 4, both ACR and DCR methods are used. The direct rating method is often called the single-stimulus method, and a comparable method using a reference is called the double-stimulus method, and is referred to as either the double-stimulus impairment scale method or double-stimulus continuous quality-scale method depending on the rating scale.

Subjective QoS assessment methods, however, are very expensive and time consuming. This is because they require many test subjects for opinion rating measurements, an expert team of subjects for loudness and intelligibility measurements, a large amount of equipment including a reference system such as ARAEN (which simulates characteristics of a one-meter air space between a speaker's mouth and a listener's ear), and sending and receiving booths which can be altered according to a room's acoustic conditions.

If an objective assessment method by physical metrics can be established, the amount of time needed for assessment can be reduced significantly. Therefore, it is desirable for perceptual QoS to be assessed by objective measurement methods in such a way that the results correlate closely with subjectively determined results.

2.3.2 ITU Recommendations Related to Perceptual QoS

Table 1 shows the recommendations related to perceptual QoS in multimedia communications systems. The P.8XX series is related to objective and subjective QoS assessment based on opinion ratings, and the P.5XX series is related to an objective measuring system and apparatus. In the ITU recommendations, opinion ratings based on customer satisfaction have been primarily employed to assess perceptual QoS. The P.9XX series is related to audio-visual quality and the J.14X series is related to picture and multimedia quality. These are being studied under Study Group 9 in collaboration with video quality expert group (VQEG) with the participation of mainly Study Group 9 and Study Group 12 members. The transmission planning tool is supported by the G.1XX series recommendation. Speech and video coding are supported by the G.7XX series and H.2XX series recommendations respectively under ITU-T. The quality assessment for Hi-Fi audio quality is supported by the BS.XXXX series recommendation under ITU-R (Radio Sector).

Table 1 ITU recommendations related to QoS.

	Recom.	Contents
SG 12	P.1XX	Vocabulary
	P.3XX	Subscriber's line and sets
	P.4XX	Transmission standards
	P.5XX	Objective measuring apparatus
	P.6XX	Electroacoustical measurements
	P.7XX	Measurements of loudness
	P.8XX	Objective and subjective assessments
	G.1XX	Transmission planning
Other groups than SG 12	P.9XX	Audiovisual quality
	J.14X	Picture/multimedia quality
	G.7XX	Speech coding
	H.2XX	Video coding for teleconference
	I. XXX	Performance in ISDN
	BS.XXXX	Hi-Fi audio quality assessment
	BT.5XX	Television quality assessment
	MPEG	Video coding
	MPEG audio	Audio coding

2.3.3 Opinion-Equivalent Q method for Coded Speech Quality Assessment

Speech coding distortion is one of the major factors affecting the QoS of mobile and IP-based telecommunications. To remove any fluctuations in the absolute value of the MOS due to differences in the testing date and conditions, we proposed an opinion equivalent-Q method using an MNRU (Modulated Noise Reference Unit) [6], and it was standardized as P.810 in 1984. The MNRU is a reference system that outputs a speech signal and speech amplitude-correlated noise with a flat spectrum to assess low bit-rate coding. The signal-to-speech-correlated noise ratio is denoted Q (dB), which consists of Q_N for telephone-band speech and Q_W for wideband speech. Figure 5 shows the determination algorithm for opinion equivalent Q. As shown in Fig. 5, the quality of the codec is expressed as an opinion equivalent-Q and is equivalent to the signal-to-speech-correlated noise ratio in terms of the subjective quality of the coding.

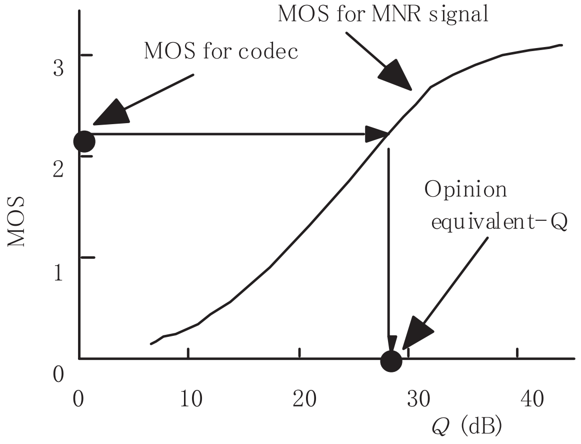


Fig. 5 Determination of opinion equivalent Q in cedec quality evaluation.

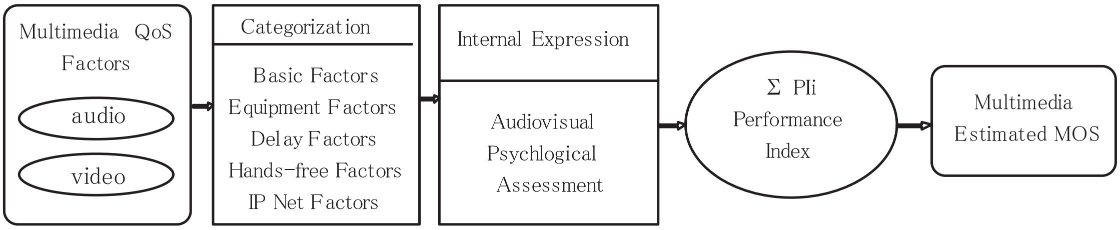


Fig. 6 Concept of opinion model for advanced mobile and IP-based telephony.

2.3.4 Reference Impairment System for Video Quality Assessment

In addition to a concept of MNRU for coded speech quality assessment, a reference system for video is standardized as ITU-T recommendation P.930 [7]. This Recommendation describes the principles of an adjustable video reference system that can be used to generate the reference conditions necessary to characterize the subjective picture quality of video produced by compressed digital video systems. A Reference Impairment System for Video (RISV) can be utilized to simulate the impairments resulting from the compression of video sequences, independent of compression scheme.

An RISV is capable of producing the following categories of distortions, either singly or in combinations, which enable independent adjustment of each impairment level:
a) Artifacts due to conversions between analog and digital formats,
b) Artifacts due to coding and compression, and
c) Artifacts due to transmission channel errors.

3. Objective Quality Prediction Scheme

An objective quality prediction scheme is proposed from the viewpoints of quality measurement tools for each quality factor and an opinion model for compound quality factors in multimedia communications of mobile and IP-based communications networks. The measurement tools are categorized into a speech-layer objective model (e.g., P.862, PESQ) and a packet-layer objective model (e.g., P.VTQ). The opinion model is based on the assumption that all the factors' contributions to quality degradation may be summed on a psychological scale (e.g., G.107, E-model).

3.1 Opinion Model for Compound Quality Factors

3.1.1 Concept of Opinion Model

Perceptual QoS is affected by various quality factors in telecommunications systems. We have proposed an opinion model for estimating the MOS from multiple quality factors in telephone-band communications. We have called this model overall performance index model for network evaluation (OPINE) [8]. The concept of OPINE is as follows. Relevant quality factors are categorized into several groups according to their quality features, and transformed into internal expressions such as the loudness model, the Bark spectrum and a psychological scale derived from an auditory-psychological process.

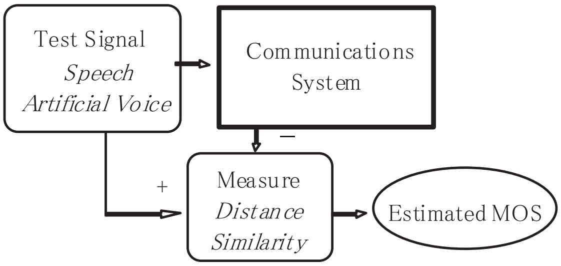


Fig. 7 Proposed model for diagnostic objective measurement.

The MOS can be estimated by summing the psychological performance indexes (PIs) for each group. The OPINE model is described in Supplement 3 of the P-series recommendation in the CCITT Blue Book [9], and is followed by the E-model conforming to ITU-T Recommendation G.107 standardized in1998.

Figure 6 shows the concept of an opinion model for multimedia communications of advanced mobile and IP-based telephony derived from the OPINE model for conventional telephone-band communications.

3.1.2 E-Model

The current Recommendation G.107 E-model, which is related to the transmission planning tool, is based on a concept given in the description of the OPINE model: Psychological factors on the psychological scale are additive. The result of any calculation with the E-model in the first step is the transmission rating factor R, which combines all transmission parameters relevant for the considered connection. This rating factor R is composed of five elements.

$$R = R_0 - I_s - I_d - I_e + A$$

R_0: the basic signal-to-noise ratio including circuit noise and room noise
I_s: a combination of all impairments which occur simultaneously with the voice signal
I_d: the impairments caused by delay
I_e: the impairments caused by low bit rate codecs
A: the advantage factor representing access advantages to the user

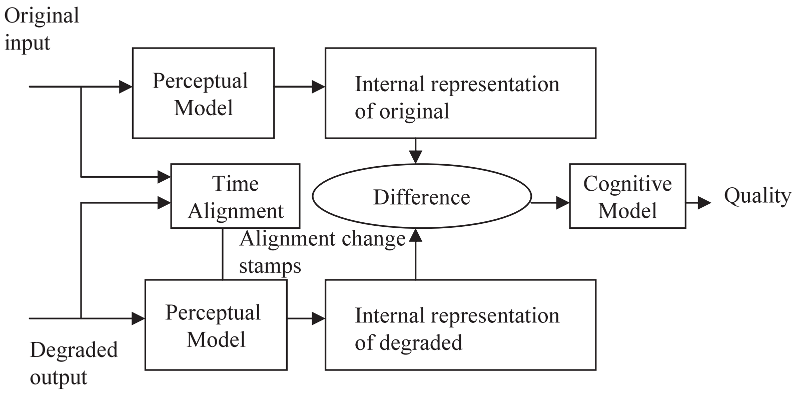


Fig. 8 Overview of PESQ.

However, the consistency of the E-model between subjective and objective assessment could not be sufficiently verified, and the new QoS factors such as wideband speech and video are not included. These points should be studied further.

3.2 Diagnostic Objective Measurement for Individual Quality Factors

3.2.1 Scheme of Objective Measurement for Speech Coding

The author proposed an objective measurement diagram for diagnostics for speech coding distortion as shown in Fig. 7 in 1982 [10]. A test signal is applied to the communications system, and the difference (distance or similarity) between the original and the distorted signal is measured. The MOS is estimated from the distance or similarity obtained.

3.2.2 Objective Measure and Test Signal

Recommendation P.861 "Objective quality measurement of telephone-band (300–3400 Hz) coded speech" using the PSQM (Perceptual Speech Quality Measure) was standardized in 1996 [11]. PSQM compares the original (input) signal with the degraded output of the device under test conditions using a perceptual mode. The key to this process is transformation of both the original and degraded signals into an internal representation (Bark spectrum) that is analogous to the psychological representation of audio signals in the human auditory system, taking perceptual frequency and loudness into account.

Based on the concept of PSQM, a new recommendation P.862 "Perceptual Evaluation of Speech Quality (PESQ)" was standardized in 2000 [12]. Figure 8 shows an overview of the basic thinking behind PESQ. Inserting the time alignment function into the PSQM structure, the PESQ can be applied to not only speech codecs but also end-to-end measurements including filtering, variable delay, channel errors, and IP packet loss. It was verified that PESQ can be used to estimate MOS for these quality factors.

We proposed an artificial voice as a test signal in the objective quality assessment of coded speech because a real speech signal depends on the speaker, and a conventional sinusoidal signal is insufficient for low bit-rate coding [13]. Figure 9 shows the generation method for the artificial voice using a PARCOR synthesizer. The artificial voice reflects the average characteristics of the spoken language such as long-term average spectra, instantaneous amplitude distribution, level distribution of segmental power, spectral distribution of segmental power, voiced/unvoiced structure of speech waveform, and short-term spectral characteristics. It was standardized as Recommendation P.50 [14].

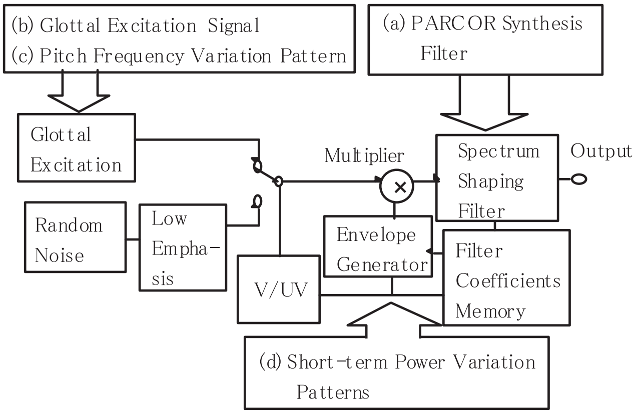

(b) Glottal Excitation Signal
(c) Pitch Frequency Variation Pattern
(a) PARCOR Synthesis Filter
Glottal Excitation
Multiplier
Spectrum Shaping Filter
Output
Random Noise
Low Empha-sis
V/UV
Envelope Generator
Filter Coefficients Memory
(d) Short-term Power Variation Patterns

Fig. 9 Generation method for artificial voice using PARCOR synthesizer.

4. Current Trials for QoS Measurement Tools and Opinion Models

This section describes the author's current trials of measurement tools and opinion models for multimedia communications of advanced mobile and IP-based telephony. These include a wideband speech quality measure, an objective audio quality measure, objective quality assessment methodologies for noise reduction algorithms in speech recognition-synthesis systems, measurement of echo canceller performance for hands-free communication, and comprehensive opinion models for audiovisual communications taking audio and video quality interaction into account.

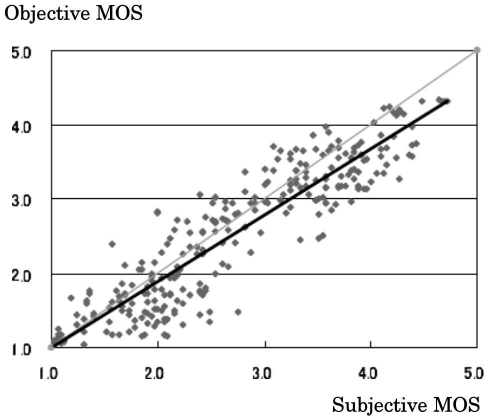


Fig. 10 Relationship between objective MOS estimated by Wideband-PESQ and subjective MOS.

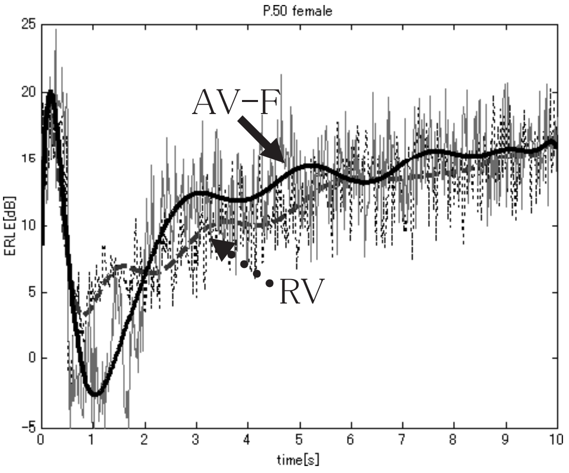


Fig. 11 ERLE characteristics measured by AV-F and RV.

4.1 Objective Quality Assessment of Wideband Speech

4.1.1 Wideband Speech Quality Measure

An objective quality measure called Wideband-PESQ has been proposed for the objective quality measurement of wideband speech coding, and is described in the draft Annex of P.862 as a way of further verifying the performance evaluation.

We have verified the availability of Wideband-PESQ from the viewpoint of consistency between subjectively evaluated MOS and objectively estimated MOS, as shown in Fig. 10 [15]. The codecs used in this experiment were G.722 “SB-ADPCM,” G.722.1 “MLT,” and G.722.2 “AMR-WB.” The correlation coefficient between subjective MOS and objective MOS is 0.913, and the RMSE (Root Mean Square Error) from the regression line is 0.442. Therefore, experimental results showed that the correlation between them is strong, and the RMSE is relatively small. It was concluded that Wideband-PESQ is a promising measure for the objective quality assessment of wideband-speech coding. Wideband-PESQ was standardized as a new recommendation at the Study Group 12 October meeting in 2005.

4.1.2 Test Signal for Wideband Speech Quality Prediction

Recent studies showed that the artificial voice conforming to Recommendation P.50 can be applied to objective quality measurement of not only the newly developed Code Excited Linear Prediction (CELP)-type coding and IP packet loss but also to wideband speech coding using Wideband-PESQ [15]. Therefore, the author proposed the use of artificial voice as well as Wideband PESQ for the objective quality measurement of wideband speech coding in ITU-T Study Group 12.

4.2 Objective Measurement of Hands-Free Quality Factors

4.2.1 Test Signal Used for Measurement of Acoustic Echo Characteristics

Many test signals are described in Recommendation P.501 [16]. However, it is not clear what test signal is appropriate for measuring residual acoustic echo characteristics in hands-free telecommunications. The author compared the performance of various test signals for the measurement of residual acoustic echo characteristics expressed as echo return loss enhancement (ERLE) [17]. The signals were a real voice, as a reference, white noise, frequency-weighted noise, an artificial voice, and a composite source signal listed in Recommendation P.501. Here, the composite source signal is composed of the above mentioned artificial voice sequence in the voiced intervals, a white noise sequence in the unvoiced intervals, and pauses, mixed at random.

Figure 11 shows an example of the ERLE characteristics of an artificial female voice (AV-F) and a real voice (RV). In the figure, smooth curves approximate each ERLE characteristic, in which the dotted curve shows RV, and the solid curve shows AV-F. It is shown that ERLE characteristics measured by the artificial female voice according to Recommendation P.50 are almost equivalent to those of the real voice.

A comparative study for these test signals were carried out from viewpoints of the consistency and RMSE of the ERLE characteristics between each test signal and real voice (as a reference) [17]. The consistency of the convergence time of echo characteristics was also compared between each test signal and real voice. From the comparative assessment, the ERLE characteristics measured using the artificial voice conforming to P.50 were found to be almost equivalent to those of the real voice and the most accurate among the test signals evaluated. It was concluded that test

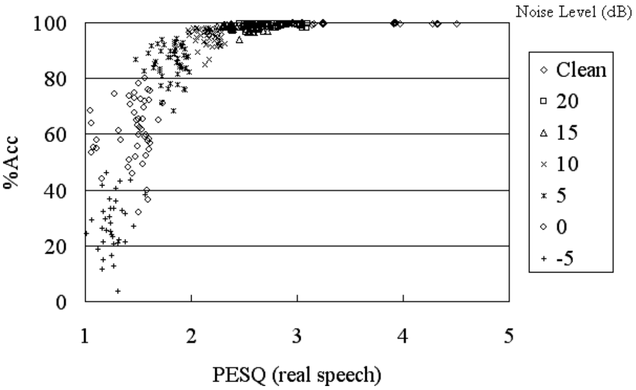


Fig. 12 Relationship between word accuracy and PESQ score in multi-condition training.

signals conforming to P.50 are satisfactory for the measurement of residual echo characteristics. Therefore, the author proposed the use of P.50 for the measurement of acoustic echo characteristics to ITU-T Study Group 12.

4.2.2 Performance Prediction of Noise Reduction Algorithm

In vehicle communication and PC-PC communication, noise reduction technology is introduced to suppress ambient noise. The author proposed a new application of PESQ and the artificial voice conforming to recommendations P.862 and P.50. This application is a methodology for the performance estimation of noise reduction algorithms used for noisy speech recognition [18].

For this purpose, recognition experiments using four noise reduction algorithms were performed on the AURORA-2J connected digit speech recognition task. The noise reduction algorithms were spectral subtraction, temporal domain SVD-based speech enhancement, GMM-based speech estimation, and KLT-based comb-filtering.

Before performing the speech recognition experiments, the conditions for the speech recognition system should be confirmed by "training" the HMM speech recognition system. The training methods used for the algorithms were "clean training" and "multi-condition training." Here, clean training implies use of speech data without environmental noise, and multi-condition training implies use of speech data in noisy speech environments. In this experiment, noisy speech in "subway," "babble," "car," and "exhibition" environments were used.

An example of the relationship between the word accuracy (percentage Acc) and the PESQ score for a number of different noise levels expressed in dB, and the appropriateness of the artificial voice for calculating the PESQ score are shown in Figs. 12 and 13 respectively. These results confirmed that there is a strong correlation between the word accuracy and the PESQ score calculated from real speech and the artificial voice. This method was proposed to Study Group 12.

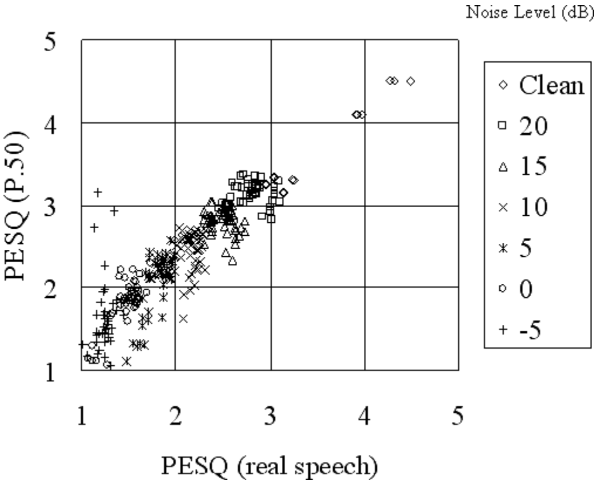


Fig. 13 Relationship between the PESQ scores calculated from real speech and artificial voice.

4.3 Multimedia in Multi-Modal Systems

4.3.1 Optimum Quality Design of Audio Coding at Fixed Low Bit Rate

Recently, high-quality and low-bit-rate audio coding schemes have been developed and used in network services such as the online distribution of music, and the playing of music on PCs. The author proposed an optimum design method for audio coding at a fixed low bit rate, based on choosing the most appropriate bandwidth for encoding the audio samples and taking into account the effects of coding distortion and the limited frequency bandwidth [19].

Also, an objective quality measurement algorithm was proposed, as shown in Fig. 14, which features arbitrary selection of the bandwidth. It was concluded that it is possible to select the most suitable bandwidth without relying on the designer's subjective intuition, which is obviously not always reliable.

4.3.2 Multimedia Opinion Model Based on Media Interaction of Audiovisual Communication

A number of PC-based applications have been developed for videoconferencing and videophone services as well as for mobile videophone services. It is likely that such applications will become the major telecommunications services in the near future since the access network is evolving rapidly. Multimedia is defined as the combination of multiple forms of media such as audio, video, text, graphics, fax, and telephony in the communication of information. The initial goal of our work is to produce an objective measurement of quality for audiovisual communications. The primary use of the model is to measure the quality of limited bandwidth services. It is considered that limited bandwidth represents a more critical level of service from the viewpoints of multimedia quality degradation.

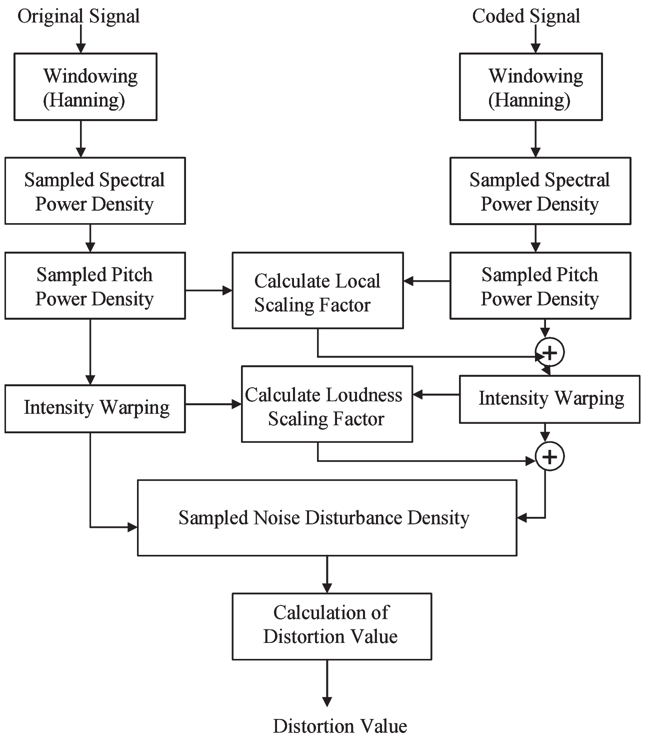

Original Signal
Coded Signal
Windowing (Hanning)
Windowing (Hanning)
Sampled Spectral Power Density
Sampled Spectral Power Density
Sampled Pitch Power Density
Calculate Local Scaling Factor
Sampled Pitch Power Density
Intensity Warping
Calculate Loudness Scaling Factor
Intensity Warping
Sampled Noise Disturbance Density
Calculation of Distortion Value
Distortion Value

Fig. 14 Objective audio quality measurement algorithm.

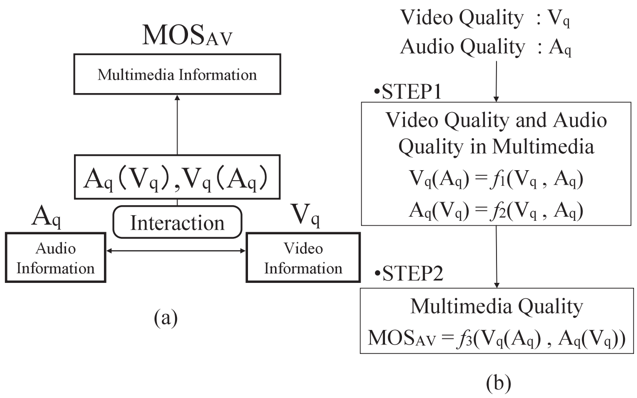


Fig. 15 Proposed multimedia opinion model (method 2).

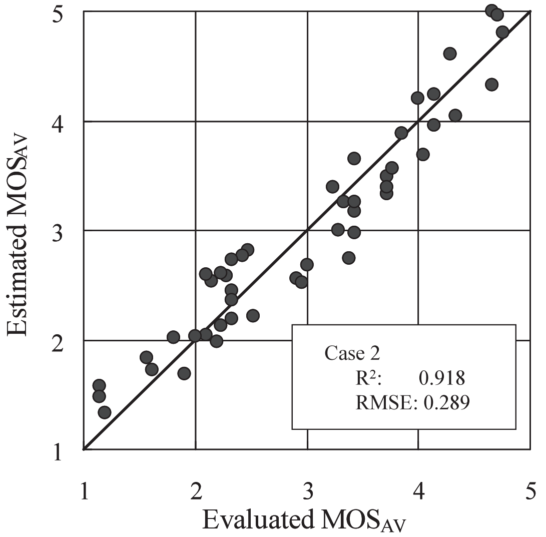


Fig. 16 Relationship between evaluated MOS_{AV} and estimated MOS_{AV}.

Perceptual models for the objective measurement of audio quality and video quality have been studied [20], [21]. In these previous studies, multimedia quality (MOS_{AV}) is estimated by conducting separate objective measurements of audio quality (A_q) and video quality (V_q). The multimedia integration function is constructed by each A_q and V_q (method 1).

However, we focused on the mutual interaction between audio and video information [22]. That is, perceived multimedia quality depends on the mutual interaction of audio and video information, and cannot be adequately assessed from independent audio and video information. We proposed a new multimedia opinion model, shown in Fig. 15, in which multimedia quality (MOS_{AV}) is estimated from $A_q(V_q)$ and $V_q(A_q)$, (method 2). Here, $A_q(V_q)$ implies an objective measure of audio quality taking into account the influence of video quality, while $V_q(A_q)$ implies a similar measure of video quality taking into account the influence of audio quality.

The proposed multimedia opinion model has been verified from the viewpoints of consistency between the subjectively evaluated MOS and the objectively estimated MOS for audiovisual communications, and a comparison of the estimation accuracy with that of the previous multimedia opinion model (method 1).

Figure 16 shows an example of the relationship between estimated MOS_{AV} from estimated $V_q(A_q)$ and $A_q(V_q)$, and evaluated MOS_{AV}. It was shown that the proposed multimedia opinion model has the best performance index. Therefore, it is concluded that for construction of a multimedia opinion model, it is important to take into account the mutual interaction between audio and video information.

However, research on the multimedia opinion model has just started, and in this experiment, treated only coding bit rates for audio and video as quality factors. Almost all the quality factors categorized in Fig. 6 are currently not included in the proposed multimedia opinion model. Further improvement is needed for these points.

4.3.3 Wideband, Multimodal, and Multiparty

Figure 17 shows the framework of quality assessment research for next-generation communications services. The key words for such services will be “wideband,” “multimodal,” and “multiparty.” As the bandwidths of core and access networks rapidly become broader, telecommunications applications will have more bandwidth available for speech, audio, and video data, and this will lead to higher-quality services.

Assessing the quality of such services based on the simple opinion ratings that have been used for the PSTN

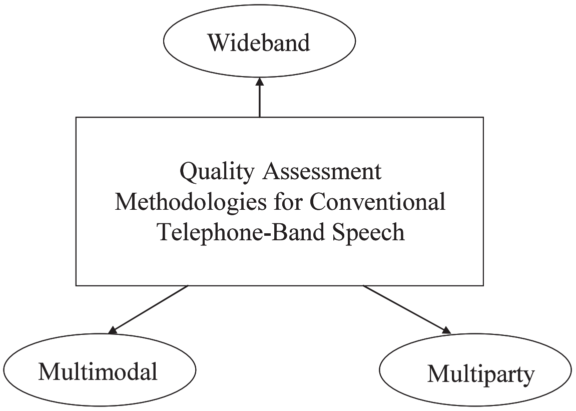

Wideband
Quality Assessment Methodologies for Conventional Telephone-Band Speech
Multimodal
Multiparty

Fig. 17 Framework for the development of quality assessment research.

and for VoIP might be insufficient. Thus, NTT proposed that services should be compared not only on a one-dimensional scale, like the MOS, but also on a multidimensional scale that characterizes the QoS in a way that takes the richness of the services into account. There have been some studies of psychological factors that affect high-quality audio and video [23]. In evaluating the quality levels of multimodal services, we need to take into account the interaction between media, as well as the quality of individual media.

The conventional targets of quality assessment have been one-to-one communications and one-way content-delivery services. IP telephony and video-streaming services are typical examples. With the advent of high-speed wired and wireless networks, multiparty communications services, such as instant messaging, teleconferencing, and distributed collaboration services are being deployed. Such services are multipoint (users are geographically dispersed), real time, and interactive. The important points in evaluating the quality of multiparty services are the heterogeneous communications environments of the users and the synchronization of user streams.

5. Conclusions

This paper has described the author's perspective on multimedia quality prediction methodologies for multimedia communications for advanced mobile and IP-based telephony, with the primary focus on objective quality assessment, and described associated practical experiments and trials.

First, the paper described the need for perceptual QoS assessment of various quality factors such as a) wideband speech, audio, and video for multimedia, b) noise reduction and speech recognition synthesis for hands-free communications, and c) IP packet loss, acoustic echo, and pure delay for IP-based telephony.

Then, an objective quality prediction scheme was proposed from the viewpoints of quality measurement tools for each quality factor and an opinion model for compound quality factors. The measurement tools are categorized into a speech-layer objective model and a packet-layer objective model for diagnostic evaluation. The opinion model is based on the assumption that the contributions of all the factors to quality degradation may be summed on a psychological scale.

Finally, the author's current trials of measurement tools and opinion models were described. These include wideband speech and audio quality measures, objective quality measurements for hands-free telephony, and opinion models for wideband and audiovisual communications. These methodologies were proposed to ITU-T Study Group12.

References

[1] N. Kitawaki, "Perceptual QoS assessment for wireless personal communications," IEEE Fourth International Symposium on Wireless Personal Multimedia Communications, Proc. WPMC'01, pp.541–546, Sept. 2001.

[2] N. Kitawaki, "Perspectives on multimedia quality prediction methodologies for advanced mobile and IP-based telephony," Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction, Proc. ETSI STQ, pp.1–8, June 2004.

[3] 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.

[4] ITU-T Rec. G.107, "The E-model, a computational model for use in transmission planning," May 2000.

[5] TTC Standard JJ-201.01 (Japan), "A method for speech quality assessment of IP telephony," April 2003.

[6] ITU-T Rec. P.810, "Modulated noise reference unit," Feb. 1996.

[7] ITU-T Rec. P.930, "Principles of a reference impairment system for video," Aug. 1996.

[8] N. Osaka, K. Kakehi, S. Iai, and N. Kitawaki, "A model for evaluating talker echo and sidetone in a telephone transmission network," IEEE Trans. Commun., vol.40, no.11, pp.1684–1692, Nov. 1992.

[9] CCITT P-Series Recommendations, Blue Book, vol.V, Supplement no.3, 1988.

[10] N. Kitawaki, K. Itoh, M. Honda, and K. Kakehi, "Comparison of objective speech quality measures for voiceband codecs," IEEE Int. Conf. on Acoust., Speech, Signal Processing, ICASSP'82, pp.S9.5.1–9.5.4, May 1982.

[11] ITU-T Rec. P.861, "Objective quality measurement of telephone-band (300–3400 Hz) speech codecs," Aug. 1996.

[12] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[13] K. Itoh, N. Kitawaki, H. Nagabuchi, and H. Irii, "A new artificial speech signal for objective quality evaluation of speech coding systems," IEEE Trans. Commun., vol.42, no.2/3/4, pp.664–672, 1994.

[14] ITU-T Rec. P.50, "Artificial voices," Sept. 1999.

[15] N. Kitawaki, K. Nagai, and T. Yamada, "Objective quality assessment of wideband speech coding," IEICE Trans. Commun., vol.E88-B, no.3, pp.1111–1118, March 2005.

[16] ITU-T Rec. P.501, "Test signals for use in telephonometry," May 2000.

[17] N. Kitawaki, T. Yamada, and F. Asano, "Comparative assessment of test signals used for measuring residual echo characteristics," IEICE Trans. Commun., vol.E86-B, no.3, pp.1102–1108, March 2003.

[18] T. Yamada and N. Kitawaki, "A PESQ-based performance prediction method for noisy speech recognition," International Congress on Acoustics, ICA2004, Proc. pp.Tu.P2.9, April 2004.

[19] N. Kitawaki, K. Kotegawa, T. Yamada, F. Asano, and Y. Hiraguri, "Optimum quality design of audio coding at fixed low bit-rate, taking account of coding distortion and bandwidth limitation," IEICE Trans. Commun. (Japanese Edition), vol.J87-B, no.11, pp.1888–1897, Nov. 2004.

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

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

[22] N. Kitawaki, Y. Arayama, and T. Yamada, "Multimedia opinion model based on media interaction of audio-visual communications," International Conference on Measurement of Speech and Audio Quality in Networks, Proc. MESAQIN2005, pp.5–10, June 2005.

[23] T. Tachi, S. Iai, and N. Kitawaki, "Proposal of selection method of test pictures in HDTV subjective quality assessments," IEEE Proc. MULTIMEDIA'92, pp.67–68, April 1992.

Nobuhiko Kitawaki was born in Aichi, Japan, on September 27, 1946. He obtained B.Eng., M.Eng. and Dr. Eng. degrees from Tohoku University, Japan, in 1969, 1971, and 1981, respectively. From 1971 to 1997, he was engaged in research on speech and acoustic information processing at the laboratories of Nippon Telegraph and Telephone (NTT) Corporation. From 1993–1997, he was the Executive Manager of NTT's Speech and Acoustics Laboratory. He currently serves as a Professor of the Graduate School of Systems and Information Engineering, and Dean of the College of International Studies, University of Tsukuba, Japan. He has contributed to ITU-T Study Group 12 from 1981 until the present, and served as a Rapporteur from 1985 to 2000. Prof. Kitawaki is a Fellow of the IEEE and a councilor of the ASJ, and a member of the IPSJ. He received paper awards from the IEICE in 1979 and 1984, and an award presented by the Minister of Posts and Telecommunications from ARIB in 1995.