ANALYSIS OF RELATIONSHIP BETWEEEN OVERALL QUALITY AND PSYCHOLOGICAL FACTORS AFFECTING HIGH-QUALITY SPEECH COMMUNICATION SERVICES

Hitoshi Aoki and Akira Takahashi

NTT Service Integration Laboratories
3-9-11, Midori-cho, Musashino-shi Tokyo 180-8585 Japan
aoki.hitoshi@lab.ntt.co.jp, takahashi.akira@lab.ntt.co.jp

ABSTRACT

This paper describes the relationship between overall speech quality and individual psychological factors, which are extracted based on conversational quality evaluation experiments, for developing an opinion model for high-quality speech communications. We propose the concept of an opinion model for such services taking into account psychological aspects, as well as overall quality. The experimental results showed that such services had four psychological factors, such as fidelity, activity, harmony, and potency. We also develop part of the model and demonstrate the validity of the proposed concept.

1. INTRODUCTION

Multi-modal high-quality real-time communication services have become more familiar to us through the spread of broadband Internet in recent years. Such services should be evaluated subjectively. From the viewpoints of quality design and management, however, a systematic means for estimating users' satisfaction should be based on physical, network, terminal, and environmental characteristics because subjective quality assessment is time-consuming and expensive. This is called an opinion model[1]. The goal of our study is to develop an opinion model for high-quality real-time speech communication services.

For speech media, service quality is expected to be enhanced by the use of wideband speech, of stereo channels, and hands-free terminals. In quality assessment of such communication services, it is important to assess not only simple one-dimensional opinions, which determines how satisfactorily given telephone connections may be expected to perform, but also impression factors, which are psychological human feelings, in order to express the service quality appropriately. Therefore, it is important for the model to describe the characteristics of the service from various psychological viewpoints.

Although the extraction of psychological factors for timbre has been studied[2][3], the interactive telecommunication has never been examined. Moreover, for the network quality planning and management, what is important is to establish an opinion model that estimates the overall speech communication quality for a combination of quality factors. The study of an opinion model for high-quality speech communication services has never examined.

In this paper, we first proposes the concept of an opinion model for high-quality speech communication services taking into account psychological aspects, as well as overall quality. To establish such an opinion model, we extract psychological factors that affect the overall quality based on extensive psycho-acoustic experiments. We also develop part of the model and demonstrate the validity of the proposed concept.

2. OPINION MODEL FOR HIGH-QUALITY SPEECH COMMUNICATIONS

A computational model that estimates the overall speech communication quality for a combination of quality factors is called the "opinion model[1]." The "E-model" that is standardized in ITU-T (International Telecommunication Union – Telecommunication Sector) Recommendation G.107 is one based on the concept that the amount of psychological degradation can be accumulated on a psychological scale. The E-model is widely used for network planning of telephone or VoIP services. Since this model assumes only the conventional telephone services, however, it cannot evaluate the overall speech communication quality taking into account wideband speech, stereo channels, and hands-free terminals etc. In addition, it cannot represent the characteristics of high-quality services determined by the contribution to individual psychological factors which form the overall quality. Therefore, we introduce psychological factors into the model.

Figure 1 shows the concept of the opinion model for high-quality speech communication services targeted in this

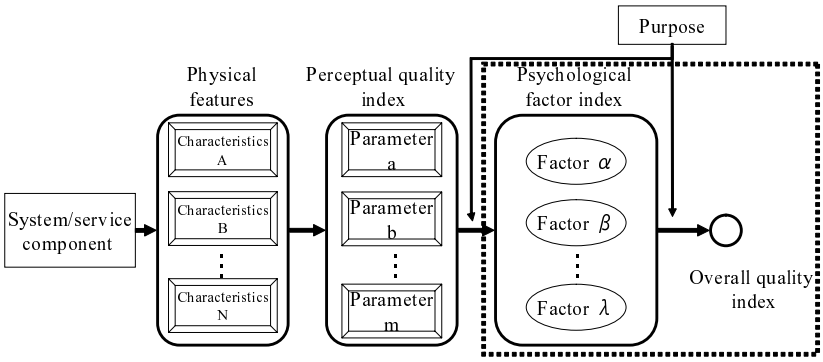

Fig. 1. Concept of opinion model for high-quality speech communication services.

study. The model first inputs the physical features (e.g., speech codec, amount of acoustic echo return loss, packet-loss rate, packet-loss pattern) in accordance with the system requirements and service components of high-quality speech communication services. It then computes the perceptual quality index (e.g., a distortion index is computed from the speech codec, packet-loss rate and packet-loss pattern; an echo quality index is computed from the amount of acoustic echo return loss and network delay time) using the physical features. Furthermore, it computes each of the psychological indices using these perceptual quality indices. Finally, it outputs an overall quality index using these psychological indices. Here, the computational models are adaptively optimized according to the communication purpose, e.g., a chat or a business meeting, when calculating the psychological indices and the overall quality index. Since the model can describe the characteristics of the service from the viewpoints of various psychological aspects, for example, a service planner can design various service qualities while keeping a constant overall quality taking into account the intentions of the users. This paper discusses the structure of psychological factors affecting high-quality speech communication services and propose a model that estimates overall quality based on psychological indices as part of a study on an opinion model for these services.

3. EXTRACTION OF PSYCHOLOGICAL FACTORS

3.1. Conversational subjective quality experiments

In our investigation, we assumed a high-quality speech communication service in which several persons participated in a hands-free communication environment where microphones and loudspeakers were used at two points. The number of users at each location is two. We conducted two different subjective experiments using different parameter settings and subjects. Table 1 shows the experimental parameters in each experiment.

In the first experiment (Exp. #1), we evaluated the effects of speech bandwidth, one-way absolute delay, acoustic echo generated between loudspeakers and microphones, and speech channels on the subjective quality for. In the second experiment (Exp. #2), we evaluated the effect of time clipping, which often occurs due to packet loss. Here, we also investigated the interactive effects among time clipping distortion, speech bandwidth, and speech channels.

Table 1. Experimental settings for Exps. #1 and #2.

Exp. #1	
Speech bandwidth	50 - 3400 Hz, 50 - 7000 Hz, - 20 kHz
Delay [ms]	100, 500
Acoustic echo return loss [dB]	∞, 40 dB@ 1kHz
Time clipping rate[%]	0
Channels	monaural, stereo
Exp. #2	
Speech bandwidth	300 - 3400 Hz, 50 - 7000 Hz, - 20 kHz
Delay [ms]	500, 750
Acoustic echo return loss [dB]	∞, 63, 52
Time clipping rate[%]	0, 3, 10
Channels	monaural, stereo

It is known that the quality evaluation score in a conversational experiment is affected by the coversation task[4]. In this investigation, we used tasks in which subjects are required to guess the shape of a figure or the name of an object by receiving oral information. The interactivity of conversation in these tasks is similar to a free conversation[4].

In these experiments, we evaluated the overall quality and quality impression using the conventional absolute category rating (ACR) method and semantic differential (SD) method proposed by Osgood[2], respectively. The associated testing conditions are summarized in Table 2.

In the evaluation of quality impression by the SD method, the subjects expressed their impression for a given condition using 45 pairs of bipolar adjectives, some of which were chosen from a study on the impression of timbre[2, 3] while others represent the impression received from speech communication services. They were scored from -3 to +3. The subjects evaluated five adjective pairs after talking for 150

Table 2. Subjective experimental conditions.

No. of subjects	40
Duration of conversation	22 min, 30 s per condition
Ambient noise	Hoth noise@35 dB(A)
Conversational task	figure shape and name guessing

s. Such a session was repeated nine times per condition to obtain a total of 45 scores. At the end of each condition the subjects were requested to evaluate the overall quality based on ACR. To control the acoustic echo return loss (AERL), a simulated echo signal was injected into outgoing speech. Here, the echo was simulated by convoluting incoming speech with the impulse response of the acoustic echo path between a microphone and a loudspeaker, which was measured in advance.

We used binaural headphones. Incoming speech was convoluted with the impulse responses of propagation paths from loudspeakers to human ears to simulate free-field listening.

3.2. Factor analysis

Factor analysis by the principal factor method (initial value of commonality was squared multiple correlation) with varimax rotation was performed on the scores obtained in the two experiments. We set the number of primary factors to four taking into account the characteristics of the eigenvalue of an initial solution, and the qualitative interpretation of the extracted factors. To avoid confusion caused by including adjective pairs affecting multiple factors, we tried to simplify the structure of the factor loading matrix in the following manner: adjective pairs were selected so that the absolute value of factor loading was more than 0.4 for only one factor. Factor analysis was repeated until the factor loading matrix met this criterion. This resulted in a smaller set of adjective pairs (26). The factor loading matrix finally obtained is shown in Table 3.

The first factor can be interpreted as *fidelity* since adjective pairs such as *strong-weak, powerful-weak* and *loud-quiet* had high loadings on it. The second factor can be interpreted as *activity* since adjective pairs such as *rounded-angular, delicate-rugged, soft-hard* and *calming-exciting* had high loadings on it. The third factor can be interpreted as *harmony* since adjective pairs such as *tight-loose, tight-slack* and *fast-slow* had high loadings on it. The fourth factor can be interpreted as *potency* since adjective pairs such as *light-heavy* and *cheerful-solemn* had high loadings on it. *Activity* and *potency* are also extracted in the study of evaluation of timbre[2]. *Fidelity* and *harmony*, however, are factors peculiar to interactive high-quality communication services.

Table 3. Factor loading matrix.

Adjective pair	Factor 1	Factor 2	Factor 3	Factor 4
strong - weak	**0.685**	0.153	0.339	-0.145
powerful - weak	**0.679**	0.260	0.323	-0.020
loud - quiet	**0.672**	0.064	0.279	-0.010
3D - 2D	**0.668**	0.239	0.135	0.041
deep - shallow	**0.616**	0.332	-0.039	0.192
thick - thin	**0.593**	0.280	0.138	-0.192
broad - narrow	**0.591**	0.308	-0.053	0.345
exciting - ordinary	**0.573**	-0.188	0.225	-0.047
wide - tight	**0.565**	0.378	0.038	0.374
fluttering - steady	**0.547**	0.193	0.358	0.076
interesting - boring	**0.502**	0.319	0.349	0.030
gentle - violent	**-0.487**	0.329	-0.166	0.115
rounded - angular	0.105	**0.777**	0.017	0.057
delicate - rugged	0.142	**0.720**	0.107	0.120
soft - hard	0.167	**0.685**	-0.090	0.030
calming - exciting	0.007	**0.681**	0.118	0.135
relaxed - tense	0.175	**0.681**	0.165	0.006
mild - intense	-0.071	**0.648**	0.177	0.181
warm - cold	0.313	**0.609**	0.083	-0.055
smooth - jerky	0.304	**0.546**	0.199	0.107
comfortable - embarassing	0.151	**0.438**	0.262	0.100
tight - loose	0.261	0.066	**0.591**	0.095
tight - slack	0.307	0.093	**0.586**	0.197
fast - slow	0.222	0.281	**0.457**	0.101
light - heavy	-0.166	0.062	0.115	**0.564**
cheerful - solemn	0.195	0.267	0.276	**0.523**
proportion of variance explained (%)	18.964	18.570	7.181	4.313
cumulative proportion of variance explained (%)	18.964	37.534	44.715	49.028

4. RELATIONSHIP BETWEEN PSYCHOLOGICAL FACTORS AND OVERALL QUALITY

4.1. Definition of psychological factor index

The psychological factor indices in Figure 1 quantitatively express the fidelity, activity, harmony, and potency of a system and/or service under evaluation. In our analysis, we defined a score for each test condition as the average score over all the subjects. Conventionally, scores are normalized so that their mean and variance are zero and one, respectively. In our investigation, however, we used raw scores in calculating the psychological factor indices. Psychological factor indices ($f_1 - f_4$) are defined as the sum of the average scores over 26 adjective pairs each weighted by the corresponding coefficient in the factor score coefficient matrix. Figure 2 shows the relationship between psychological factor indices and MOS. This figure indicates that they have negative correlation as expected.

4.2. Structuring of MOS with psychological factor indices

We saw in the previous section that high-quality speech communication services can be characterized by four psychological factors. We assumed that MOS, which expresses overall conversational speech quality, could be formulized by a linear combination of the psychological factor indices ($f_1 - f_4$). The following regression equation was obtained

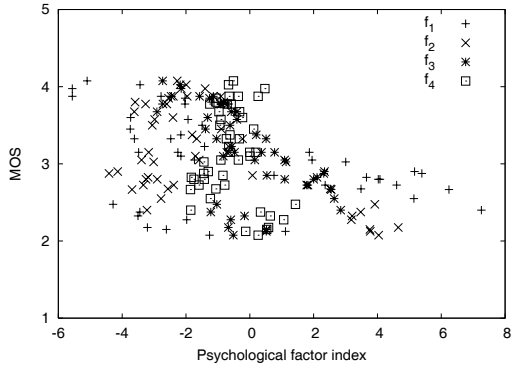

Fig. 2. Relationship between psychological factor index and MOS.

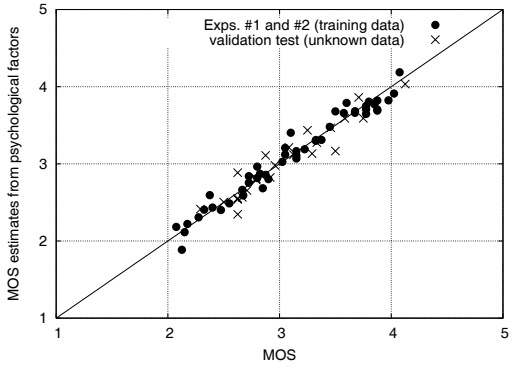

Fig. 3. Relationship between MOS and estimated MOS obtained from psychological factors.

as a result of the multiple regression analysis with respect to $f_1 - f_4$ as explanatory variables and MOS as a criterion variable.

$$MOS = c_0 - c_1 f_1 - c_2 f_2 - c_3 f_3 - c_4 f_4. \qquad (1)$$

Here, c_0, c_1, c_2, c_3, and c_4 are constants and positive real numbers. The coefficient of determination adjusted for the degree of freedom, which indicates the explanation rate of the variation in the explanatory variables, was 0.96, showing that a linear combination of these four indices can explain the overall quality. The ANOVA result also supports this conclusion ($F_{4,43} = 288.2, p < 0.01$). Partial regression coefficients were significant ($p < 0.05$).

5. VALIDITY OF THE PROPOSED MODEL

We tested the validity of the proposed model. Table 4 shows the validation test conditions. The subjective testing procedure was the same as Exps. #1 and #2 as summarized in Table 2 except for the number of subjects. The number of subjects here was 24 and different subjects from those used in Exps. #1 and #2 were used. The validation test evaluated 23 conditions, which are combinations of the parameters in Table 4. This data is unknown to the model. Figure 3 demonstrates the relationship between MOS and estimated MOS by using the model shown in equation (1) for both training and validation data sets. The figure demonstrates that the proposed model well estimates the overall quality even for an unknown data set. The cross-correlation coefficient of the training data set and the unknown data set were 0.98 and 0.95, respectively. As the scatter plot and the cross-correlation coefficient indicate, the proposed model estimates the MOS accurately even for an unknown data set.

Table 4. Experimental settings for validation test.

Speech bandwidth	300 - 3400 Hz, 50 - 7000 Hz
Delay [ms]	190, 300, 600
Acoustic echo return loss [dB]	∞, 63, 54, 50
Time clipping rate[%]	0, 5, 15
Clipping patterns	random, burst
Channels	monaural, stereo

6. CONCLUSION

To develop an opinion model for high-quality speech communication serivices, we analyzed psychological factors that characterize such services based on conversation quality evaluation experiments. The factor analysis extracted four psychological factors. Then, we proposed a model which estimates MOS from the psychological factor indices. Finally, we verified the validity of the model by applying it to an unknown data set.

7. REFERENCES

[1] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," *IEEE Communications Magazine*, pp. 28–34, July 2004.

[2] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The measurement of meaning*, University of Illinois press, 1957.

[3] O. Kitamura, S. Namba, and R. Matsumoto, "Factor analysis research of tone color," in *Proceedings of the 6th International Congress on Acoustics*, 1968, vol. A-5-11.

[4] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE Jounal of Selected Areas in Communications*, vol. 9, no. 4, pp. 586–593, May 1991.