

# Perceptual QoS Assessment Technologies for VoIP

Akira Takahashi and Hideaki Yoshino, NTT Service Integration Laboratories  
Nobuhiko Kitawaki, University of Tsukuba

## ABSTRACT

Since quality is not generally guaranteed in an IP network, the proper design and management of networks and/or terminals for high-quality voice over IP services and maintenance of service levels is important. In terms of quality design and management, methodologies for appropriately and effectively evaluating the perceptual QoS of VoIP are indispensable. This article gives an overview of the state of the art of quality assessment technologies for VoIP, including recent work on improving their accuracy.

## INTRODUCTION

Voice over IP (VoIP), the integration of conventional telephone services with the growing number of other IP-based applications, is seen as one of the most important technologies for telecommunications providers. In addition to the cost reduction achieved by the sharing of network resources, VoIP is expected to accelerate the development of rich multimedia services.

Since quality is not generally guaranteed in an IP network, it is important that we properly design the networks and/or terminals before providing the service, constantly monitor the quality of the service, and take action as necessary to maintain the level of service. In achieving these goals, methodologies for evaluating the perceptual quality of service (QoS) of VoIP services are indispensable.

In this article we start by introducing the quality assessment of telephony services, and describe how objective factors determine the perceived quality of a VoIP system. Next, we give an overview of subjective quality assessment methodologies in which users' perceptions of communications quality are directly measured through psycho-acoustic experiments. Then we discuss objective quality assessment methodologies, in which subjective quality is estimated from measured physical characteristics of the terminals and networks. Study Group 12 of the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T SG12) is responsible for investigating

such methodologies, and we review the standards introduced to date and the group's current activities. We also introduce our recent study on improving the accuracy of objective quality assessment. Finally, we discuss future research needed on quality assessment of high-quality multimedia communications services.

## AN OVERVIEW OF PERCEPTUAL QOS ASSESSMENT

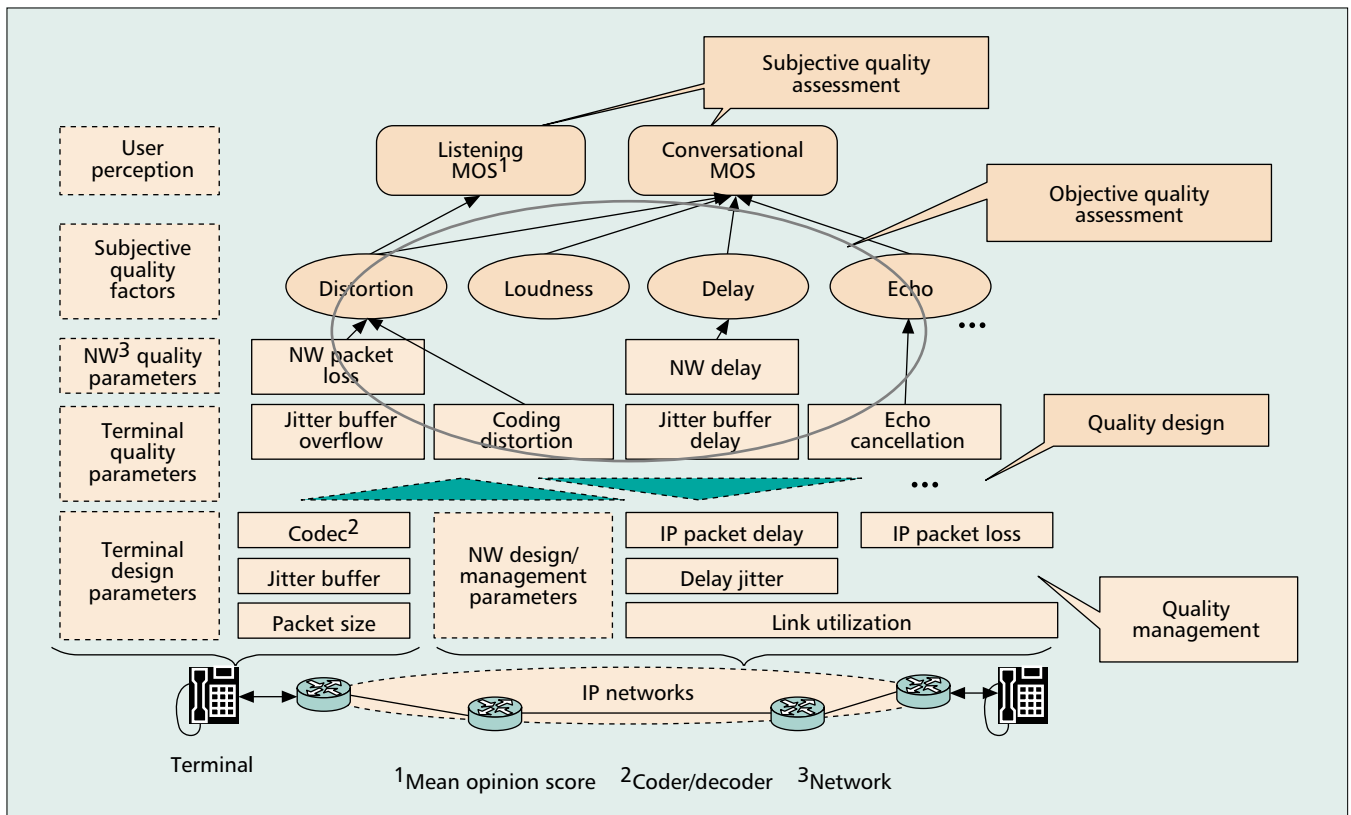
### DETERMINANTS AND ASPECTS OF QUALITY IN VOIP SYSTEMS

Figure 1 shows the various aspects of the perceptual QoS of a VoIP system and how these are determined. In conventional telephony until the 1980s, where the signal bandwidth was a fixed 0.3–3.4 kHz, the impairment factors were transmission loss, frequency distortion, stationary circuit noise, and, in digital systems, signal-correlated quantization noise associated with pulse code modulation (PCM), and so on. These properties are usually described in terms of a signal-to-noise ratio (SNR). Since VoIP systems are based on new coding technologies and a new transmission technology, the primary determinants of the perceptual QoS of a VoIP service are distortion caused by speech coding and packet loss, loudness, delay (network and terminal delay), and echo.

### PERCEPTUAL QOS ASSESSMENT, DESIGN, AND MANAGEMENT

The prime criterion for the quality of audio and video communications services is subjective quality, the users' perceptions of service quality. This can be measured through *subjective quality assessment*. The most widely used metric is the mean opinion score (MOS). However, while subjective quality assessment is the most reliable method, it is also time-consuming and expensive. Methods for estimating subjective quality from physical quality parameters are thus desirable. This process is called *objective quality assessment*.

The subjective quality factors are mapped to network and terminal quality parameters as



■ **Figure 1.** Factors that determine the quality of a VoIP service.

shown in Fig. 1. Since service providers use quality assessment technologies in order to design and manage QoS in a way that takes users' perceptions into account, they need to further map these quality parameters to parameters that are designed and/or managed.

## SUBJECTIVE QUALITY ASSESSMENT

### OPINION RATING

The most widely used subjective quality assessment methodology is *opinion rating*, which is defined in ITU-T Recommendation P.800. The performance of the system under test is rated directly (absolute category rating, ACR) or relative to the subjective quality of a reference system (degradation category rating, DCR).

The following opinion scale used in an ACR test is the most frequently used in ITU-T: excellent (5), good (4), fair (3), poor (2), and bad (1). Equivalent wording should be used in languages other than English, which might result in small variations in the original score. The arithmetic mean of all the opinion scores collected is the MOS.

In a DCR test, the subjects are instructed to rate the conditions according to this five-point degradation category scale: degradation is inaudible (5), audible but not annoying (4), slightly annoying (3), annoying (2), and very annoying (1). The mean value of the results is called the degradation mean opinion score (DMOS).

From the viewpoint of the factors taken into account in subjective experiments, subjective quality is categorized into *listening quality* and

*conversational quality*, as shown in Fig. 1. Clearly, conversational quality assessment must involve two-way communication, while in listening quality assessment, subjects are simply provided with recorded speech material. Although the overall quality of VoIP must be discussed in terms of conversational quality, listening quality assessment is also quite helpful in diagnosing the effects of individual quality factors such as speech coding and packet loss.

### OPINION EQUIVALENT-Q METHOD

The MOS is experiment-dependent due to differences in the testing date and the mix of quality levels in the experiment. For example, if we employ a lot of good-quality conditions, the MOS for a certain condition becomes lower than that obtained when we use fewer good-quality conditions. Therefore, we need to remove these effects from the MOS. We have proposed the opinion equivalent-Q method, in which the modulated noise reference unit (MNRU) is used. This approach was standardized as ITU-T Recommendation P.810 in 1984. MNRU is a reference system that outputs a speech signal and speech-amplitude-correlated noise with a flat spectrum. The ratio of signal to speech correlated noise in dB is called the Q value.

The opinion equivalent-Q is defined as the Q-value of MNRU speech with quality equivalent to that of the speech under evaluation. Since we expect the relative quality between MNRU and target speech to be preserved across subjective experiments, the resulting equivalent-Q value is reproducible.

The accuracy of E-model prediction is being thoroughly studied by ITU-T. This might lead to a revision of the existing Recommendation (G.107). We can also expect the integration of opinion models with speech-layer and/or packet-layer objective models.

Category	Opinion models (e.g., G.107, E-model)	Speech-layer objective models (e.g., P.862, PESQ)	Packet-layer objective models (e.g., P.VTQ)
Aim	Network planning	Benchmarking/management	Management
Measurement procedure	(Planning value)	Active/passive	Passive
Input information	Quality parameters	Speech signals	IP packet (not payload)
Estimated quality	Conversational MOS	Listening MOS	Listening MOS

■ **Table 1.** Three categories for objective quality assessment methodologies.

## OBJECTIVE QUALITY ASSESSMENT

Since subjective quality assessment is time-consuming and expensive, we need a method for estimating subjective quality by measuring the physical characteristics of the terminals and networks. In a wide sense, all such methods are forms of objective quality assessment.

Objective quality assessment methodologies can be categorized into several groups from the viewpoints of aim, measurement procedure, input information, and MOS for estimation (Table 1). In this article objective quality assessment methodologies that exploit network and terminal quality parameters, as shown in Fig. 1, and produce estimates of conversational MOS are called *opinion models*. On the other hand, those that require speech signals as inputs and produce estimates of listening MOS are called *speech-layer objective models*, and those that exploit IP packet characteristics and produce estimates of listening MOS are called *packet-layer objective models*. Although the speech- and packet-layer objective models estimate the same thing (i.e., the listening quality), they are used in different scenarios. For example, if it is impossible or difficult to obtain actual speech samples via in-service quality monitoring, we should use packet-layer objective models. Conversely, if it is difficult to capture necessary packet information or we need to obtain quality estimates that are as accurate as possible, we should use speech-layer objective models.

Figure 2 is a review of the history of these technologies in the ITU. We analyze trends in research and standardization of these three forms of technology in the following subsections.

### OPINION MODELS

Opinion models have long been studied in the ITU-T, and several models were proposed in the early 1980s as candidates for an ITU-T standard. However, the working group was unable to settle on a single algorithm as the international standard and consequently ended up creating an informative document in which four different models were introduced.

A new model called the *E-model* was proposed for ITU-T standardization in the 1990s. The E-model is based on the OPINE model from NTT [1], in which all factors responsible for quality degradation are summed on a psychological scale. The E-model and the TR model from AT&T produce similar output: a score on a psychological scale is produced as an index of overall quality. In the E-model, the quality degradation introduced by speech coding, bit

error, and packet loss is treated collectively as an *equipment impairment factor*. ITU-T standardized the E-model as Recommendation G.107 in 1998. It was also adopted by the European Telecommunications Standards Institute (ETSI) and Telecommunications Industry Association (TIA) as a network planning tool [2, 3] and has become the most widely used opinion model in the world.

The E-model has 20 input parameters that represent the terminal, network, and environmental quality factors. Its output is called the *R-value*, which is a function of the 20 input parameters. First, the degrees of quality degradation due to individual quality factors such as loudness, echo, delay, and distortion are calculated on the same psychological scale. Then these values are subtracted from the reference value.

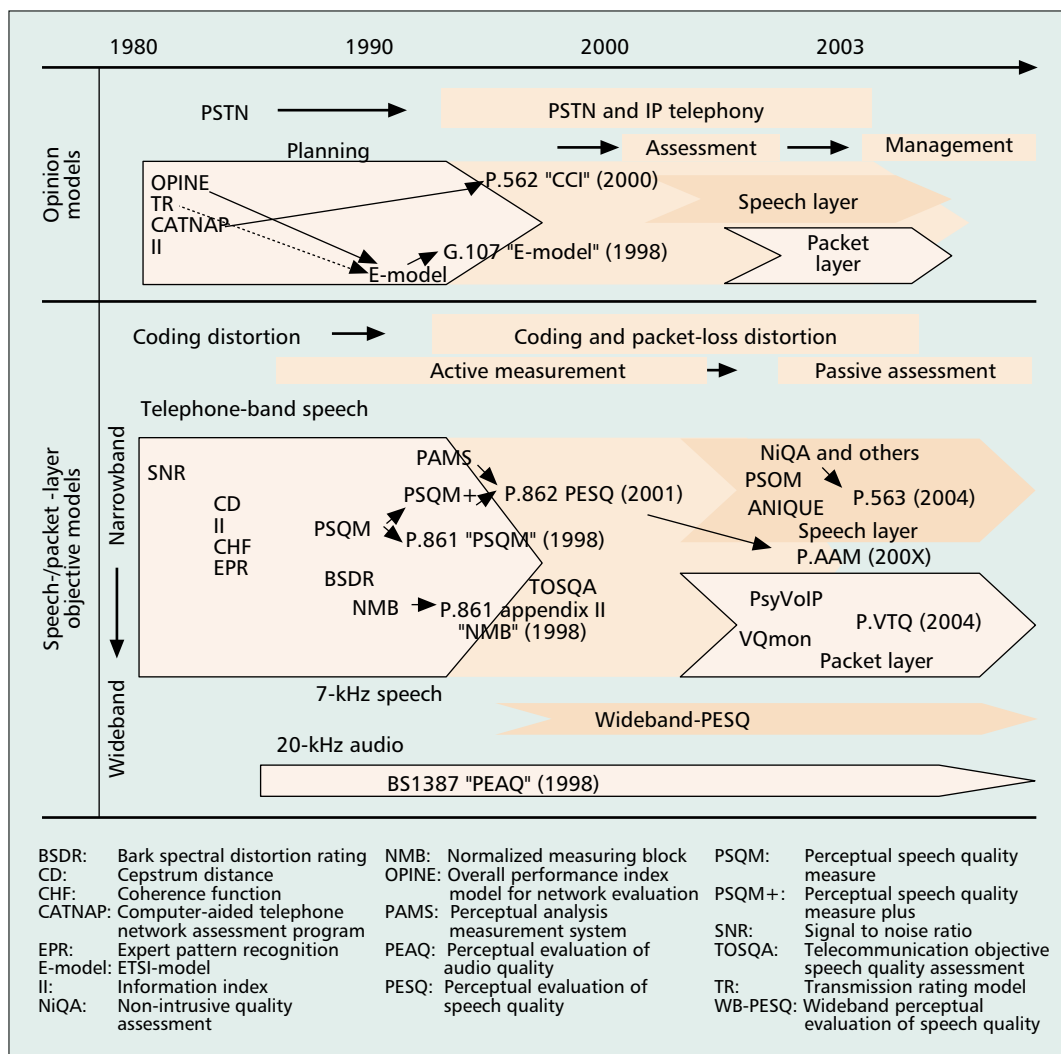
Recommendation G.107 provides a set of default values, which can be used when network planners assume that terminals and the usage environment are normal. An index of overall quality should thus represent the perceptions of a user using a normal terminal under normal circumstances.

Since a basic assumption for the E-model is telephone band (300–3400 Hz) handset communications, it is inapplicable to the evaluation of hands-free or wideband (e.g., 150–7000 Hz) communications. Taking into account the quality evaluation of future speech and multimedia communications services, it is quite important to expand the scope of the E-model.

Although the R-values produced by the E-model have some correlation with subjective conversational MOS and are useful in network planning, they are not necessarily accurate as estimators of subjective quality. In particular, the validity of the additive property assumed in the E-model is sometimes questionable [4, 5]. This is discussed in detail later. The accuracy of E-model prediction is being thoroughly studied by the ITU-T. This might lead to a revision of the existing Recommendation (G.107). We can also expect the integration of opinion models with speech-layer and/or packet-layer objective models, which are introduced in the next subsection. One such example is ITU-T Recommendation P.834, which provides a means of converting the results of a speech-layer objective model into a quantity that can be incorporated in Recommendation G.107, the E-model, although its scope is currently limited to the evaluation of error-free speech.

### SPEECH-LAYER OBJECTIVE MODELS

The study of speech-layer objective models started with the use of signal-to-noise ratio (SNR) as a means of evaluating PCM-coded speech. In the



■ **Figure 2.** Standardization of objective quality assessment: history and current state.

Some problems regarding the implementation of Recommendation P.862 have recently been reported to ITU-T. An applications guide will now be published as a new Recommendation, P.862.2..

latter half of the 1980s, several objective models that exploited spectral distortion rather than waveform distortion were proposed as objective quality assessment methods more applicable to the evaluation of low-bit-rate codecs. However, due to their lack of accuracy in estimation, none was standardized as an ITU-T Recommendation. Later, a model based on Bark spectral distortion provided adequate accuracy, and was the basis for Recommendation P.861, “Perceptual Speech Quality Measure (PSQM),” in 1998.

Since this approach shows sufficient performance only under error-free coding conditions, it is inapplicable to the evaluation of VoIP speech in general, which often suffers from packet loss. Therefore, the next target for standardization was the development of a method that is applicable to the evaluation of the effects of discontinuous forms of degradation such as packet loss in VoIP and bit errors in mobile communications. Consequently, a compromise between the algorithms of the Perceptual Analysis Measurement System (PAMS), which utilizes different perceptual modeling than PSQM and has a quite sophisticated time-alignment scheme, and PSQM+, which is an extension of PSQM, was standardized as Recommendation P.862, “Perceptual Evalua-

tion of Speech Quality (PESQ),” in 2001. Since this is a “full-reference” objective model (i.e., it requires input test speech as a reference), its main application is active measurement in which test speech samples are fed into the system under test, and the original speech is compared to the post-transmission speech.

Some problems regarding the implementation of Recommendation P.862 have recently been reported to ITU-T. An applications guide will now be published as a new Recommendation, P.862.2, so equipment makers and network operators and providers will be aware of the problems and able to use the Recommendation appropriately.

The current targets for standardization in the field of speech-layer objective models are:

- P.563, a “no-reference” model, in which standard electro-acoustic characteristics are assumed for the terminals
- P.AAM, a full-reference model that takes into account the electro-acoustic characteristics of terminals

#### PACKET-LAYER OBJECTIVE MODELS

ITU-T is also trying to standardize an objective quality assessment methodology based solely on IP packet information (not speech in the pay-

## IMPROVING THE ACCURACY OF OBJECTIVE QUALITY ASSESSMENT

This section describes the estimation accuracy of the E-model for the major quality factors in VoIP (i.e., speech distortion, delay, talker echo, and loudness) [4]. Then we introduce a modified model based on the E-model but with improved performance, and demonstrate its validity in estimating the conversational quality of practical VoIP systems.

We conducted two subjective quality experiments to investigate the performance of the original E-model for the individual quality factors as well as for the interaction between speech distortion and delay. For further details of the experimental conditions, see [4]. The subjective experiments were conducted using the conversational ACR method defined in ITU-T Recommendation P.800.

Although ITU-T Recommendation G.107 Annex B provides a relationship between the R-value produced by the E-model and the estimated MOS (MOS-CQE, where CQE stands for conversational quality estimated by a network planning tool, as defined in ITU-T Recommendation P.800.1), it would be inappropriate to directly apply this to the Japanese-language experimental results. This is because of the systematic difference between scores for the major western European languages and for Japanese. Therefore, we used the following transformation to estimate the Japanese MOS-CQE, hereafter denoted MOS-CQE-j, from the R-value:  $MOS-CQE-j = 0.8681MOS-CQE + 0.0271$ . This was derived by analyzing the MOS scores for various languages, including Japanese, provided by Supplement 23 to ITU-T P-series Recommendations. We transformed the R-value to MOS-CQE by applying the formula defined in ITU-T Recommendation G.107 Annex B.

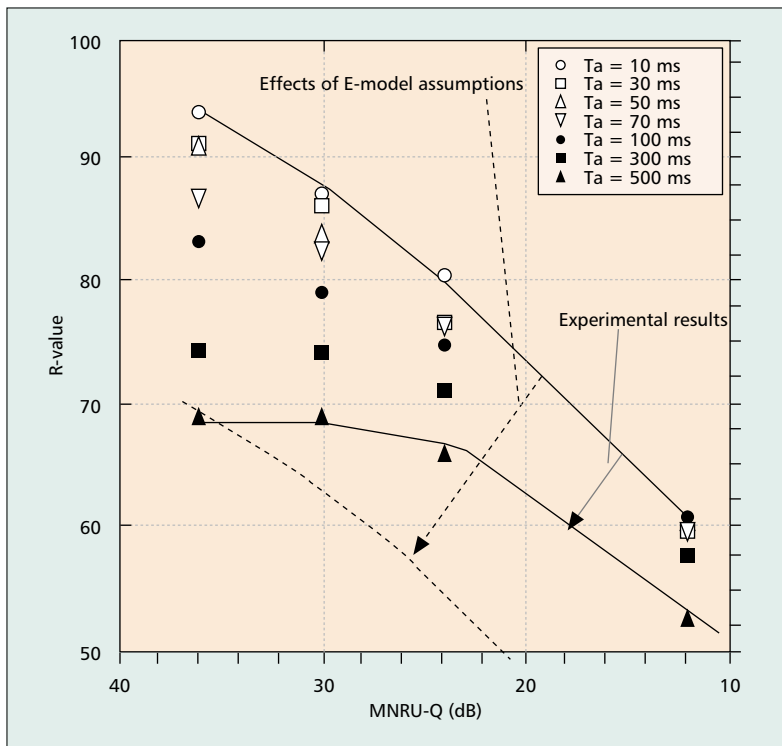
### PERFORMANCE OF THE E-MODEL

**Interaction between Delay and Speech Distortion** — The E-model assumes that individual quality factors such as loudness, delay, talker echo, and speech distortion have mutually independent effects on the psychological scale. We investigate the interaction between delay and speech distortion.

Figure 3 shows the effect of speech distortion on the R-value for various one-way delay ( $T_a$ ) conditions in terms of MOS-CQS, which stands for the MOS for conversational quality obtained by subjective experiments. This figure implies that speech distortion is dependent on delay in a way that is affected by the R-value. That is, the additive property assumed for the E-model does not necessarily hold, and this is especially so in the low-quality region.

**Individual Quality Factors** — We found the following points in the evaluation of individual quality factors:

- Loudness: The E-model estimated the subjective quality fairly well at values around the typical overall loudness rating (OLR) value, which was 10 dB.



■ Figure 3. Interaction between pure delay and speech distortion.

load) for use in real-time quality monitoring. This is provisionally called P.VTQ. The process is now in the algorithm selection phase, and there are two candidates: PsyVoIP [6] and VQmon [7]. As of the September 2003 meeting of SG12, the performance evaluation tests still had not been completed.

The P.VTQ procedure starts with estimation of intermediate quality parameters such as packet loss rate, packet loss pattern, and delay jitter from the Real-Time Transport Protocol (RTP) header and/or Real-Time Transport Control Protocol (RTCP) information. The listening MOS is then estimated. The intermediate quality parameters form a subset of entities defined in the IETF. It is then easy to construct a quality management system if RTCP-XR is implemented in the VoIP system. That is, since a terminal which can handle RTCP-XR provides the necessary information, the listening quality can be estimated by applying the second-stage P.VTQ algorithm. RTCP-XR can also be used for reporting the estimated listening quality.

In cases where the terminals cannot handle RTCP-XR, the intermediate quality parameters can be estimated from RTP and/or RTCP packets (the first stage of the P.VTQ algorithm). The second stage of the algorithm is then used to estimate listening quality. Here, the problem is that the packet loss rate differs, even given the same delay jitter conditions, according to the implementation of the jitter buffers of terminals. Psytechnics [6] solves this problem by preparing a calibration file for each terminal type in advance, containing a description of the characteristics of the jitter buffer in the terminal.

- Delay: Although an assumption of the E-model is that no degradation due to delay occurs at delay values up to 100 ms, real subjective quality shows different characteristics. In addition, the E-model tends to underestimate the quality for delays beyond 300 ms.
- Echo: The estimates tend to provide a good match with the subjective quality under “good” (MOS-CQS > 3.0) talker echo conditions. For conditions in which the talker echo is annoying, however, the estimate diverges from the actual subjective quality.

### THE MODIFIED OPINION MODEL

We took the problems pointed out in the previous subsection into account in modifying the E-model so that it better estimates subjective quality. In this article, we call the result the *modified E-model*. We did not modify the E-model equations for loudness, because the E-model estimates for loudness seem to be reasonable in the typical OLR range.

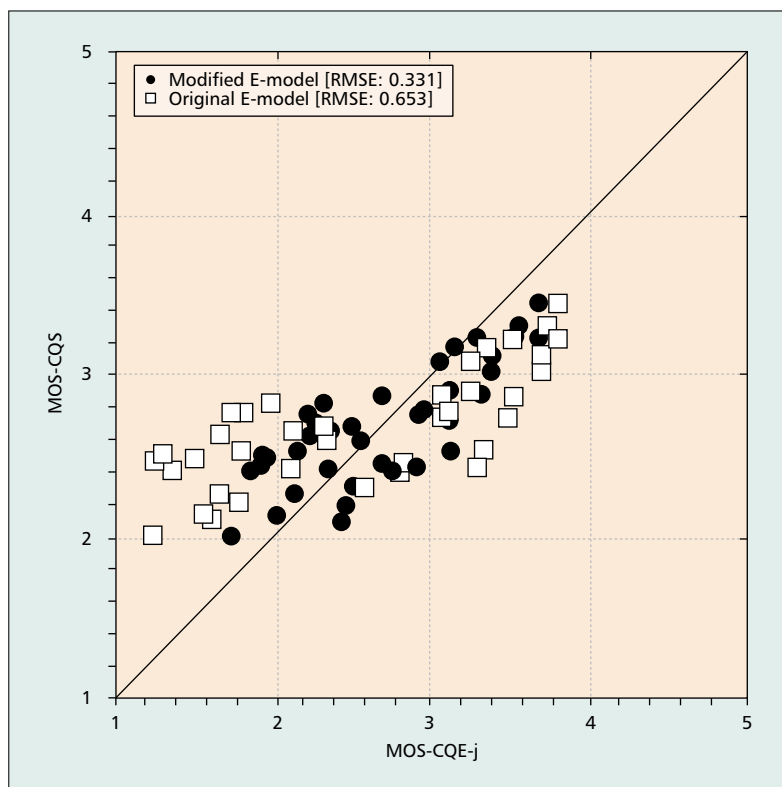
**Effects of Delay and Talker Echo** — We modeled the impairment due to delay (i.e., *I<sub>dd</sub>* in Recommendation G.107) as a second-order polynomial in terms of delay. After investigating the effect of talker echo as a function of echo path delay (*T*) and echo loudness (*TELR*), we modeled the talker echo impairment factor (i.e., *I<sub>dte</sub>* in Recommendation G.107) as a combination of logarithmic and exponential functions of *T* and *TELR*. Here, we assumed typical terminal and environmental conditions as given by the default parameters of the E-model. Modeling *I<sub>dte</sub>* for other conditions is a subject for further study. The assumption of default values is, however, practical in most cases of the evaluation of handset communications.

**Interaction between Delay and Speech Distortion** — For the E-model, the effects of delay and speech distortion (i.e., *I<sub>e,eff</sub>* in Recommendation G.107) are assumed to be independent. However, we model the overall effect as a quadratic equation in terms of *I<sub>dd</sub>* and *I<sub>e,eff</sub>*, because some dependence between them was obvious, and we found the quadratic equation to be quite accurate. Here, we apply the new *I<sub>dd</sub>* defined above.

### VALIDITY OF THE MODIFIED MODEL

We compared the performance of the original and modified E-models by applying them to the evaluation of conversational quality in practical VoIP systems with various delay, echo, codec, and packet loss conditions [4]. These data are unknown to both the original and modified E-models.

We used commercial VoIP gateway products with analog two-wire interfaces. A network emulator inserted between the gateways controlled the packet delay and packet loss. The combinations of packet loss rate (*Ppl* [%]) and delay (*T<sub>a</sub>* [ms]) were used to test the G.711 PCM codec with a packet loss concealment (PLC) algorithm, the G.729 codec, and the G.711 codec without PLC. The G.729 codec was used with various values of *TELR*, delay, and packet loss rate. The packet length was 20 ms for each codec. For further details of the experimental conditions see [4].



■ **Figure 4.** The relationship between subjective and estimated quality.

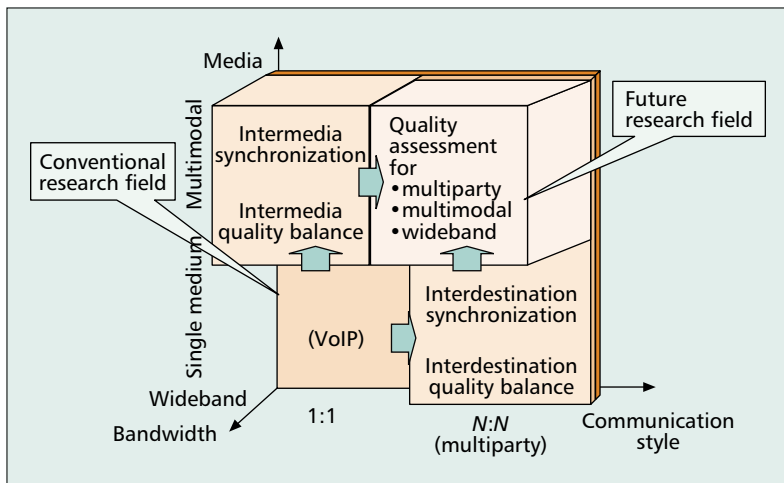
The estimation accuracy of the original and modified E-models is demonstrated in Fig. 4. The figure shows that even if the actual subjective MOS is 2.5, for example, its E-model estimates may vary from 1.2 to 3.4. In terms of root mean square error (RMSE), the modified E-model roughly halves the error in estimation. The coefficients of cross-correlation with subjective quality were 0.763 and 0.793 for the original and modified E-models, respectively.

## PERCEPTUAL QOS FOR NEXT-GENERATION COMMUNICATIONS SERVICES

Figure 5 shows the framework of quality assessment research for next-generation communications services. We think the key words for such services will be *wideband*, *multimodal*, and *multi-party*. The conventional quality assessment methodologies do not appear to be directly applicable to such services.

### TAKING WIDEBAND AND MULTIMODAL INTO ACCOUNT

With the bandwidths of core and access networks rapidly becoming broader, telecommunications applications have more bandwidth available for speech, audio, and video data, which leads to higher-quality services. Assessing the quality of such services based on the simple opinion ratings that have been used for the PSTN and VoIP might be insufficient. That is, we need to compare services not only on a one-



■ Figure 5. A framework for the development of quality assessment research.

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 (e.g., [8]). However, further study is needed to take interactivity and multimodality into account. In addition, establishing methodologies for quality design and management based on such quality assessment methodologies is extremely important.

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. Although intermedia synchronization has been studied for various applications [9], the quality-balance issue has seldom been considered. Such studies should clarify the effects of individual quality parameters, determined on intra- and intermedia bases, on the overall quality. This should lead to the creation of an opinion model for wideband and multimodal services.

### MULTIPARTY CONNECTIONS

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 issues in evaluating the quality of multiparty services are the heterogeneous communications environments of the users and synchronization of user streams.

### CONCLUSION

In this article we describe the state of the art of perceptual QoS assessment methodologies for VoIP systems. We primarily focus on objective

quality assessment, and introduce the past and current activities of the ITU in this area.

We also introduce our own recent work on improving the accuracy of the E-model, which is the opinion model most widely used in network planning for the VoIP systems of today. Experimental results showed that the E-model can be enhanced so that it better estimates users' perceptions of VoIP services.

Finally, we look at perceptual QoS assessment methodologies for the multimedia communications systems of the next generation. We point out three important characteristics of upcoming services: wideband, multimodality, and multiparty connection, and discuss how these will direct future research on perceptual QoS assessment.

### REFERENCES

- [1] N. Osaka et al., "A Model for Evaluating Talker Echo and Sidetone in a Telephone Transmission Network," *IEEE Trans. Commun.*, vol. 40, no. 11, 1992, pp. 1684-92.
- [2] ETSI ETR250, "Speech Communication Quality from Mouth to Ear for 3.1 kHz Handset Telephony Across Networks," July 1996.
- [3] TIA/EIA TSB116, "Voice Quality Recommendations for IP Telephony," Mar. 2001.
- [4] A. Takahashi, "Opinion Model for Estimating Conversational Quality of VoIP," *Proc. IEEE ICASSP '04*, vol. III, May 2004, pp. 1072-75.
- [5] A. Raake, "Speech Quality of Heterogeneous Networks Involving VoIP: Are Time-Varying Impairments Additive to Classical Stationary Ones?," *Proc. 1st ISCA Tutorial and Research Wksp. Auditory Quality of Sys.*, Apr. 2003, pp. 63-70.
- [6] S. Broom and M. Hollier, "Speech Quality Measurement Tools for Dynamic Network Management," *MESAOIN 2003*.
- [7] A. Clark, "Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality," *IP Telephony Wksp. 2001*.
- [8] T. Tachi, S. Iai, and N. Kitawaki, "Proposal of Selection Method of Test Pictures in HDTV Subjective Quality Assessments," *IEEE Multimedia '92*, Apr. 1992, pp. 67-68.
- [9] R. Steinmetz, "Human Perception of Jitter and Media Synchronization," *IEEE JSAC*, vol. 14, no. 1, Jan. 1996.

### BIOGRAPHIES

AKIRA TAKAHASHI [M'04] (takahashi.akira@lab.ntt.co.jp) received a B.S. degree in mathematics from Hokkaido University in Japan in 1988 and his M.S. degree in electrical engineering from the California Institute of Technology in 1993. He joined NTT Laboratories in 1988, and has been engaged in the quality assessment of speech and audio telecommunications. Currently, he is primarily working on the quality assessment of speech over IP networks.

HIDEAKI YOSHINO [M'03] (yoshino.hideaki@lab.ntt.co.jp) received his B.S. and M.S. degrees in information science from the Tokyo Institute of Technology in 1983 and 1985, respectively. He joined NTT Laboratories in 1985 and has been engaged in teletraffic research. From 1990 to 1991 he was a visiting scholar at the University of Stuttgart. He currently serves as a group manager of NTT Laboratories, where he is conducting research on traffic and service quality for future communications services.

NOBUHIKO KITAWAKI [M'87, SM'03] (kitawaki@is.tsukuba.ac.jp) received B.E., M.E. and Ph.D. degrees from Tohoku University, Japan, in 1969, 1971, and 1981. From 1971 to 1997 he was engaged in research on speech and acoustics information processing at NTT Laboratories. He currently serves as a professor at the Institute of Information Sciences and Electronics and as dean of the College of International Studies, University of Tsukuba, Japan. He is a Fellow of IEICE Japan.