SNR-dependent Background Noise Compensation of PESQ Values for Cellular Phone Speech

Kengo Fujita, Tsuneo Kato, Hideaki Yamada and Hisashi Kawai

KDDI R&D Laboratories Inc.
2-1-15 Ohara, Kamifukuoka, Saitama, 356-8502 Japan
fujita@kddilabs.jp

Abstract

To evaluate the speech quality of actual cellular phone systems with an objective assessment, PESQ values were compared with MOS values for speech with background noises via four cellular phone systems used in Japan. As PESQ value errors were observed to be SNR-dependent, two SNR-dependent background noise compensation methods for PESQ values are proposed. Applying the compensation methods to the speech for four cellular phone systems, the RMSEs between MOS and compensated PESQ values were reduced to less than half of the original RMSEs for all four cellular phone systems. They were equal to the level of RMSE of MOS values.

1. Introduction

Evaluation of speech quality is primarily based on a subjective quality assessment. For evaluation of telephone speech quality, the MOS (Mean Opinion Score) test is standardized in ITU-T Recommendation P.800 [1]. However, MOS tests are time-consuming and expensive. To save the time and expenses, various objective assessments, such as PSQM [2], PAMS [3] and PESQ [4], were proposed as substitute assessments for the MOS test [5]. Among them, PESQ standardized in ITU-T Recommendation P.862 has the strongest correlation with MOS values and is widely used.

To evaluate the speech quality of cellular phone systems, evaluation of speech with background noise is essential because cellular phones are frequently used in noisy environments. In cellular phone systems, input speech with background noises first passes through a noise suppressor, then is coded by a low bit-rate speech codec and transmitted to a cellular network. ITU-T Recommendation P.862 describes that noises on input of speech and cellular phone codecs (e.g. G.729, AMR, EVRC) are factors for which PESQ has demonstrated acceptable accuracy. Nevertheless, the effect of noise suppressors is a factor for further study. Takahashi et al. reported the effect of measurement noise [6]. However the effectiveness of PESQ for speech with background noise in actual cellular phone systems is unclear.

In this paper, the effectiveness of PESQ for speech with background noise in actual cellular phone systems is evaluated, and SNR-dependent background noise compensation methods for reducing estimation errors of PESQ values are proposed. In the comparison of MOS and PESQ values for speech with background noise, MOS and PESQ values still had a strong correlation. The correlation coefficient was 0.94, which was nearly equal to the level described in P.862. However there were errors in the PESQ values, which depended on SNR and the noise types. Therefore, two noise compensation methods, one of which depends on SNR and the other depends on the ratio of loudness [7] of speech to that of noise (SNRL in short) are proposed to reduce errors.

In Section 2, the accuracy of MOS values is first investigated to get a target level of the estimation error of PESQ values. Next, PESQ values are compared with MOS values. In Section 3, two compensation methods of PESQ are proposed, and examined with speech via the four cellular phone systems.

2. Accuracy of MOS Values and PESQ Values

2.1. Accuracy of MOS values

The accuracy of MOS needs to be evaluated since MOS values are used as references for PESQ evaluation. To know the accuracy of MOS values for cellular phone speech with background noise, first, two MOS tests with different, yet sufficient subjects were carried out, and two MOS values for common conditions were compared.

Table 1 shows experimental conditions for the MOS tests. The ACR (Absolute Category Rating) method prescribed in IUT-T P.800 was used. The background conditions were five types of background noise with three different SNRs plus one clean environment (i.e. sixteen conditions in total). Four speech codecs were tested. The tested speech was a standard Japanese speech database. For each test condition, 40 speech samples by four male and four female talkers were used. The source speech samples were input to cellular terminals set on HATS (Head And Torso Simulator) [8]. The tested speech samples were digitally recorded through the cellular phone network and ISDN, and filtered with modified IRS receive characteristics [9] for the handset. Subjects listened with one ear that they preferred in a sound proof room. For each tested background noise condition, four votes were made by 80 subjects, for a total of 320 votes. A MOS value was calculated as the average of the 320 votes.

The accuracy of MOS values was evaluated by RMSE (Root Mean Square Error) calculated with equation (1)

\[ \text{RMSE}_{\text{MOS}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\text{MOS}_i(i) - \text{MOS}_o(i))^2} \]  

where \( \text{MOS}_i(i) \) and \( \text{MOS}_o(i) \) represent the MOS values for condition \( i \) in two different MOS tests, and \( N \) is the number of compared conditions. As two MOS tests have two same tested codecs, \( N \) is 32 (16 noise conditions for two codecs).

The RMSE of 32 conditions was 0.12. This value gives a target level of the errors for PESQ values.
2.2. Evaluation of PESQ values for cellular phone speech with background noise

PESQ values were compared with MOS values for cellular phone speech tested in the previous section. The PESQ values are calculated as the average of the 40 speech samples used for the MOS tests.

Table 2 shows correlation coefficients and RMSEs between MOS and PESQ values. Generally, correlation coefficients are still large with background noise. However RMSEs are from 2 to 5 times the RMSE of MOS values (0.12).

Figure 1 shows the correlation between MOS and PESQ values for speech of Codec A on different noise conditions shown in Table 1. Three plotted points along with a line correspond to 9dB, 15dB and 21dB of a noise type from left to right, respectively. Although PESQ values should ideally be equal to MOS values, PESQ values have errors from MOS values for all noise types. However three plotted points on a single noise type individually creates a straight line.

The proportionality constants of all noise types are smaller than 1.0, which causes the errors.

3. SNR-dependent Background Noise Compensation of PESQ Values

3.1. SNR dependencies of PESQ values

To apply the PESQ to speech in background noise conditions, it needs to be compensated for so that:
- Reduce PESQ value errors.
- Reduce the noise type dependencies.

Figure 2 shows the correlation between SNRs of codec input and PESQ value errors. The SNRs were calculated on the signal sources as the average of the 40 speech samples used for the MOS tests. The SNR was the ratio of average power of speech to that of noise. The average of the errors monotonously increases with SNR. Therefore the error can be reduced by adding an SNR-dependent compensation term. However there are still large variances due to noise types.

Low bit-rate speech codecs used in cellular phone systems have a noise suppressor at its front-end for transmitting speech clearly. As the noise suppressor does not work equally for all types of noise, the SNRs of codec output change depending on the noise even if SNRs of codec input are equal. The SNR of codec output can affect the subjective assessment of speech more directly than the SNR of codec input. Figure 3 shows the correlation between SNRs of codec output and the errors of PESQ values. In Figure 4, a similar correlation is seen as shown in Figure 3.
3.2. SNR-dependent compensation of PESQ values

Based on the experimental results, two compensation methods are proposed.

1) An additional linear function of SNR of codec output

   \[ MOS - PESQ \approx c_0 \cdot SNR + c_1 \]  

   where \( c_0 \) and \( c_1 \) are empirical coefficients obtained from experimental results related to SNRs and the errors. Hence the compensated PESQ value \( cPESQ_{SNR} \) is derived from,

   \[ cPESQ_{SNR} = PESQ + c_1 \cdot SNR + c_0 \]  

2) An additional linear function of SNRL of codec output

   In method 1), PESQ value errors are approximated using the following linear function,

   \[ MOS - PESQ \approx c_0 \cdot SNR + c_1 \cdot SNR + c_0 \]  

   In terms of method 2), SNRL is derived from the ratio of loudness of speech to that of noise. The loudness was calculated using B&K software, which implements Zwicker’s study on loudness. Similarly to equation (3), compensated PESQ value \( cPESQ_{SNRL} \) is derived from,

   \[ cPESQ_{SNRL} = PESQ + c_1 \cdot SNRL + c_0 \]  

3.3. Evaluation of the compensation methods

To evaluate the proposed compensation methods, a cross validation test was conducted on the MOS test results described in Section 2.1. Subjects of the MOS test were divided into a development set and a test set. The coefficients \( c_0 \) and \( c_1 \) are calculated with the development set and compensated PESQ values \( cPESQ_{SNR} \) and \( cPESQ_{SNRL} \) are evaluated with the development set. Ten pairs of the development set and the test set were created.

Figure 5 shows correlations between MOS-PESQ, MOS-\( cPESQ_{SNR} \) and MOS-\( cPESQ_{SNRL} \) for Codec A. In Figure 5, the distributions of both \( cPESQ_{SNR} \) and \( cPESQ_{SNRL} \) values are closer to the dashed straight line than PESQ values, which indicates the errors of \( cPESQ_{SNR} \) and \( cPESQ_{SNRL} \) values are smaller than that of PESQ values. Both \( cPESQ_{SNR} \) and \( cPESQ_{SNRL} \) values estimate MOS values more accurately than PESQ values, and estimation errors were reduced. The other nine pairs of the development set and the test set showed similar correlations. Estimation of \( cPESQ_{SNR} \) and \( cPESQ_{SNRL} \) values was also accurate for Codec B, C and D.

Table 3 shows the correlation coefficients between MOS-PESQ, MOS-\( cPESQ_{SNR} \) and MOS-\( cPESQ_{SNRL} \). The coefficients were calculated as the average of ten pairs of the development set and the test set. For all the codecs, \( cPESQ_{SNR} \) and \( cPESQ_{SNRL} \) values have stronger correlations with MOS values than PESQ values. Table 4 shows RMSEs between MOS-PESQ, MOS-\( cPESQ_{SNR} \) and MOS-\( cPESQ_{SNRL} \). RMSEs between MOS and \( cPESQ_{SNR} \) values and MOS and \( cPESQ_{SNRL} \) values were reduced to less than half. As a result, the RMSEs were reduced to the level of RMSE of MOS values (0.12) described in Section 2.1, and the noise type dependencies were reduced. The effects of \( cPESQ_{SNR} \) and \( cPESQ_{SNRL} \) values were almost equal.
Conclusions

In this paper, PESQ values were compared with MOS values for cellular phone speech with background noise via four cellular phone systems used in Japan. PESQ values were found to have errors from MOS values on all noise types. The errors depended on the noise type. To reduce errors and the noise type dependencies, two background noise compensation methods for PESQ values were proposed. One method depends on SNR, and the other depends on the ratio of loudness of speech and that of noise. The evaluation of the compensation methods showed that both compensated PESQ values estimate MOS values more accurately than PESQ values. The RMSEs between MOS and compensated PESQ values were reduced to less than half of the original RMSEs for all the four cellular phone systems. They were equal to the level of RMSE of MOS values.

References