Word Intelligibility Estimation of Noise-Reduced Speech

Takeshi Yamada, Masakazu Kumakura, Nobuhiko Kitawaki

Graduate School of Systems and Information Engineering
University of Tsukuba, Tsukuba, Japan
takeshi@cs.tsukuba.ac.jp

Abstract

It is indispensable to establish an objective test methodology for noise-reduced speech. This paper proposes a new methodology which estimates word intelligibility of the noise-reduced speech from PESQ MOS (subjective MOS estimated by the PESQ). To evaluate the effectiveness of the proposed methodology, a word intelligibility test of the noise-reduced speech was performed by using four noise reduction algorithms and word lists which take word difficulty into account, and then the word intelligibility was estimated by the proposed methodology. The results confirmed that the word intelligibility can be estimated well from the PESQ MOS without distinguishing the noise reduction algorithms and the noise types.

Index Terms: word intelligibility, objective estimation, PESQ.

1. Introduction

Hands-free speech communication is becoming increasingly necessary for teleconferences, in-car phones, and PC-based IP telephony. In these communication systems, most users prefer not to use a close-talk (headset) microphone but a more distant microphone. However, there is the problem that 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 is to remove the noise component from the noisy input speech without affecting the speech component. However, there is a trade-off between the speech distortion and the residual noise. For example, aggressive algorithms are effective in suppressing the noise component, but also tend to increase the speech distortion. Furthermore, the characteristics of the speech distortion and the residual noise vary according to the principle of the noise reduction used. It is therefore essential to establish an objective test methodology for the noise-reduced speech.

Recently, we have shown that the PESQ (Perceptual Evaluation of Speech Quality), which was standardized by the ITU-T as Rec. P.862 [1], gives a relatively accurate estimate of the subjective MOS (Mean Opinion Score) of noise-reduced speech [2]. However, the noise-reduced speech should be evaluated from the viewpoint of intelligibility in addition to the subjective quality. We therefore propose a new methodology which estimates the word intelligibility of the noise-reduced speech from the PESQ MOS (the subjective MOS estimated by the PESQ).

The rest of this paper is organized as follows. Section 2 describes a word intelligibility test on the noise-reduced speech obtained using four noise reduction algorithms. In this paper, word lists which take word difficulty into account [3] are used, since word intelligibility depends strongly on word difficulty. Section 3 gives an overview of the proposed methodology. The effectiveness of the proposed methodology is evaluated in terms of the consistency between the true word intelligibility and the estimated word intelligibility. Section 4 summarizes the contributions of this paper.

Table 1: *The 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 moras
Noise	Subway, Car
SNR	Clean, 20dB, 15dB, 10dB, 5dB, 0dB
Channel	G.712

2. Word intelligibility test

2.1. Test conditions

Word intelligibility depends strongly on word difficulty. We therefore adopted word lists developed by Sakamoto *et al.* [3]. In each individual word list, 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 the following four word familiarity ranks:

(F4) 7.0 to 5.5 (high word familiarity),
(F3) 5.5 to 4.0 (middle-high word familiarity),
(F2) 4.0 to 2.5 (middle-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 shows the speech samples used for the word intelligibility test. We used a speech database collected 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 moras. The speech samples were mixed with the noise samples included in the AURORA-2J [4]. In this test, the noise-reduced speech samples were prepared using the following noise reduction algorithms.

(B) Baseline (Noise reduction was NOT implemented for this case.)
(G) GMM-based speech estimation [5]
(S) Spectral subtraction with smoothing of the time direction [6]
(T) Temporal domain SVD-based speech enhancement [5]

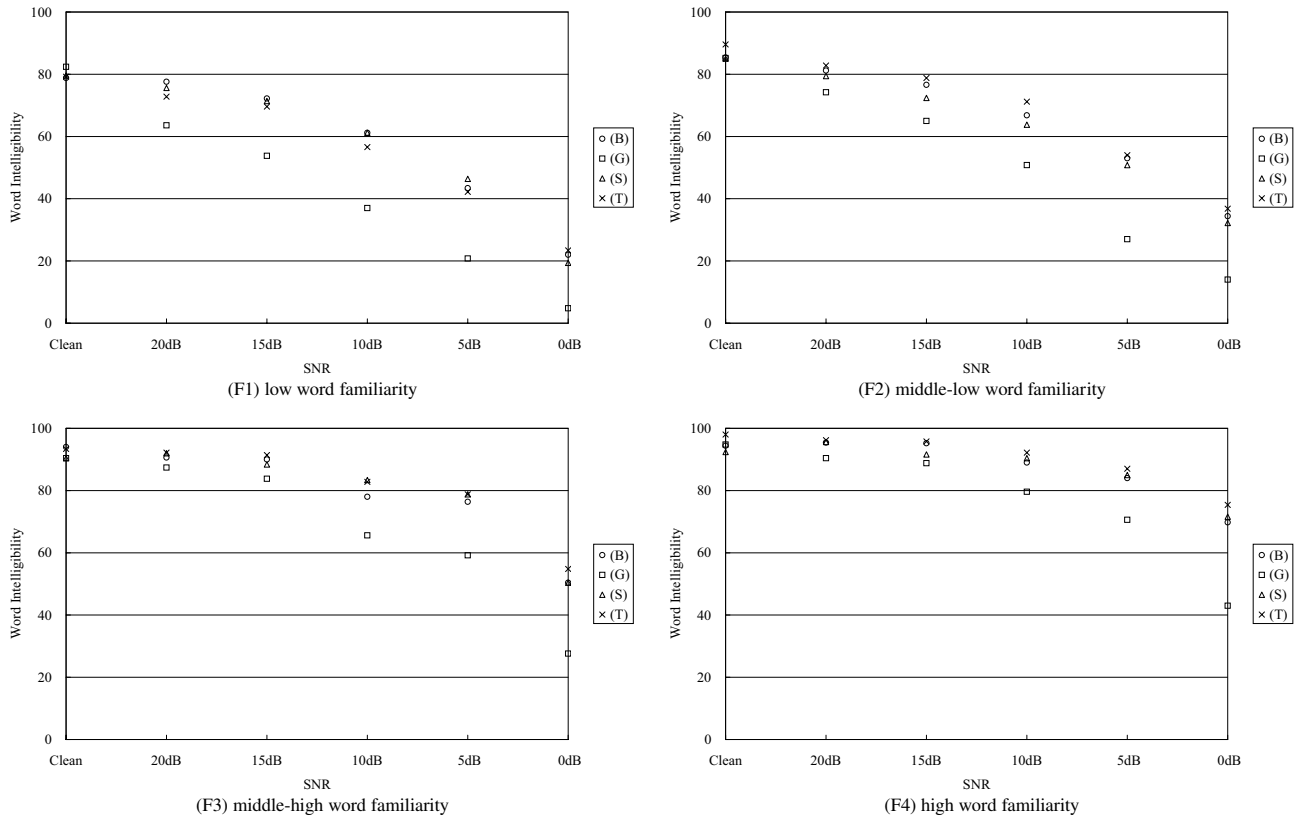

Figure 1: *The word intelligibility for each word familiarity rank in the case of the Car noise.*

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) $\times$ 500 (utterances) $\times$ 2 (noise types) $\times$ 6 (SNR values) $\times$ 4 (algorithms).

The word intelligibility test was performed in a soundproof 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), and most of them had not participated in such a test previously. The subjects were divided into two groups: one for the Subway noise and the other for the Car noise. The number of the speech samples for each subject is 4,800 (96 word lists), that is, 4 (familiarity ranks) $\times$ 50 (utterances) $\times$ 1 (noise type) $\times$ 6 (SNR values) $\times$ 4 (algorithms). In this test, each individual word list was used only once. The word intelligibility, which is defined by the ratio of the number of correct words to the total number of words, was calculated for each word list.

2.2. Results

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. It can be seen that the word familiarity strongly affects the word intelligibility. In particular, the degradation of the word intelligibility due to the noise increases as the word familiarity rank becomes low. We can also see that the word intelligibility for algorithm (T) is generally higher than that for (B), where (T) has causes little degradation of the speech component while the residual noise is relatively loud. On the other hand, (G) seriously degrades the word intelligibility in most cases.

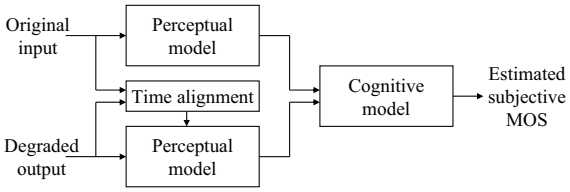

Figure 2: *The calculation process for the PESQ MOS.*

The reason is that (G) increases the speech distortion instead of considerably removing the noise component, especially under low SNR conditions.

3. Estimation of Word intelligibility

3.1. Overview of the proposed methodology

Figure 2 represents the calculation process for the PESQ MOS. First, the degraded sample and its original version are transformed to an internal representation based on perceptual frequency (Bark) and loudness (Sone) by using the perceptual model. Second, the cognitive model gives the estimated subjective MOS, which has a range of -0.5 to 4.5, by evaluating the difference between the degraded and the original samples.

In this paper, the word intelligibility is estimated by using an estimator expressed in the following form.

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

where y and x represent the estimated word intelligibility and the

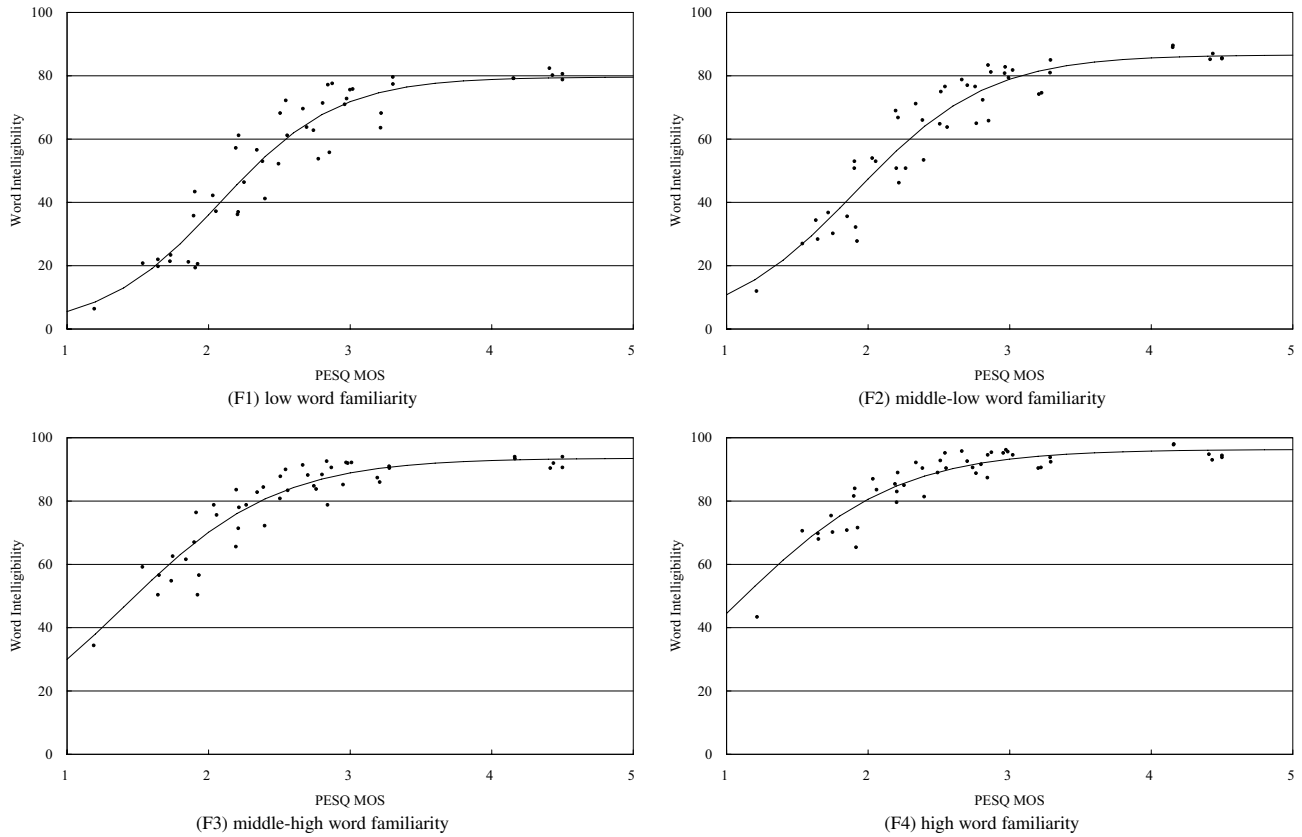

Figure 3: *The relationship between the word intelligibility and the PESQ MOS for each word familiarity rank.*

Table 2: *The constants used in the estimator 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

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. In this paper, the estimators used were optimized for each individual word familiarity rank without distinguishing the noise reduction algorithms and the noise types.

3.2. Effectiveness of the proposed methodology

Figure 3 shows the relationship between the word intelligibility and the PESQ MOS for each word familiarity rank. In this figure, each point represents the PESQ MOS and the word intelligibility obtained using one of the noise reduction algorithms for one of the noise types and a particular value of SNR. The solid line is the estimator mentioned above. The constants used in the estimator for each word familiarity rank are summarized in Table 2.

Figure 4 shows the relationship between the true word intelligibility and the estimated word intelligibility for each word familiarity rank. The coefficient of determination and the RMSE for each word familiarity rank are summarized in Table 3. From

Table 3: *The coefficient of determination and the 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

Figure 4 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 word intelligibility can be estimated well from the PESQ MOS without distinguishing the noise reduction algorithms and the noise types.

4. Conclusions

This paper has proposed a methodology which estimates the word intelligibility of the noise-reduced speech from the PESQ MOS. To evaluate the effectiveness of the proposed methodology, a word intelligibility test was performed on noise-reduced speech using four different noise reduction algorithms and word lists which take word difficulty into account. The word intelligibility was then estimated using the proposed methodology. The results confirmed that word intelligibility can be estimated well from the PESQ MOS without distinguishing the noise reduction algorithms and noise types.

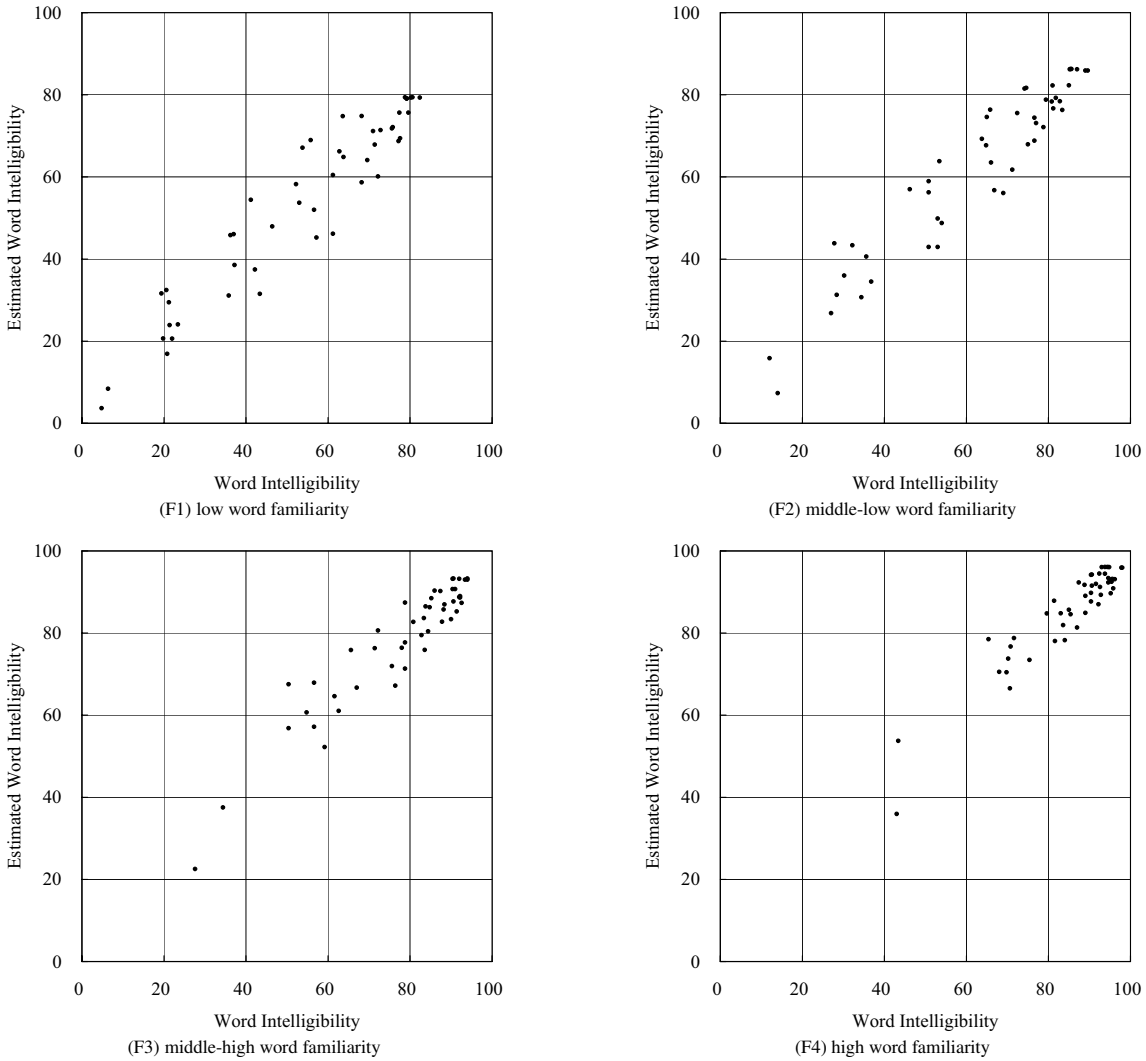

Figure 4: *The relationship between the true word intelligibility and the estimated word intelligibility for each word familiarity rank.*

5. 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 conducted using the AURORA-2J database developed by IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group. This work was supported in part by the Ministry of Public Management, Home Affairs, Posts and Telecommunications of Japan.

6. 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, 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] S. Sakamoto, Y. Suzuki, S. Amano, 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 2001.

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

[5] 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. European Conference on Speech Communication and Technology, EUROSPEECH2003, pp. 1781–1784, Sep. 2003.

[6] N. Kitaoka, 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, Sep. 2002.