Performance Estimation of Noisy Speech Recognition Using Spectral Distortion and SNR of Noise-reduced Speech

Guo Ling, Takeshi Yamada, Shoji Makino and Nobuhiko Kitawaki

Graduate School of Systems and Information Engineering, University of Tsukuba

1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan

e-mail: guoling@mmlab.cs.tsukuba.ac.jp

***Abstract*—To ensure a satisfactory QoE (Quality of Experience) and facilitate system design in speech recognition services, it is essential to establish a method that can be used to efficiently investigate recognition performance in different noise environments. Previously, we proposed a performance estimation method using the PESQ (Perceptual Evaluation of Speech Quality) as a spectral distortion measure. However, there is the problem that the relationship between the recognition performance and the distortion value differs depending on the noise reduction algorithm used. To solve this problem, we propose a novel performance estimation method that uses an estimator defined as a function of the distortion value and the SNR (Signal to Noise Ratio) of noise-reduced speech. The estimator is applicable to different noise reduction algorithms without any modification. We confirmed the effectiveness of the proposed method by experiments using the AURORA-2J connected digit recognition task and four different noise reduction algorithms.**

Keywords—performance estimation; noisy speech recognition; noise reduction; spectral distortion; SNR

I. INTRODUCTION

In recent years, speech recognition technology has been considerably improved by applying a statistical framework. However, current speech recognition systems still have the serious problem that their recognition performance is degraded in the presence of ambient noise. The degree of the performance degradation depends on the nature of ambient noise and the type of noise reduction used as a preprocessing stage in speech recognition. To ensure a satisfactory QoE (Quality of Experience) and facilitate system design in speech recognition services, it is essential to establish a method that can be used to efficiently investigate recognition performance in different noise environments.

One typical approach is to collect noisy speech data in a target noise environment and then perform a recognition experiment. However, this requires a skilled engineer and is labor and time-consuming. An alternative approach is to estimate recognition performance from a distortion value, which represents a spectral distortion between noisy (or noise-reduced) speech and its original clean version as shown in Fig. 1 [1-3].

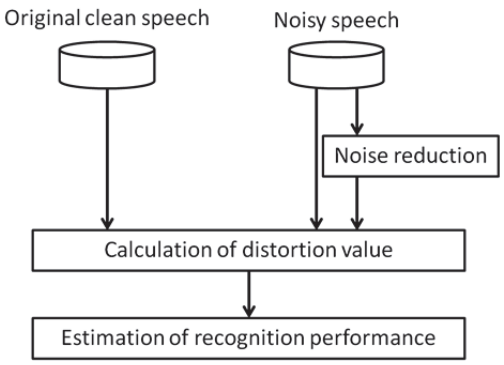


Fig. 1 Estimation of the recognition performance from the distortion value

Previously, we proposed a performance estimation method using the PESQ (Perceptual Evaluation of Speech Quality) [4] as a distortion measure [3]. In this method, an estimator, which is a function of the distortion value, is obtained by approximating the relationship between the recognition performance and the distortion value. Although this method gives accurate estimates of the recognition performance, there is the problem that the relationship between the recognition performance and the distortion value differs depending on the noise reduction algorithm used. This means that each individual noise reduction algorithm requires the special estimator.

To solve this problem, we propose a novel performance estimation method using the PESQ and the output SNR (Signal to Noise Ratio) that is the SNR of noise-reduced speech. In the proposed method, an estimator, which is a function of the distortion value and the output SNR, is introduced. The estimator is applicable to different noise reduction algorithms without any modification. We evaluate the effectiveness of the

proposed method by experiments using the AURORA-2J connected digit recognition task and four different noise reduction algorithms.

II. Proposed Method

The conventional method estimates the recognition performance by using the estimator expressed in the following form [3].

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

where y and x represent the estimated word accuracy and the PESQ score, respectively. The PESQ score has a range of 0.5 to 4.5. Note that the PESQ score of 4.5 means no spectral distortion. The constants a, b, and c correspond to the word accuracy for clean speech, the slope of the performance degradation, and the robustness against the spectral distortion, respectively. These constants are determined by approximating the relationship between the word accuracy and the PESQ score obtained in various noise environments.

Fig. 2 shows the relationship between the word accuracy and the PESQ score. In this figure, each point represents the PESQ score and the word accuracy obtained by one of five noise reduction algorithms in one of 28 noise conditions. The details of the experimental conditions are given in Section III. From Fig. 2, we can confirm that the relationship between the word accuracy and the PESQ score differs depending on the noise reduction algorithm used. This means that each individual noise reduction algorithm requires the special estimator.

To solve this problem, we focus on the output SNR that is equivalent to the SNR of noise-reduced speech. In this paper, the output SNR is approximately calculated by

$$\text{Output SNR} = 10\log_{10}\frac{\overline{|x(i)|^2}}{\overline{|n(i)|^2}}, \quad (2)$$

where *x(i)* and *n(i)* are the speech segments and noise segments, respectively. Fig. 3 represents the relationship between the word accuracy, the PESQ score, and the output SNR. This figure is the same as Fig. 2 except that the type of marker represents not the algorithm but the range of the output SNR. From Fig. 3, we can see that the word accuracy tends to be higher as the output SNR becomes large, when comparing with two points with the same PESQ score. This implies that the word accuracy can be estimated by using both the PESQ score and the output SNR.

Based on this finding, we propose the estimator expressed in the following form.

$$y = \frac{a}{1 + e^{-b_1 x_1 - b_2 x_2 + c}}, \quad (3)$$

where y, x_1, and x_2 indicate the estimated word accuracy, the PESQ score, and the output SNR, respectively. The constants a, b_1, b_2, and c are determined by approximating the relationship between the word accuracy, the PESQ score, and the output SNR obtained using different noise reduction algorithms in various noise environments.

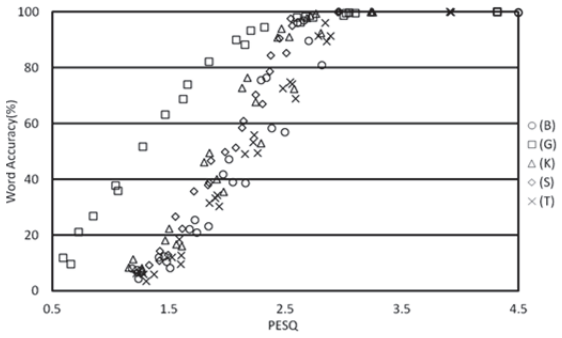


Fig. 2 Relationship between the word accuracy and the PESQ score

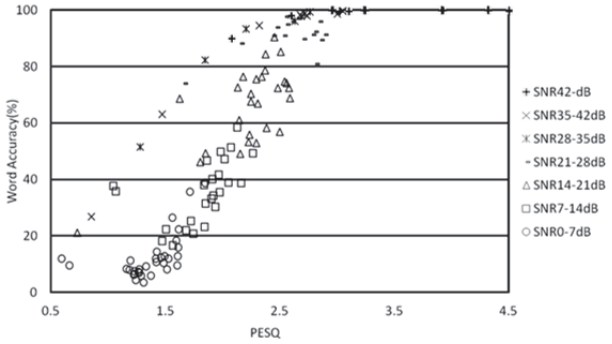


Fig. 3 Relationship between the word accuracy, the PESQ score, and the output SNR

III. Verification of Proposed Method

A. Experimental conditions

We use four noise reduction algorithms, in addition to the reference case of no such algorithm.

(B) Baseline (noise reduction is NOT implemented)

(S) Spectral subtraction with smoothing of time direction [5]

(T) Temporal domain SVD-based speech enhancement [6]

(G) GMM-based speech estimation [6]

(K) KLT-based comb-filtering [7]

In these algorithms, the speech component in the noisy speech frames is extracted by using the noise component estimated from the non-speech frames. The algorithm (S) is one of the most popular algorithms. The algorithm (G), and algorithms (T) and (K) are based on the Wiener filter and the subspace method, respectively. The characteristics of a speech distortion and residual noise in noise-reduced speech vary depending on the noise-reduction algorithm used.

Recognition experiments were performed on the AURORA-2J connected digit recognition task [8]. The training and test sets for the AURORA-2J task are shown in Table I.

TABLE I. EXPERIMENTAL CONDITIONS

Training and Test sets	Speech	Noise	Channel	SNR(dB)
Clean training	8,440 utterances of 110 people	None	G.712	Clean
Test set A	4,004 utterances of 104 people	Subway, Babble, Car, Exhibition		Clean, 20, 15, 10, 5, 0, -5
Test set B		Restaurant, Street, Airport, Station		

In this paper, the clean training set is used for training HMMs (Hidden Markov Models) and the test sets A and B for evaluation. The test sets A and B are based on the same 4,004 clean speech signals. For each set these signals are divided into four groups of 1,001 speech signals, corresponding to four different types of noise. Then, for each group, the appropriate noise signal is artificially added to the speech signals at different seven values of SNR. As a result, each test set has 28 noise conditions.

The speech signal sampled at 8 kHz is windowed by a 25ms Hamming window every 10ms. The feature vector has 39 components consisting of 12 MFCCs (Mel-Frequency Cepstral Coefficients) together with log power, and their first and second derivatives. The digit models and the silence model are the HMMs with 16 states and 3 states, respectively. The number of Gaussians per state is 20 for the digit models and 36 for the silence model. The HMM of the short pause is the 1 state tied with the 2nd state of the silence model. The set of HMMs is trained for each individual noise reduction algorithm, where all the training data are processed by the algorithm before training.

The conditions for determination of the estimator's constants are described below.

(C1) The estimators of the conventional method and the proposed method are optimized for the test set A with all the noise reduction algorithms. They are evaluated using the test set A (Noise-closed test).

(C2) The estimators are the same as those in the condition (C1), but they are evaluated using the test set B (Noise-open test).

The determined estimators of the conventional method and the proposed method were

$$y = \frac{102.7779}{1 + e^{-2.807325x + 5.618707}} \tag{4}$$

and

$$y = \frac{100.8289}{1 + e^{-0.088535x_1 - 1.738907 \times x_2 + 4.737119}}, \tag{5}$$

respectively.

B. Results

Fig. 4 shows the relationship between the word accuracy and the overall distortion calculated by $b_1x_1+b_2x_2$ in the proposed method. We can see the consistent relationship between the word accuracy and the overall distortion for all the noise reduction algorithms used.

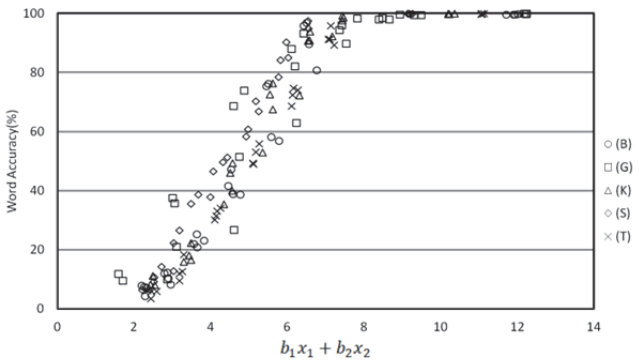


Fig. 4 Relationship between the word accuracy and the overall distortion calculated by $b_1x_1+b_2x_2$ in the proposed method

We then estimated the recognition performance by using the estimators expressed by Eqs. (4) and (5). Figs. 5 and 6 show the relationship between the true word accuracy and the estimated word accuracy in the conditions (C1) and (C2), respectively. It can be seen that the proposed method gives more accurate estimates than the conventional method. The coefficient of determination, R^2, and the RMSE (Root Mean Square Error), which are given by

$$R^2 = 1 - \frac{\sum(True\ \%Acc - Estimated\ \%Acc)^2}{\sum(True\ \%Acc - \overline{True\ \%Acc})^2} \tag{6}$$

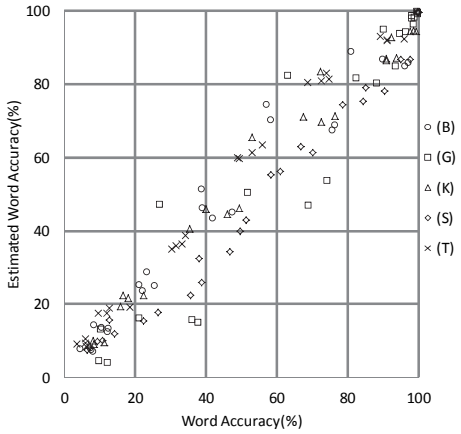
and

$$RMSE = \sqrt{\frac{\sum(True\ \%Acc - Estimated\ \%Acc)^2}{N}} \tag{7}$$

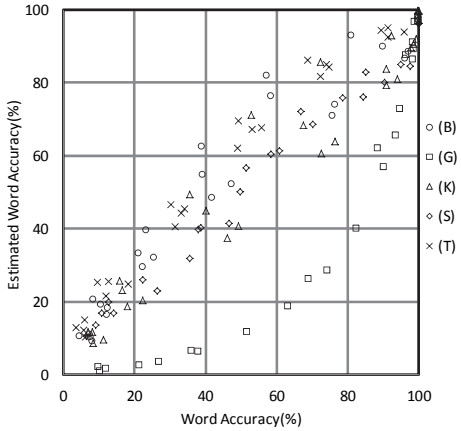
are shown in Table II. It can be confirmed that the RMSE of the proposed method decreased almost by half to 7.0 from 13.2 in the condition (C1) and to 7.37 from 14.0 in the condition (C2). The difference in noise type has little effect on the estimation performance. These results confirmed that the proposed method give accurate estimates of the recognition performance without preparing the special estimator for each individual noise reduction algorithm.

IV. CONCLUSIONS

Previously, we proposed a performance estimation method using a spectral distortion measure. However, there is the problem that the relationship between the recognition performance and the distortion value differs depending on the noise reduction algorithm used. To solve this problem, we proposed a novel performance estimation method that uses an estimator defined as a function of the distortion value and the SNR (Signal to Noise Ratio) of noise-reduced speech. The experimental results confirmed that the proposed method give accurate estimates of the recognition performance without preparing the special estimator for each individual noise reduction algorithm.

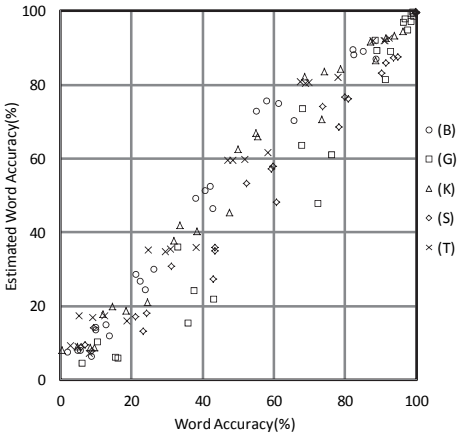


(a) Proposed method

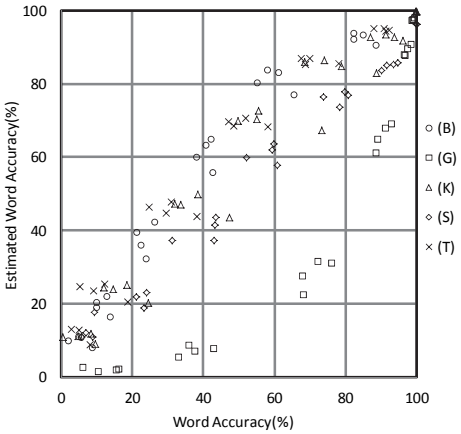


(b) Conventional method

Fig. 5 Relationship between the true word accuracy and the estimated word accuracy in the condition (C1)



(a) Proposed method



(b) Conventional method

Fig. 6 Relationship between the true word accuracy and the estimated word accuracy in the condition (C2)

REFERENCES

[1] H. Sun, L. Shue, J. Chen, "Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP2004, Vol. 1, pp. 865-868, May 2004.

[2] M. Kondo, K. Takeda, F. Itakura, "Predicting the degradation of speech recognition performance from sub-band dynamic ranges," IPSJ Journal, Vol. 43, No. 7, pp. 2242-2248, July 2002.

[3] T. Yamada, M. Kumakura, N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 6, pp. 2006–2013, Nov. 2006.

[4] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.

[5] 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, 2002.

TABLE II. R^2 AND RMSE

	(C1)		(C2)	
	R^2	RMSE	R^2	RMSE
Conventional Method	0.86	13.2	0.85	14.0
Proposed Method	0.96	7.0	0.96	7.37

[6] 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, 2003.

[7] S.-J. Park, M. Ikeda, K. Takeda, 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, 2002.

[8] 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.