FULL-REFERENCE OBJECTIVE QUALITY EVALUATION FOR NOISE-REDUCED SPEECH CONSIDERING EFFECT OF MUSICAL NOISE

Yuki Fujita[†], *Takeshi Yamada*[†], *Shoji Makino*[†‡], *and Nobuhiko Kitawaki*[*]

[†]Graduate School of Systems and Information Engineering, University of Tsukuba
[‡]Life Science Center of TARA, University of Tsukuba
[*]Department of Global Activities, University of Tsukuba
Tsukuba-shi, Ibaraki, 305–8573, Japan

ABSTRACT

In this paper, we propose a full-reference objective quality evaluation method for noise-reduced speech. In the proposed method, acoustical features are extracted from input noise-reduced speech and its original clean version, and then the speech quality and the noise quality are estimated separately from these features. Finally, the overall quality is determined from the estimates of the speech quality and the noise quality. The basic idea of the proposed method is similar to that of the non-reference method that we previously proposed but has two attractive features. The first is to use an improved overall quality estimation model. The other is to consider the effect of musical noise, which is known as annoying residual noise with tonal components. Experimental results confirmed that the proposed method gives accurate estimates of the overall quality.

Index Terms— Noise-reduced speech, objective quality evaluation, ITU-T Rec. P.835, overall quality estimation model, musical noise

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 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 not estimate precisely for noise-reduced speech [2]. The other approach is the non-reference approach that uses only the input speech for extracting the acoustical features. We previously proposed a non-reference objective quality evaluation method for noise-reduced speech using an overall quality estimation model and showed its effectiveness [3].

In this paper, we propose a full-reference objective quality evaluation method for noise-reduced speech. The basic idea of the proposed method is similar to that of our previous non-reference method [3] but has two attractive features. The first is to use an improved overall quality estimation model. The other is to consider the effect of musical noise, which is known as annoying residual noise with tonal components. We evaluate the effectiveness of the proposed method.

2. OVERVIEW OF THE PROPOSED METHOD

The proposed method, as well as our previous method, was inspired by the subjective evaluation method described in ITU-T Rec. P.835 [4], in which subjects are instructed to determine the overall quality after individually rating the speech quality of only the speech component and the noise quality of only the noise component.

Fig. 1 illustrates the overview of the proposed method. In the proposed method, the acoustical features are first extracted from the input noise-reduced speech and its original clean version. Second, from these features, the speech quality and the noise quality are estimated separately. Finally, the

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

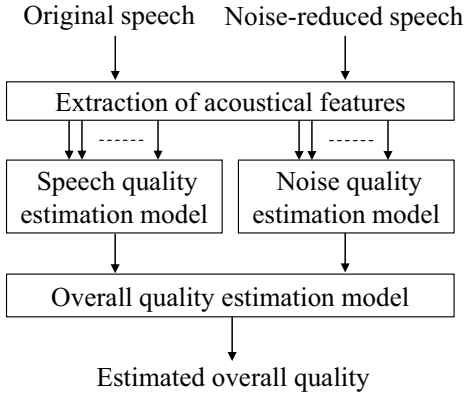

Fig. 1. Overview of the proposed method.

overall quality is determined by substituting the estimates of the speech quality and the noise quality in the overall quality estimation model.

The proposed method has two attractive features as mentioned above. The first is to use an improved overall quality estimation model, which is described in Sect. 3. The other is to consider the effect of the musical noise, which is discussed in Sect. 4.

3. IMPROVEMENT OF THE OVERALL QUALITY ESTIMATION MODEL

3.1. Subjective Test

A subjective test was conducted in accordance with ITU-T Rec. P.835 [4]. 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.

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. For noise, we used the in-car noise, the exhibition hall noise, and the train noise included in the Denshikyo noise database [5], 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.

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

Sampling rate	8kHz
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

(E) Noise suppressor embedded in the EVRC (Enhanced Variable Rate Codec) standardized by EIA TIA [6],

(S) Noise suppressor based on mutual control of spectral subtraction and spectral amplitude suppression [7], which was the first technique to be endorsed by 3GPP,

(T) Temporal domain singular value decomposition-based noise reduction [8],

(G) Gaussian mixture model-based Wiener filtering [8], 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. 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) $\times$ 5 (algorithms) $\times$ 4 (noise types) $\times$ 5 (SNRs) plus 4 (samples) $\times$ 5 (algorithms) in the Clean speech case.

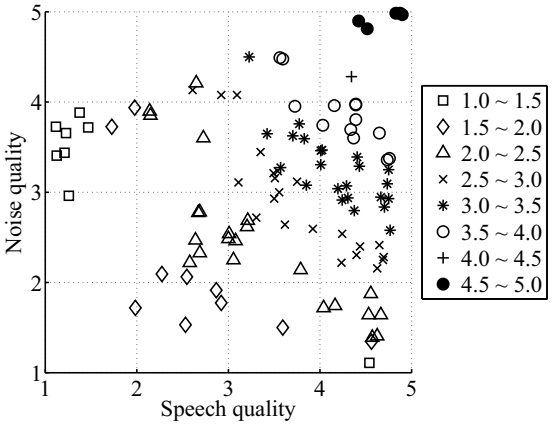

Fig. 2. Results of the subjective test.

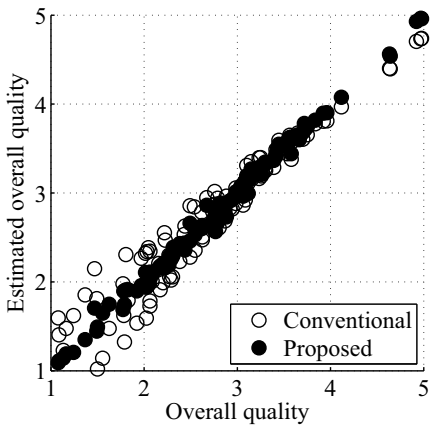

Fig. 3. Relationship between true overall quality and estimated overall quality.

Fig. 2 illustrates the 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 four speech samples based on their 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. 2 we can see that the subjects determined the overall quality considering the balance of the speech quality and the noise quality.

3.2. Overall quality estimation model

Based on the result described in Sect. 3.1, we previously defined the overall quality estimation model[3] by

$$Q_o = 0.6303 \times Q_s + 0.6125 \times Q_n - 1.3917, \quad (1)$$

where Q_o is the overall quality, Q_s the speech quality and Q_n the noise quality.

From Fig. 2, however, we can see that the subjects tend to rate the overall quality low when either the speech quality or the noise quality is especially low. Considering the finding, we propose the new overall quality estimation model expressed by

$$Q_o = \sum_{i=1}^{2} c_{s_i}(Q_s)^i + \sum_{i=1}^{2} c_{n_i}(Q_n)^i + c_{sn}Q_sQ_n + c, \quad (2)$$

where $c_{s_1} = 0.3582$, $c_{s_2} = -0.0696$, $c_{n_1} = -0.0751$, $c_{n_2} = -0.0271$, $c_{sn} = 0.2228$ and $c = 0.5091$. The constants in (1) and (2) were determined by applying least-square-based data fitting to the results of the subjective test described in Sect. 3.1.

3.3. Evaluation

To verify the effectiveness of the proposed model, we estimated the overall quality from the speech quality and the noise quality obtained by the subjective test in Sect. 3.1.

Table 3. Speech samples and noise types used for the subjective test.

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

Fig. 3 shows the relationship between the true overall quality and the estimated overall quality. The coefficient of determination and the RMSE (Root Mean Square Error) for the conventional model are 0.93 and 0.23, respectively, while for the proposed model they are 0.99 and 0.07, respectively. We can see that the proposed model gives more accurate estimates compared with the conventional model.

4. EFFECT OF MUSICAL NOISE

4.1. Subjective Test

To investigate the effect of the musical noise on the speech quality and the noise quality, we first conducted a subjective test in accordance with ITU-T Rec. P.835. The test conditions are similar to those in Sect. 3.1 as shown in Table 3. In this subject test, the noise reduction algorithms (E), (S), and (N), in addition to (W) the time-domain Wiener filter-based noise reduction. We adjusted the internal parameters of the algorithms (S) and (W) so that the musical noise is generated. The total number of the samples used for the subjective test was 208, that is, 4 (samples) $\times$ 4 (algorithms) $\times$ 3 (noise types) $\times$ 4 (SNRs) plus 4 (samples) $\times$ 4 (algorithms) in the Clean speech case.

Table 5. Correlation coefficients among the three qualities.

	Speech quality	Noise quality	Musical noise quality
Speech quality	1.00	-	-
Noise quality	0.34	1.00	-
Musical noise quality	0.09	0.78	1.00

Table 4. Quality rating scale for the musical noise quality.

Score	Musical noise quality
5	NOT NOTICEABLE
4	SLIGHTLY NOTICEABLE
3	NOTICEABLE BUT NOT INTRUSIVE
2	SOMEWHAT INTRUSIVE
1	VERY INTRUSIVE

After that, we explained what the musical noise is to the subjects and conducted an additional subjective test to obtain the musical noise quality for the speech samples used in the above test. The subjects were instructed to focus only on the musical noise component and rate the musical noise quality on the quality rating scale shown in Table 4.

4.2. Results of the subjective test

Table 5 show the correlation coefficients among the speech quality, the noise quality, and the musical noise quality. Note that the cases where the musical noise quality is more than 4.0 are removed in advance. From Table 5, we can see that there is a strong correlation between the musical noise quality and the noise quality, while a weak correlation between the musical noise quality and the speech quality. Fig. 4 illustrates the relationship between the musical noise quality and the noise quality. Focusing on the points that the musical noise score is less than 4.0, there are variations in the overall relationship. This implies that the subject determines the noise quality considering the effect of the musical noise.

4.3. Estimation of the speech quality and the noise quality

In the proposed method, the following five acoustical features are extracted from the input noise-reduced speech and its original clean version to estimate the speech quality and the noise quality.

- Additive distortion in the speech period (X_1)
- Additive distortion in the non-speech period (X_2)
- Subtractive distortion in the speech period (X_3)
- Subtractive distortion in the non-speech period (X_4)
- Average log power in the non-speech period (X_5)

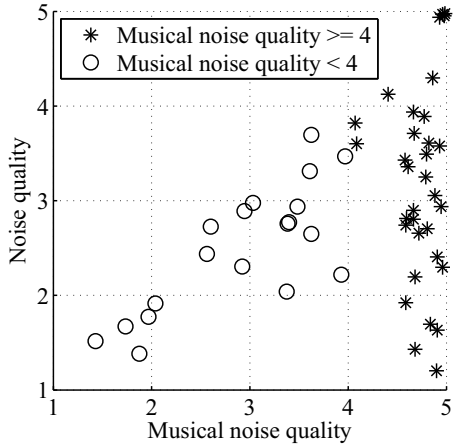

Fig. 4. Relationship between musical noise quality and noise quality in the subjective test.

The distortion measure used is the same as that in the PESQ method. The speech quality estimation model is expressed by

$$Q_s = \sum_{i=1}^{5}(\alpha_i X_i) + \alpha_6, \tag{3}$$

where Q_s is the speech quality and X_i the i-th feature mentioned above. $\alpha_1 = -0.1195$, $\alpha_2 = 0.2396$, $\alpha_3 = -0.2593$, $\alpha_4 = -0.0303$, $\alpha_5 = 0.5389$ and $\alpha_6 = 4.0771$. The noise quality estimation model is also expressed by

$$Q_n = \sum_{i=1}^{6}(\beta_i X_i) + \beta_7, \tag{4}$$

where Q_n is the noise quality and X_i the i-th feature. $\beta_1 = -0.1141$, $\beta_2 = -0.0013$, $\beta_3 = -0.0087$, $\beta_4 = 0.0667$, $\beta_5 = -0.6978$, $\beta_6 = -1.1215$ and $\beta_7 = 5.7035$. Note that the amount of the musical noise, X_6, is added as the 6th feature, which corresponds to the logarithmic kurtosis ratio [9]. This is based on the result described in Sect. 4.2. The constants α_i and β_i in (3) and (4) are determined by applying the least-square based data fitting.

5. EVALUATION OF PROPOSED METHOD

The speech quality and the noise quality obtained by the subjective test in Sect. 4.1 were estimated by using the speech

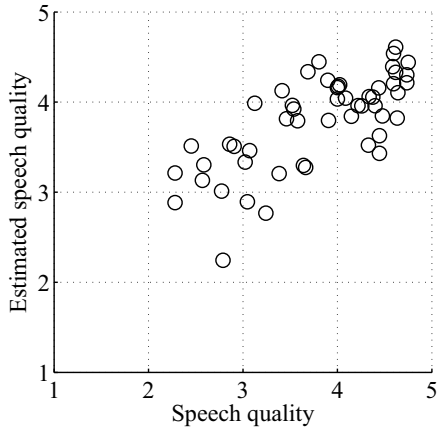

Fig. 5. Relationship between true speech quality and estimated speech quality.

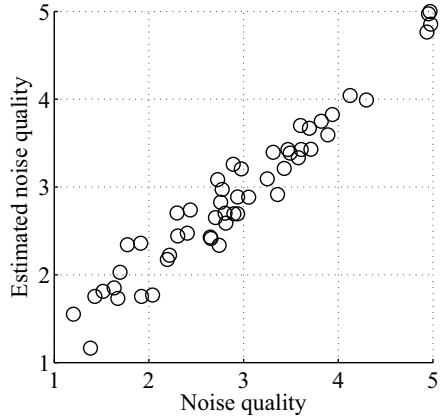

Fig. 6. Relationship between true noise quality and estimated noise quality.

quality estimation model and the noise quality estimation model, respectively. Fig. 5 shows the relationship between the true speech quality and the estimated speech quality. The coefficient of determination and the RMSE are 0.52 and 0.50, respectively. Fig. 6 also represents the relationship between the true noise quality and the estimated noise quality. The coefficient of determination and the RMSE are 0.94 and 0.24, respectively. We can see that the estimation of the noise quality is more successful than that of the speech quality.

Finally, the overall quality obtained by the subjective test in Sect. 4.1 was estimated by the proposed method. In the proposed method, the overall quality was determined by substituting the estimates of the speech quality and the noise quality mentioned above. Fig. 7 shows the relationship between the true overall quality and the estimated overall quality. The coefficient of determination and the RMSE for the proposed

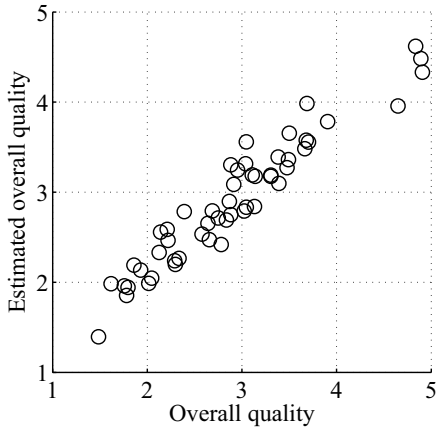

Fig. 7. Relationship between true overall quality and estimated overall quality.

method are 0.91 and 0.26, respectively, while for the PESQ method they are 0.85 and 0.50, respectively. The RMSE value of 0.26 for the proposed method is close to the target value of 0.13, which is defined by the 95 percent confidence interval for the overall scores obtained in the subjective test. The results confirmed that the proposed method gives accurate estimates of the overall quality.

6. CONCLUTIONS

In this paper, we proposed the full-reference objective quality evaluation method for noise-reduced speech. First, we described the improved overall quality estimation model. Second, we clarified that the noise quality rather than the speech quality is affected by the musical noise and then, based on this finding, designed the way to estimate the speech quality and the noise quality. Finally, we confirmed that the proposed method gives accurate estimates of the overall quality. As future work, we have a plan to improve the estimation accuracy of the speech quality estimation model.

7. ACKNOWLEDGMENT

This work was supported in part by The Telecommunications Advancement Foundation.

8. 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] T. Yamada, Y. Kasuya, Y. Shinohara, and N. Kitawaki, "Non-reference objective quality evaluation for noise-reduced speech using overall quality estimation model," *IEICE Transactions on Communications*, vol. E93-B, no. 6, pp. 1367–1372, June 2010.

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

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

[6] 3GPP2 C.S0014-A Version 1.0, "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems," April 2004.

[7] 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. 99–102, April 2004.

[8] 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. Eurospeech2003*, pp. 1781–1784, 2003.

[9] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," *Proc. of International Workshop on acoustic Echo and Noise Control*, 2008.