PAPER Special Section on Quality of Communication Networks and Services

Non-reference Objective Quality Evaluation for Noise-Reduced Speech Using Overall Quality Estimation Model

Takeshi YAMADA[†a)], ***Member***, **Yuki KASUYA**[†*], **Yuki SHINOHARA**[†], ***Nonmembers***, ***and*** **Nobuhiko KITAWAKI**[†], ***Fellow***

SUMMARY This paper describes non-reference objective quality evaluation for noise-reduced speech. First, a subjective test is conducted in accordance with ITU-T Rec. P.835 to obtain the speech quality, the noise quality, and the overall quality of noise-reduced speech. Based on the results, we then propose an overall quality estimation model. The unique point of the proposed model is that the estimation of the overall quality is done only using the previously estimated speech quality and noise quality, in contrast to conventional models, which utilize the acoustical features extracted. Finally, we propose a non-reference objective quality evaluation method using the proposed model. The results of an experiment with different noise reduction algorithms and noise types confirmed that the proposed method gives more accurate estimates of the overall quality compared with the method described in ITU-T Rec. P.563.

key words: *non-reference objective quality evaluation, noise-reduced speech, overall quality estimation model, ITU-T Rec. P.835, ITU-T Rec. P.563*

1. Introduction

Hands-free speech communication has gained increased importance in modern communication systems, including teleconferences, in-car phones, and desktop IP phones. However, it still has the serious problem that speech acquired by a hands-free microphone is corrupted by ambient noise. To provide users with natural and intelligible speech, the use of a noise reduction algorithm, which reduces the noise component in the noisy input speech, can be effective. It is, however, well-known that any noise reduction algorithm unavoidably produces speech distortion and residual noise. Here, the critical issue is that the characteristics of these undesired by-products vary according to the noise reduction algorithm used and the type of noise to be reduced. To facilitate QoE (Quality of Experience) design and monitoring, it is essential to establish an objective method that can be used to efficiently evaluate the quality of noise-reduced speech.

In general, objective quality evaluation methods extract the acoustical features that reflect the quality of the input speech, and then estimate the subjective MOS (Mean Opinion Score) from these features. The ways to extract the acoustical features are divided into two types of approach. One is the full-reference approach that requires a separate reference corresponding to the original clean version of the input speech to exactly calculate the spectral distortion. The PESQ (Perceptual Evaluation of Speech Quality) method, standardized as ITU-T Rec. P.862 [1], is the most widely-used of this type. However, the PESQ method can give poor estimates for noise-reduced speech, as we reported in [2]. Egi et al. recently proposed a full-reference objective quality evaluation method for noise-reduced speech and showed its effectiveness [3].

The other approach is the non-reference approach that uses only the input speech for extracting the acoustical features, such as unnaturalness of speech, noise annoyance, and temporal interruption of speech. This approach is especially useful for situations where the reference is unavailable, including QoE monitoring and realtime selection of an optimal noise reduction algorithm. The method described in ITU-T Rec. P.563 [4] belongs to this approach. However, the P.563 method is not appropriate for evaluating noise-reduced speech, as is shown in Sect. 4.

In this paper, we propose a non-reference objective quality evaluation method for noise-reduced speech. The proposed method first involves estimating the quality of only the speech component and the quality of only the noise component from the acoustical features extracted from the input speech, and then estimating the quality (hereafter referred to as 'overall quality') of the input speech from the previously estimated speech quality and noise quality. For the latter estimation, we propose an overall quality estimation model. The unique point of the proposed model is that the estimation of the overall quality is done only from the previous estimates of speech quality and noise quality. In contrast, conventional models, including those used in the PESQ method and the P.563 method, utilize the acoustical features extracted. The proposed model, therefore, has the following advantages.

- The proposed model is applicable to both the full-reference approach and the non-reference approach, because it is independent of the acoustical features.
- We can concentrate on estimating the speech quality and the noise quality. This is easier than estimating the overall quality directly from the acoustical features.
- Even if the acoustical features are added to or changed to improve the performance of estimating the speech quality and the noise quality, the proposed model does not require any modification.

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

[†]The authors are with the Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba-shi, 305-8573 Japan.
[*]Presently with the KDDI Corporation.
a) E-mail: takeshi@cs.tsukuba.ac.jp

DOI: 10.1587/transcom.E93.B.1367

The proposed model was inspired by the subjective evaluation method described in ITU-T Rec. P.835 [5]. In determining the overall quality of noise-reduced speech, subjects tend to weight either the speech component or the noise component. This weighting can lead to wide variance in the overall quality. The aim of the P.835 method is to reduce this uncertainty. In the P.835 method, subjects are instructed to determine the overall quality after individually rating the speech quality and the noise quality. This implies the possibility of estimating the overall quality from the separate estimates of speech quality and the noise quality.

The rest of this paper is organized as follows. Section 2 describes the overall quality estimation model. First, a subjective test is done in accordance with ITU-T Rec. P.835 to obtain the speech quality, the noise quality, and the overall quality of noise-reduced speech. We then define the overall quality estimation model based on the results. Section 3 provides an overview of the P.563 method and the proposed method. Section 4 verifies the effectiveness of the proposed method by comparing it with the P.563 method. Section 5 summarizes the paper and gives suggestions for future work.

2. Overall Quality Estimation Model

2.1 Subjective Test

A subjective test was conducted in accordance with ITU-T Rec. P.835 [5]. Thirty two subjects listened to noise-reduced speech samples through headphones in a sound-proofed room. In evaluating one speech sample, the subjects first focus only on either the speech component or the noise component, and rate its quality on the quality rating scale specified for that component. The subjects then focus only on the other component and rate its quality. The subjects finally rate the overall quality taking account of the preceding two ratings. The quality rating scales used are shown in Table 1. To ensure consistency in the rating, the subjects evaluated speech samples corrupted by the MNRU (Modulated Noise Reference Unit) [6] at the beginning of the subjective test.

Table 2 summarizes the speech samples and the noise types used for the subjective test. We used four speech samples, comprising two male and two female voices, where one speech sample consisted of a pair of Japanese sentences. These speech samples fulfill the requirements on speech samples in ITU-T Rec. P.563 as shown in Table 3. For noise, we used the in-car noise, the exhibition hall noise, and the train noise included in the Denshikyo noise database [7], in addition to white noise, with which most noise reduction algorithms can work well. The noisy speech samples were generated by artificially adding the noise sample to the speech sample at six different values of SNR. We used the four noise reduction algorithms described below, in addition to the reference case of no such algorithm.

(E) Noise suppressor embedded in the EVRC (Enhanced Variable Rate Codec) standardized by EIA/TIA [8],

(S) Noise suppressor based on mutual control of spectral subtraction and spectral amplitude suppression [9], which was the first technique to be endorsed by 3GPP,

(T) Temporal domain singular value decomposition-based noise reduction [10],

(G) Gaussian mixture model-based Wiener filtering [10], and

(N) No noise reduction algorithm.

We chose the noise reduction algorithms so that they can cover a wide range of the speech quality and the noise quality. In the algorithms (E) and (S), the speech quality and the noise quality are relatively balanced. In the algorithm (T), the speech quality tends to be higher than the noise quality. In the algorithm (G), the noise quality tends to be higher than the speech quality. The characteristics of the noise-reduced speech samples vary according to the noise reduction algorithm used and the type of noise to be reduced. The total number of the samples used for the subjective test was 420, that is, 4 (samples) × 5 (algorithms) × 4 (noise types) × 5 (SNRs) plus 4 (samples) × 5 (algorithms) in the Clean speech case.

Figure 1 illustrates the results of the subjective test.

Table 2 Speech samples and noise types used for the subjective test.

Sampling rate	8 kHz
Quantization	16 bit linear PCM
Speech samples	4 pairs of sentences
Noise types	In-car noise, exhibition hall noise train noise, and white noise
SNRs	Clean, 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB

Table 3 Requirements on speech samples in ITU-T Rec. P.563 and actual values.

Requirement		Actual value	
Active speech	≥ 3.0 sec	Max. Min.	5.5 sec 4.4 sec
Signal length	≤ 20.0 sec	Max. Min.	7.6 sec 6.4 sec
Speech activity ratio	≥ 25% and ≤ 75%	Max. Min.	73% 69%

Table 1 Quality rating scales in ITU-T Rec. P.835.

Score	Speech quality	Noise quality	Overall quality
5	NOT DISTORTED	NOT NOTICEABLE	EXCELLENT
4	SLIGHTLY DISTORTED	SLIGHTLY NOTICEABLE	GOOD
3	SOMEWHAT DISTORTED	NOTICEABLE BUT NOT INTRUSIVE	FAIR
2	FAIRLY DISTORTED	SOMEWHAT INTRUSIVE	POOR
1	VERY DISTORTED	VERY INTRUSIVE	BAD

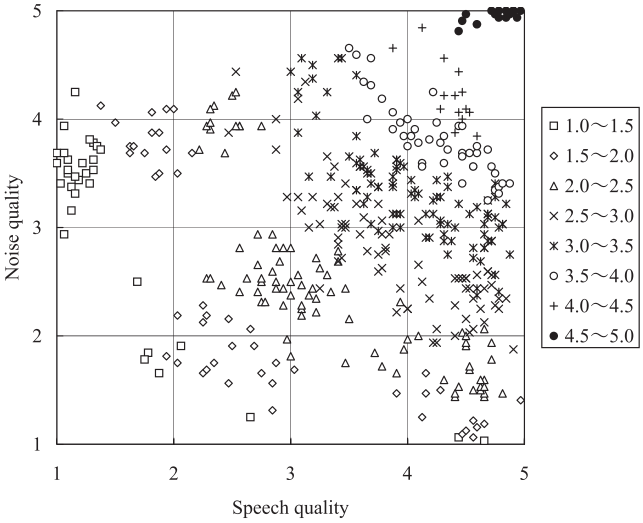

Fig. 1 Results of the subjective test.

The x-axis and the y-axis are the speech quality and the noise quality, respectively. Each point represents the average quality of one speech sample based on its rating by all individual participants. The speech quality and the noise quality can be found from the position of the point and the range of the overall quality from the type of marker.

From Fig. 1 we can see that the subjects determined the overall quality considering the balance of the speech quality and the noise quality. A linear regression line can be drawn from upper left to lower right when focusing on each marker type. The average of the correlation coefficient between the speech quality and the noise quality for each marker type was 0.80. It is especially interesting that the lines are considered to be substantially parallel and equally spaced.

2.2 Proposed Model and Its Effectiveness

Based on the findings described in Sect. 2.1, we define the overall quality estimation model by

$$\textit{Overall quality} = a \times \textit{Speech quality} + b \times \textit{Noise quality} + c, \qquad (1)$$

where a, b, and c are constants. By applying least-square-based data fitting to the results shown in Fig. 1, we determined the overall quality model as

$$\textit{Overall quality} = 0.6303 \times \textit{Speech quality} + 0.6125 \times \textit{Noise quality} - 1.3917. \qquad (2)$$

When applying the proposed model to an objective quality estimation method, first the speech quality and the noise quality are estimated individually from the acoustical features extracted, and then the overall quality is obtained by substituting the estimates of the speech quality and the noise quality in Eq. (2). The advantages of the proposed model mentioned in Sect. 1 come from the fact that the overall quality is not directly estimated from the acoustical features. Although the individual estimation of the speech quality and the noise quality depends on the features used, they do not affect the overall quality estimation model.

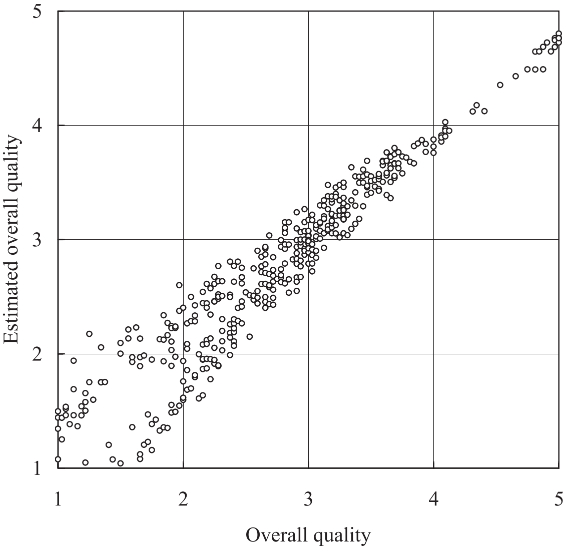

Fig. 2 Relationship between the true overall quality and the estimated overall quality in the closed test.

To verify the effectiveness of the proposed model, we first estimated the overall quality from the speech quality and the noise quality obtained by the subjective test in Sect. 2.1. This corresponds to a closed test in the sense that the test data set is the same as that used for determining Eq. (2). Figure 2 shows the relationship between the true overall quality and the estimated overall quality. The x-axis is the true overall quality and the y-axis the estimated overall quality. The coefficient of determination and the RMSE (Root Mean Square Error) are 0.91 and 0.26, respectively. We can see that the proposed model gives accurate estimates of the overall quality.

An additional subjective test was conducted for an open test. The test conditions were almost the same as those in Sect. 2.1, except that the noisy continuous digit utterances included in the AURORA-2J database [11] were used and the two different noise reduction algorithms described below were adopted instead of the algorithms (E) and (S).

(SS) Spectral subtraction with smoothing of the time direction [12] and
(K) KLT-based comb-filtering [13].

In the algorithms (SS) and (K), the speech quality and the noise quality are relatively balanced. In typical applications, it is required that the speech quality and the noise quality are well-balanced. We therefore decided to replace the algorithms (E) and (S) to the algorithms (SS) and (K). We estimated the overall quality from the speech quality and the noise quality obtained by this subjective test. Figure 3 illustrates the relationship between the true overall quality and the estimated overall quality. The coefficient of determination and the RMSE are 0.95 and 0.26, respectively, so we

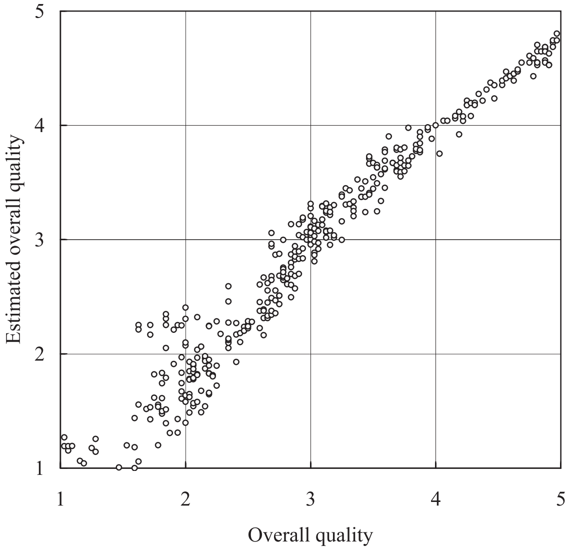

Fig. 3 Relationship between the true overall quality and the estimated overall quality in the open test.

conclude that the proposed model also achieved good performance in the open test.

3. Non-reference Objective Quality Evaluation

3.1 ITU-T Rec. P.563

This section briefly explains the non-reference objective quality evaluation method described in ITU-T Rec. P.563 [4]. The P.563 method had demonstrated acceptable accuracy for quality factors, including codec characteristics, IP packet loss, and environmental noise at the sending side, but not including the effect of noise reduction.

Figure 4 shows an overview of the P.563 method. In the P.563 method, the acoustical features associated with the speech component and the noise component in the input speech are first extracted, and then the intermediate speech quality is calculated. Finally, the estimation of the overall quality is made from these features and the intermediate speech quality. The number of the features is about 50. The overall quality estimation model used is based on the optimal approximation of the relationship between the overall quality, the intermediate speech quality, and the acoustical features. The shortcoming of this model is that it needs to be reconstructed when adding a new acoustical feature to cope with the effect of noise reduction. In contrast, the proposed model utilizes only the intermediate quality corresponding to the speech quality and the noise quality described in ITU-T Rec. P.835 to decide the overall quality.

3.2 Proposed Method

Figure 5 shows an overview of the proposed method. In the proposed method, the acoustical features from the input speech are first extracted. In this paper, the acoustical features of basic speech descriptors, unnatural speech, noise analysis and interruptions/mutes described in P.563 are used. The use of a common set of features enables us to make a fair comparison of the performance of the proposed method with that of the P.563 method. We use all 27 features of basic speech descriptors and unnatural speech for estimating the speech quality, and all 24 features of noise analysis and interruptions/mutes for estimating the noise quality.

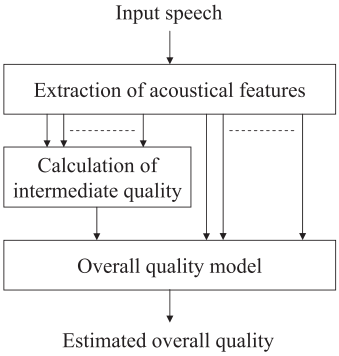
Input speech
Extraction of acoustical features
Calculation of intermediate quality
Overall quality model
Estimated overall quality

Fig. 4 Overview of the P.563 method.

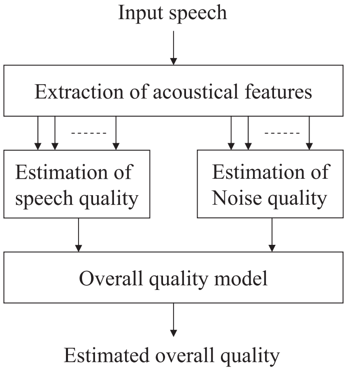
Input speech
Extraction of acoustical features
Estimation of speech quality
Estimation of Noise quality
Overall quality model
Estimated overall quality

Fig. 5 Overview of the proposed method.

Second, from these features, the speech quality and the noise quality are estimated separately. The estimation is done using a simple linear regression given by

$$Speech\ quality = \sum_{n=1}^{27}(\alpha_n X_n) + \beta, \tag{3}$$

$$Noise\ quality = \sum_{m=1}^{24}(\gamma_m Y_m) + \delta, \tag{4}$$

where X_n is the n-th feature that reflects the quality of the speech component and Y_m the m-th feature that reflects the quality of the noise component. The constants, α_n, β, γ_m, and δ are determined by least-square-based data fitting.

Finally, the overall quality is obtained by substituting the estimates of the speech quality and the noise quality in Eq. (2).

4. Verification of the Proposed Method

First, the speech quality obtained by the subjective test in

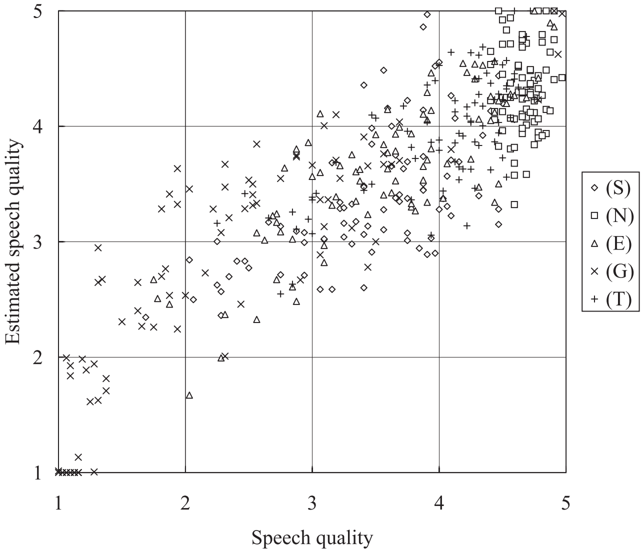

Fig. 6 Relationship between true speech quality and estimated speech quality.

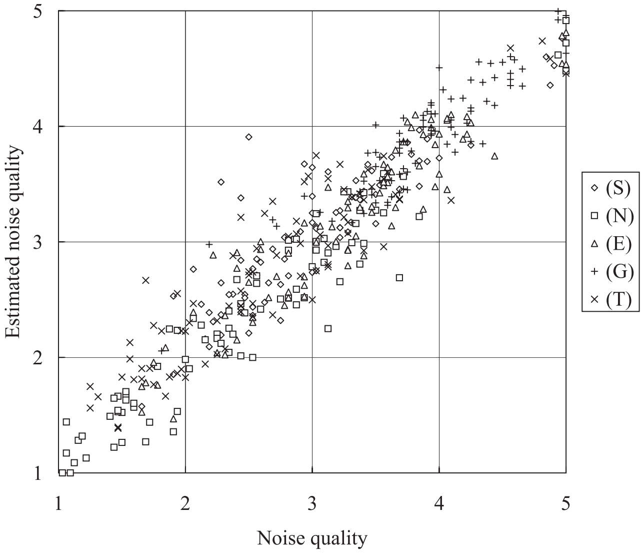

Fig. 7 Relationship between true noise quality and estimated noise quality.

Sect. 2.1 was estimated using the proposed method. Figure 6 illustrates the relationship between the true speech quality and the estimated speech quality. The marker type indicates the noise reduction algorithm used. The coefficient of determination and the RMSE are 0.73 and 0.54, respectively. In fact, the wide variance can be seen in Fig. 6.

Next, the noise quality obtained by the subjective test in Sect. 2.1 was also estimated by the proposed method. Figure 7 shows the relationship between the true noise quality and the estimated noise quality. The coefficient of determination and the RMSE are 0.89 and 0.30, respectively. We can see that the estimation of the noise quality is more successful than that of the speech quality.

Finally, we verified the effectiveness of the proposed method by comparing it with the P.563 method. The overall quality obtained by the subjective test in Sect. 2.1 was estimated by the proposed method and also by the P.563 method. In the proposed method, the overall quality was obtained by substituting the estimates of the speech quality and the noise quality mentioned above in Eq. (2). Figure 8 illustrates the relationship between the true overall quality and the estimated overall quality for both methods. The coefficient of determination and the RMSE for the P.563 method are 0.56 and 0.69, respectively, while for the proposed method they are 0.71 and 0.47, respectively. We can see that the proposed method gives more accurate estimates compared with the P.563 method. As described in Sect. 3.2, the acoustical features used in the proposed method were the same as those in the P.563 method. This fact shows the effectiveness of the overall quality estimation model used in the proposed method. However, the absolute performance of the proposed method is insufficient for practical use. This problem would be solved by improving the estimation of the speech quality.

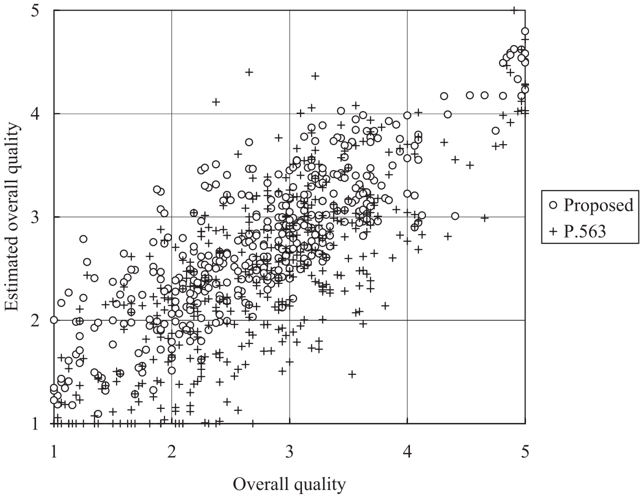

Fig. 8 Relationship between the true overall quality and the overall quality estimated by the proposed method and the P.563 method.

5. Conclusion

This paper described a non-reference objective quality evaluation method for noise-reduced speech. First, a subjective test was done in accordance with ITU-T Rec. P.835 to obtain the speech quality, the noise quality, and the overall quality of noise-reduced speech. Based on the results, we then proposed an overall quality estimation model. The unique point of the proposed model is that the estimation of the overall quality is made only from separate estimates of speech quality and noise quality, whereas, in contrast, conventional models utilize the acoustical features extracted. Finally, we proposed a non-reference objective quality evaluation method using the proposed model. The results of an experiment with different noise reduction algorithms and noise types confirmed that the proposed method gives more accurate estimates of the overall quality than the method described in ITU-T Rec. P.563.

As future work, we have plans to improve the performance of the proposed method and to develop a full-reference objective quality evaluation method for noise-reduced speech using the proposed model.

Acknowledgments

The authors would like to thank Dr. K. Takeda, Dr. N. Kitaoka, and Dr. M. Fujimoto for providing their noise reduction programs. This work was partially supported by Grant-in-Aid for Scientific Research (19300271) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. This work was conducted using the AURORA-2J database.

References

[1] 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.

[2] T. Yamada, M. Kumakura, and N. Kitawaki, "Subjective and objective quality assessment of noise reduced speech signals," Proc. IEEE-EURASIP International Workshop on Nonlinear Signal and Image Processing, NSIP2005, pp.328–331, May 2005.

[3] N. Egi, H. Aoki, and A. Takahashi, "Objective quality evaluation method for noise-reduced speech," IEICE Trans. Commun., vol.E91-B, no.5, pp.1279–1286, May 2008.

[4] ITU-T Rec. P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," May 2004.

[5] ITU-T Rec. P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," Nov. 2003.

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

[7] Denshikyo noise database, http://research.nii.ac.jp/src/list/detail.html#JEIDA-NOISE.

[8] 3GPP2 C.S0014-A Version 1.0, "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems," April 2004.

[9] S. Furuta, S. Takahashi, and K. Nakajima, "A noise suppression method based on mutual control of spectral subtraction and spectral amplitude suppression," Systems and Computers in Japan, vol.8, no.14, pp.90–102, Sept. 2007.

[10] M. Fujimoto and Y. Ariki, "Combination of temporal domain SVD based speech enhancement and GMM based speech estimation for ASR in noise — Evaluation on the AURORA2 task," Proc. European Conference on Speech Communication and Technology, EUROSPEECH2003, pp.1781–1784, 2003.

[11] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.535–544, March 2005.

[12] N. Kitaoka and S. Nakagawa, "Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task," Proc. International Conference on Spoken Language Processing, ICSLP2002, pp.465–468, 2002.

[13] S.-J. Park, M. Ikeda, K. Takeda, and F. Itakura, "Improvement of the ASR robustness using combinations of spectral subtraction and KLT based adaptive comb-filtering," IPSJ SIGNotes, SLP-44-3, pp.13–18, 2002.

Takeshi Yamada was born in Osaka, Japan. He received the B.Eng. degree from Osaka City University, Japan, in 1994, and the M.Eng. and Dr.Eng. degrees from Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. He is presently an associate professor with the Graduate School of Systems and Information Engineering, University of Tsukuba, Japan. His research interests include robust speech recognition, sound scene understanding, multi-channel signal processing, and media quality assessment. He is a member of the IEEE, the IPSJ, and the ASJ.

Yuki Kasuya received the B. Eng. and M. Eng. degrees from University of Tsukuba, Japan, in 2007 and 2009, respectively. He is presently with the KDDI Corporation, Japan. His research field includes objective quality assessment of hands-free speech communication.

Yuki Shinohara received the B.Eng. degree from University of Tsukuba, Japan, in 2008. He is presently a master course student at University of Tsukuba, Japan. His research field includes objective quality assessment of hands-free speech communication.

Nobuhiko Kitawaki was born in Aichi, Japan. 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 Provost of the School of Social and 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, 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.