Subjective and Objective Quality Assessment for Noise Reduced Speech

Nobuhiko Kitawaki, and Takeshi Yamada

Graduate School of Systems and Information Engineering,
University of Tsukuba
1-1-1, Tennoudai, Tsukuba-shi, 305-8573 Japan.
{kitawaki, takeshi}@cs.tsukuba.ac.jp}

ETSI Workshop on Speech and Noise in Wideband Communication,
22nd and 23rd May 2007 – Sophia Antipolis, France

Abstract- Recent development in telecommunication services, such as VoIP in NGN and car communications has become increasingly necessary for use of hands-free communication system using separate microphones and loudspeakers. Hands-free system has been largely affected by noisy circumstances. This paper describes experimental results and perspectives for subjective and objective quality assessment of noise reduced speech.

In this paper, noisy circumstances, such as subway, babble, car, and exhibition were considered, and as the noise reduced algorithms, spectral subtraction with smoothing of the time direction, temporal domain SVD-based speech enhancement, GMM-based speech estimation, and KLT-based comb-filtering were used.

Noise reduced speech were subjectively assessed by word intelligibility and opinion rating methods. Objective quality assessment was studied from viewpoints of objective quality measure using perceptual evaluation of speech quality (PESQ) in the auditory domain. First, this paper describes subjective and objective quality assessment from viewpoints of the opinion rating methods, and then, word intelligibility.

Index Terms- Noisy Environment, Noise Reduction, Noisy Speech Recognition, Articulation and MOS, Objective Measure, PESQ

1. Introduction

Hands-free speech communication is becoming increasingly necessary for teleconferences, in-car phones, and PC-based IP telephony. In communication systems, most users prefer not a close-talk (headset) microphone but a distant-talk microphone. However, there is a problem that speech acquired by the distant-talk microphone is generally corrupted by ambient noise. To solve this problem, many systems adopt noise reduction algorithms as a front-end processing.

There is a trade-off between the speech distortion and the residual noise. Aggressive algorithms tend to increase the speech distortion while the noise component is suppressed considerably. Furthermore, the characteristics of the speech distortion and the residual noise are mutually different according to the principles of the noise reduction. It is therefore indispensable to evaluate the noise-reduced speech [1].

This paper describes experimental results and perspectives for subjective and objective quality assessment of noise reduced speech. Noisy circumstances, such as subway, babble, car, and exhibition were considered, and as the noise reduced algorithms, spectral subtraction with smoothing of the time direction, temporal domain SVD-based speech enhancement, GMM-based speech estimation, and KLT-based comb-filtering were used [2].

Noise reduced speech were subjectively assessed by word intelligibility and opinion rating methods. The subjective MOS (Mean Opinion Score) corresponds to the “overall” quality in ITU-T Rec. P.835 “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm” [3]. In this paper, the word intelligibility test is performed by using word lists which take word difficulty into account [4] since the word intelligibility strongly depends on the word difficulty.

Then, objective quality assessment was studied from viewpoints of objective quality measure using perceptual evaluation of speech quality (PESQ) in the auditory domain [5]. The PESQ was verified for evaluating speech distorted by codecs, filtering, variable delay, and short localized distortions. However, it has not been clarified whether the PESQ is applicable to evaluating the noise-reduced speech. We therefore investigate the applicability of the PESQ by experiments using four noise reduction algorithms.

First, this paper describes subjective and objective quality assessment from viewpoints of the opinion rating methods, and then, word intelligibility.

2. Opinion Rating

2.1 Subjective Test

The subjective test was performed in accordance with ITU-T Rec. P.800 [6]. Twenty subjects listened noisy speech samples and noise-reduced speech samples through a headphone in a soundproofing room and rated them by the five-level ACR (Absolute Category Rating).

Table 1 shows the speech samples used for the subjective test. The speech samples of 2 male and 2 female were randomly selected from the test set A of the AURORA-2J [7]. The utterances are the Japanese seven-digit numbers. The noise-reduced speech samples were prepared through the noise reduction algorithms shown in Table 2.

Table 1 Speech samples used for the subjective test.

Speaker	2 male and 2 female
Speech sample	4 samples for each noise
Utterance	Japanese seven-digit numbers
Noise	Subway, Babble, Car, Exhibition
SNR	Clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB
Channel	G.712

Fig.1 shows the subjective MOS for “Subway” noise for an example, where the x-axis is the SNR of the input noisy samples. We can see that the subjective MOS varies according to the noise reduction algorithms. In this test, the baseline (B) gives the highest subjective MOS. The reason is that the noise reduction algorithms used in this test were originally developed for automatic speech recognition, which requires to reduce the residual noise rather than the speech distortion. On the other hand, it can be seen that the subjective MOS is less sensitive to the noise types compared with the noise reduction algorithms.

Table 2 Noise-Reduction Algorithms

Symbol	Noise-reduction algorithm	Ref.
(B)	Baseline	
(G)	GMM-based speech estimation	[8]
(K)	KLT-based comb-filtering	[9]
(S)	Spectral subtraction with smoothing of the time direction	[10]
(T)	Temporal domain SVD-based speech enhancement	[8]

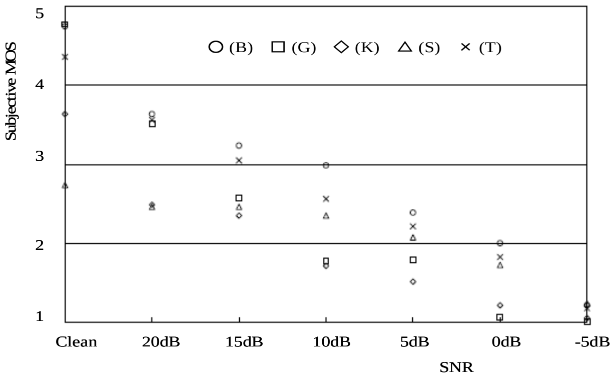

Fig.1 Subjective MOS for “Subway” noise.

2.2 Objective Estimation

Fig.2 shows the relationship between the subjective MOS and the PESQ MOS estimated by using PESQ measure, where each point is specified by the noise reduction algorithms, the noise types, and the SNRs. The coefficient of determination, R^2, and the RMSE (Root Mean Square Error) for each noise reduction algorithm are summarized in Table 3. We can see that the PESQ MOS correlates relatively well with the subjective MOS. For (B) and (T), however, it can be seen that the subjective MOS is under-estimated. In (B) and (T), the speech distortion is small but the residual noise is loud compared with the other algorithms. This fact implies that the PESQ tend to evaluate the noise effect to an excessive degree especially when the residual noise is loud. This problem would be solved by the improvement of the distortion evaluation part in the PESQ.

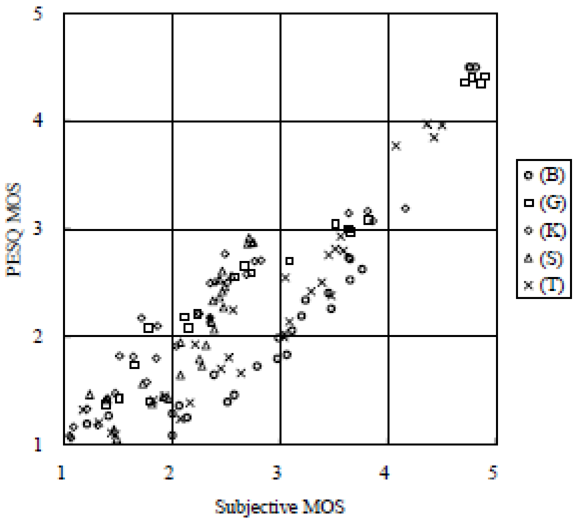

Fig.2 Subjective MOS and estimated PESQ MOS.

Table 3 Coefficient of determination and RMSE.

Symbol	R^2	RMSE
All	0.84	0.56
(B)	0.89	0.87
(G)	0.96	0.39
(K)	0.90	0.32
(S)	0.83	0.29
(T)	0.90	0.68

3. Word Intelligibility

3.1 Word Intelligibility Test

The word intelligibility strongly depends on the word difficulty. We therefore adopted word lists developed by Sakamoto et al. [4]. In the individual word lists, the word difficulty is controlled appropriately by word familiarity, which is the index of how subjectively familiar the word is. All entry words are classified into four word familiarity ranks:

(F4) 7.0-5.5 high word familiarity,
(F3) 5.5-4.0 middle-high word familiarity,
(F2) 4.0-2.5 middle-low word familiarity, and
(F1) 2.5-1.0 low word familiarity.

There are 20 word lists for each word familiarity rank, where one word list includes 50 words. Table 4 shows the speech samples used for the word intelligibility rest.

Table 4 Speech samples used for the intelligibility test.

Speaker	1 male
Subjects	20
Speech sample	500 for each familiarity rank
Utterance	Japanese words of four mora
Noise	Subway, Car
SNR	Clean, 20dB, 15dB, 10dB, 5dB, 0dB
Channel	G.712

Fig.3 shows the word intelligibility for low word familiarity rank as an example in the case of Car noise, where the x-axis is the SNR of the input noisy speech sample. We can see that the word intelligibility for (T) is mostly higher than that for (B), where (T) gives little effect on the speech component while the residual noise is relatively loud. As the experimental results, the word familiarity strongly affects the word intelligibility.

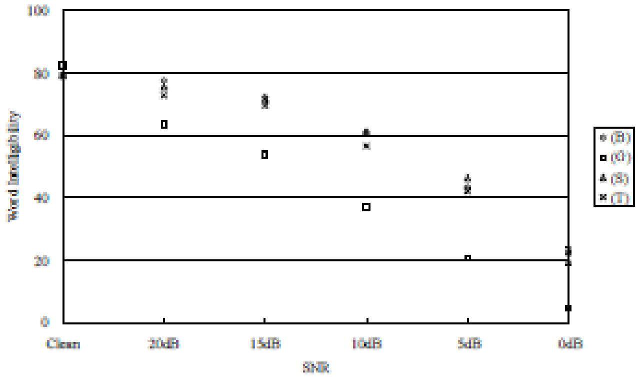

Fig.3 Word intelligibility for low familiarity rank.

3.2 Objective Estimation

In this paper, the word intelligibility is estimated by using a relation between word intelligibility and PESQ MOS shown in Fig.4 as an example. The estimators were prepared for the individual word familiarity ranks without distinguishing the noise reduction algorithms and the noise types. Table 5 shows the coefficient of determination and RMSE for each word familiarity rank. Finally, Fig.5 shows the true word intelligibility and the estimated word intelligibility as an example.

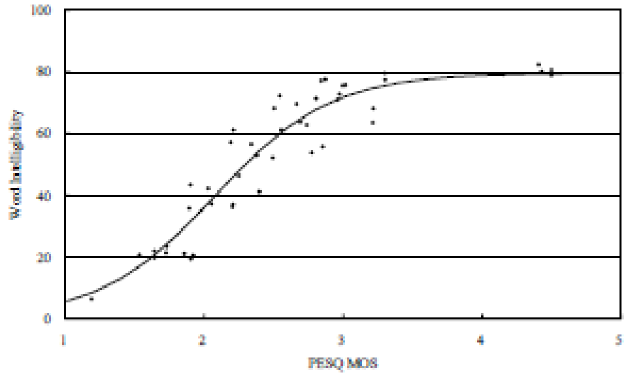

Fig.4 Word intelligibility and PESQ MOS for low familiarity.

Table 5 Coefficient of determination and RMSE.

Familiarity	R^2	RMSE
(F1)	0.90	7.0
(F2)	0.91	6.6
(F3)	0.89	5.3
(F4)	0.88	4.2

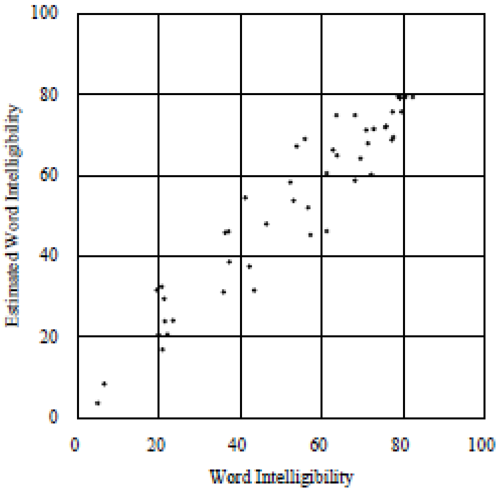

Fig.5 Evaluated and estimated word intelligibility.

4 Conclusion

This paper described the subjective and objective quality assessment for noise-reduced speech from the viewpoints of opinion rating and word intelligibility. The results confirmed that the PESQ MOS correlates relatively well with the subjective MOS. We proposed the objective test methodology for estimating the word intelligibility from the PESQ MOS and evaluated its effectiveness.

Reference

[1] Takeshi Yamada, Masakazu Kumakura, and Nobuhiko Kitawaki, “Performance Estimation of Speech Recognition System Under Noise Conditions Using Objective Quality Measure and Artificial Voice,” IEEE Trans. Audio, Speech, and Language Processing, Vol.14, No.6, pp.2006-2013, Nov. 2006.

[2] Takeshi Yamada, Masakazu Kumakura, and Nobuhiko 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] ITU-T Rec. P.835, “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm,” Nov. 2003.

[4] S. Sakamoto, Y. Suzuki, S. Amano, and T. Kondo, “Speech intelligibility by use of new word-lists with controlled word familiarities and a phonetic balance,” Proc. International Congress on Sound and Vibration, ICSV8, pp.2461-2466, July 2006.

[5] 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.

[6] ITU-T Rec. P.800, “Methods for subjective determination of transmission quality,” Aug. 1996.

[7] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishimura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, “AURORA-2J: An evaluation framework for Japanese noisy speech recognition,” IEICE Trans. Information and Systems, Vol.E88-D, No.3, pp.535-544, Mar.2005.

[8] M. Fujimoto, 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, Sep.2003.

[9] S.J. Park, M. Ikeda, 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, Dec. 2002.

[10] N. Kitaoka, and S. Nakagawa, “Evaluation of spectral subtraction with smoothing of time direction on the AURORA2 task,” Proc. ICSLP2002, pp.465-468, Sep.2002.