LETTER

Objective Estimation of Word Intelligibility for Noise-Reduced Speech

Takeshi YAMADA[†a)], ***Member***, **Masakazu KUMAKURA**[††], ***Nonmember***, ***and*** **Nobuhiko KITAWAKI**[†], ***Fellow***

SUMMARY It is essential to ensure a satisfactory QoS (Quality of Service) when offering a speech communication system with a noise reduction algorithm. In this paper, we propose a new obejective test methodology for noise-reduced speech that estimates word intelligibility by using a distortion measure. Experimental results confirmed that the proposed methodology gives an accurate estimate with independence of noise reduction algorithms and noise types.

key words: *objective estimation, word intelligibility, noise reduction, PESQ*

1. Introduction

Hands-free speech communication is becoming increasingly necessary for teleconferencing, in-car phones, and PC-based IP telephony. In these communication systems, most users prefer not to use a close (headset) microphone but a more distant microphone. However, this results in a problem, because speech acquired by a distant microphone is generally corrupted by ambient noise. To solve this problem, many systems adopt a noise reduction algorithm as a front-end processing stage.

The aim of the noise reduction algorithm is to remove the noise component from noisy input speech without affecting the speech component. However, there is a trade-off between the speech distortion which results from this processing and the residual noise. It is therefore essential to establish an objective test methodology for noise-reduced speech.

In general, noise-reduced speech is evaluated from the viewpoints of subjective quality and intelligibility. Steady progress has been made toward the objective estimation of subjective quality in recent years [1], [2]. However, researches for intelligibility have not yet been conducted. In this paper, we propose a new objective test methodology that estimates word intelligibility by using a distortion measure, and evaluate its effectiveness [3].

2. Word Intelligibility Test

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

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

There are 20 word lists for each word familiarity rank, and each list contains 50 words.

Table 1 Speech samples used for the word intelligibility test.

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

Table 1 describes the speech samples used for the word intelligibility test. We used the speech database assembled in accordance with the word lists mentioned above, which has been released by NTT Advanced Technology Corporation. The speech samples of 1 male were selected from this database, and 10 word lists for each word familiarity rank were selected randomly. The utterances were Japanese words of four mora. The speech samples were mixed with the noise samples included in the AURORA-2J database [5]. In this test, the noise-reduced speech samples were prepared using the following noise reduction algorithms: (B) Baseline (No noise reduction was implemented for this case.), (G) GMM-based speech estimation [7], (S) Spectral subtraction with smoothing in the time domain [8], and (T) Temporal domain SVD-based speech enhancement [7]. The characteristics of the noise-reduced speech samples differ according to the noise reduction algorithm used. The total number of the speech samples was 96,000, that is, 4 (familiarity ranks) × 500 (utterances) × 2 (noise types) × 6 (SNR values) × 4 (algorithms).

The word intelligibility test was performed in a sound-proof room. Subjects listened to the noisy speech samples and the noise-reduced speech samples through headphones, and then wrote down the words they heard. The number of the subjects was twenty (10 male and 10 female). The subjects were divided into two groups: one for the Subway noise and the other for the Car noise. The number of speech

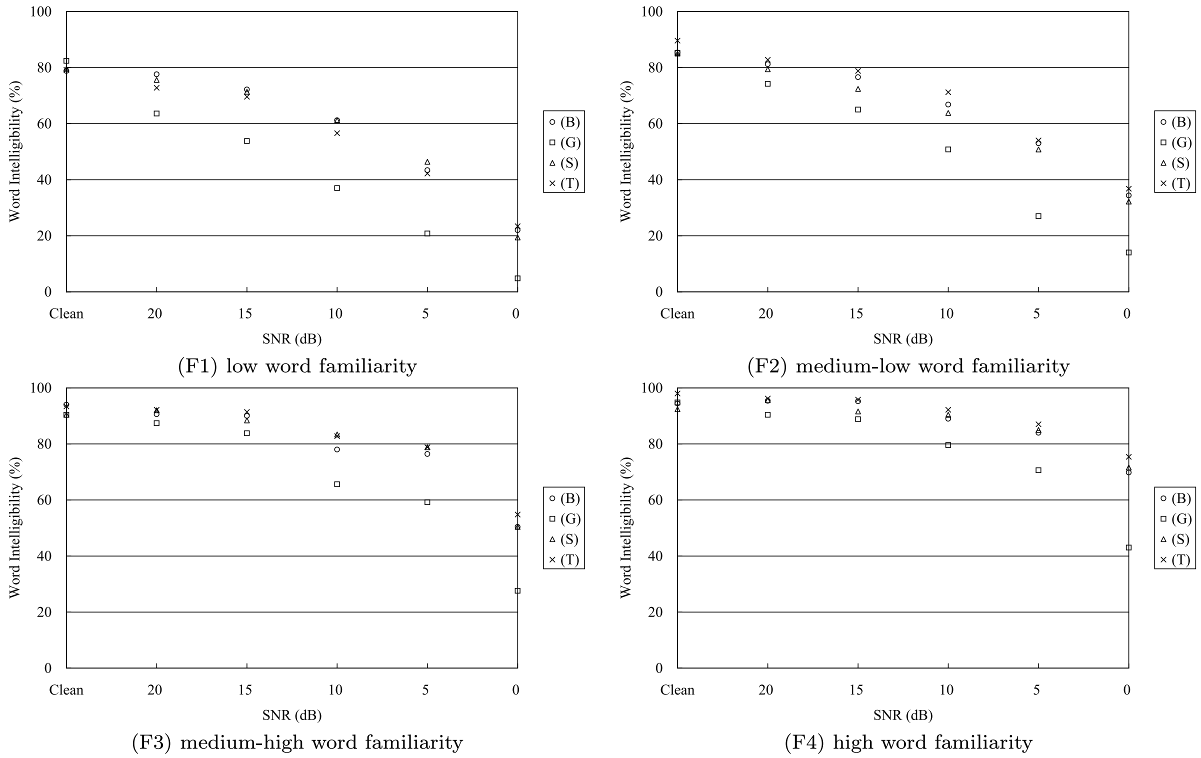
Manuscript received April 10, 2008.

[†] The authors are with the Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba-shi, 305-8573 Japan.

[††] The author is with Fujitsu Limited, Kawasaki-shi, 211-8588 Japan.

a) E-mail: takeshi@cs.tsukuba.ac.jp

DOI: 10.1093/ietcom/e91–b.12.4075



(F1) low word familiarity

(F2) medium-low word familiarity

(F3) medium-high word familiarity

(F4) high word familiarity

Fig. 1 Word intelligibility for each word familiarity rank in the case of the Car noise.

samples for each subject was 4,800 (96 word lists), that is, 4 (familiarity ranks) × 50 (utterances) × 1 (noise type) × 6 (SNR values) × 4 (algorithms). In this test, each individual word list was used only once.

Figure 1 shows the word intelligibility for each word familiarity rank in the case of the Car noise, where the x-axis is the SNR of the noisy input speech samples. The word intelligibility, which is defined as the ratio of the number of words recognized correctly to the total number of words, was calculated for each word list. It can be seen that the word familiarity strongly affects the word intelligibility. In particular, the degradation of word intelligibility due to the noise increases as the word familiarity rank becomes low. We can also see that the word intelligibility for (T) is generally higher than that for (B), except for the low word familiarity. The reason is that (T) causes little degradation of the speech component, although the residual noise is relatively loud. On the other hand, (G) seriously degrades the word intelligibility in most cases. The reason is that, while reducing the noise component effectively, (G) increases the speech distortion, especially under low SNR conditions. Similar results were obtained for the case of the Subway noise.

3. Estimation of Word Intelligibility

The proposed methodology estimates the word intelligibility by using a distortion measure. In this paper, we adopt PESQ (Perceptual Evaluation of Speech Quality), standardized by the ITU-T as Rec. P.862 [6], as a distortion measure. PESQ measures a speech distortion and outputs its value as a PESQ MOS (Mean Opinion Score) ranging from −0.5 to 4.5.

The word intelligibility is estimated by using the estimator expressed in the following form.

$$y = \frac{a}{1 + e^{-b(x-c)}}, \tag{1}$$

where y and x represent the estimated word intelligibility and the PESQ MOS, respectively, and a, b, and c are constants which are determined according to the relationship between the word intelligibility and the PESQ MOS. The estimators used were optimized for each individual word familiarity rank with independence of noise reduction algorithms and noise types.

Figure 2 shows the relationship between the word intelligibility and the PESQ MOS for each word familiarity rank. In this figure, each point represents the result for one of the noise reduction algorithms, with one of the noise types, and one of the SNR values. The solid lines are the estimators mentioned above. The constants used in the estimators for each word familiarity rank are summarized in Table 2. Figure 3 shows the relationship between the true word intelligibility and the estimated word intelligibility for each word familiarity rank. The coefficient of determination, R^2, and the RMSE (Root Mean Square Error) for each word familiarity rank are summarized in Table 3. From Fig. 3 and Table 3, it can be seen that the estimated word intelligibility correlates well with the true word intelligibility, while the word familiarity rank slightly affects the estimation accuracy. These results confirmed that the word intelligibility can be estimated well from the PESQ MOS with independence of noise re-

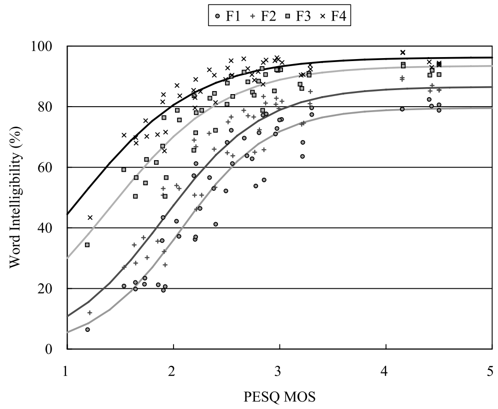


Fig. 2 Relationship between the word intelligibility and the PESQ MOS for each word familiarity rank.

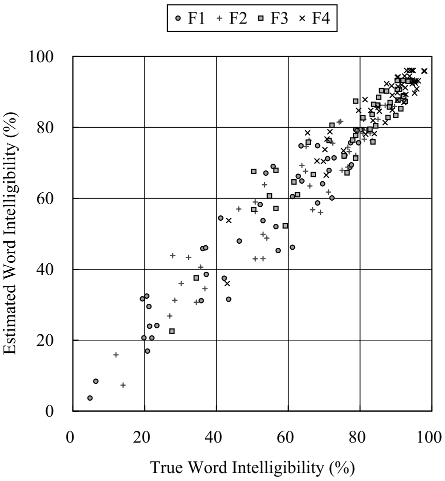


Fig. 3 Relationship between the true word intelligibility and the estimated word intelligibility for each word familiarity rank.

Table 2 Values of constants used in the estimators for each word familiarity rank.

	a	*b*	*c*
(F1)	79.6095	2.4113	2.0783
(F2)	86.6129	2.1388	1.9107
(F3)	93.5692	1.8478	1.4058
(F4)	96.3088	1.7859	1.0848

duction algorithms and noise types. As future work, it is necessary to evaluate the applicability of our methodology to unknown noise reduction algorithms and noise types.

4. Conclusion

We have proposed the obejective test methodology for noise-reduced speech that estimates the word intelligibility by using the distortion measure, PESQ. The experimental results confirmed that the proposed methodology gives an accurate estimate with independence of noise reduction algorithms and noise types.

Table 3 Coefficient of determination and RMSE for each familiarity rank.

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

Acknowledgments

The authors would like to thank Dr. N. Kitaoka, and Dr. M. Fujimoto for providing their noise reduction programs. This work was supported in part by the Ministry of Public Management, Home Affairs, Posts and Telecommunications of Japan.

References

[1] T. Yamada, M. Kumakura, and N. Kitawaki, "Subjective and objective quality assessment of noise reduced speech signals," Proc. NSIP, pp.328–331, May 2005.

[2] N. Egi, H. Aoki, and A. Takahashi, "Objective quality evaluation method for noise-reduced speech," Proc. MESAQIN, June 2007.

[3] T. Yamada, M. Kumakura, and N. Kitawaki, "Word intelligibility estimation of noise-reduced speech," Proc. ICSLP, pp.169–172, Sept. 2006.

[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. ICSV8, pp.2461–2466, July 2001.

[5] 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.

[6] 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.

[7] 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. EUROSPEECH, pp.1781–1784, Sept. 2003.

[8] N. Kitaoka and S. Nakagawa, "Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task," Proc. ICSLP, pp.465–468, Sept. 2002.