Performance Estimation of Noisy Speech Recognition Considering Recognition Task Complexity

[1]*Takeshi Yamada,* [1]*Tomohiro Nakajima,* [1]*Nobuhiko Kitawaki,* [1,2]*Shoji Makino*

[1]Graduate School of Systems and Information Engineering, University of Tsukuba, Japan
[2]Center for Tsukuba Advanced Research Alliance, University of Tsukuba, Japan
takeshi@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 a spectral distortion measure. However, there is the problem that recognition task complexity affects the relationship between the recognition performance and the distortion value. To solve this problem, this paper proposes a novel performance estimation method considering the recognition task complexity. We confirmed that the proposed method gives accurate estimates of the recognition performance for various recognition tasks by an experiment using noisy speech data recorded in a real room.

Index Terms: performance estimation, noisy speech recognition, recognition task difficulty

1. 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. 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 based on a distortion value, which represents a spectral distortion between noisy speech and its original clean version [1, 2, 3].

Previously, we proposed a performance estimation method using the PESQ (Perceptual Evaluation of Speech Quality) [4] as a distortion measure. 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 [3]. It is, however, well-known that recognition performance varies according to recognition task complexity. This means that each individual recognition task requires the special estimator.

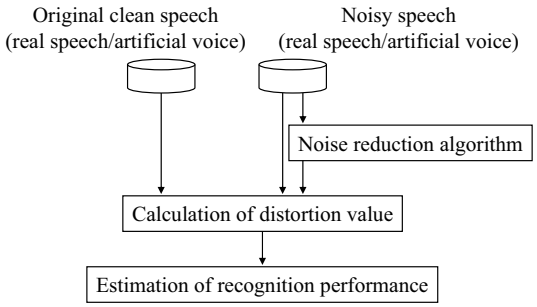

Figure 1: Overview of the recognition performance estimation.

To solve this problem, we propose a novel performance estimation method considering the recognition task complexity. In the proposed method, an estimator, which is the function of both the distortion value and the recognition task complexity, is introduced. It can estimate the recognition performance by giving both the distortion value and the task complexity, and can provide the special estimator for each individual recognition task by giving only the task complexity. We evaluate the effectiveness of the proposed method by an experiment using noisy speech data recorded in a real room.

2. Proposed method

Fig. 1 illustrates the overview of the recognition performance estimation. First the distortion value that represents the spectral distortion between the noisy speech and its original clean version is calculated. Then the recognition performance is estimated by using the estimator ex-

pressed in the following form [3].

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

where y and x represent the estimated recognition performance and the distortion value, respectively. The constants a, b, and c correspond to the recognition performance 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 recognition performance and the distortion value for various noise environments. However, as mentioned above, the recognition performance varies according to recognition task complexity. Each individual recognition task therefore requires the special estimator.

To solve this problem, we propose the estimator expressed in the following form.

$$y = f(x, \alpha) = \frac{a(\alpha)}{1 + e^{-b(\alpha)(x-c(\alpha))}}, \tag{2}$$

where α is the recognition task complexity. In Eq. (2), each constant in Eq. (1) is replaced by the function of α. This is motivated by the hypothesis that the recognition task complexity affects only the constants in Eq. (1). The proposed estimator can estimate the recognition performance by giving both the distortion value and the task complexity, and can provide the special estimator for each individual recognition task by giving only the task complexity.

In this paper, we adopt the SMR-Perplexity (Square Mean Root- Perplexity) [5] as a measure of the recognition task complexity. The SMR-Perplexity is expressed in the following form.

$$P_{\mathrm{SMR}} = \left\{ \frac{1}{n+1} \left(\sqrt{\frac{1}{P(w_1|\cdot)}} + \sqrt{\frac{1}{P(w_2|w_1)}} + \cdots + \sqrt{\frac{1}{P(\cdot|w_1 \cdots w_n)}} \right) \right\}^2, \tag{3}$$

where $P(\cdot|\cdot)$ is the word occurrence probability. Fig. 2 shows the relationship between the constant a in Eq. (1) and the SMR-Perplexity. Each point represents the SMR-Perplexity calculated for one of the recognition tasks and the constant a in the special estimator for the recognition task. The details of the recognition tasks are described in Section 3. It can be seen that the constant a can be represented by an exponential function of the SMR-Perplexity. The similar tendency was observed for the constants b and c. We therefore decided to use the following estimator.

$$y = f(x, \alpha) = \frac{p_1 \alpha^{q_1} + r_1}{1 + e^{-(p_2 \alpha^{q_2} + r_2)(x - (p_3 \alpha^{q_3} + r_3))}}. \tag{4}$$

The constants $p.$, $q.$, and $r.$ are determined by approximating the relationship between the recognition performance, the distortion value, and the SMR-Perplexity for various noise environments and recognition tasks.

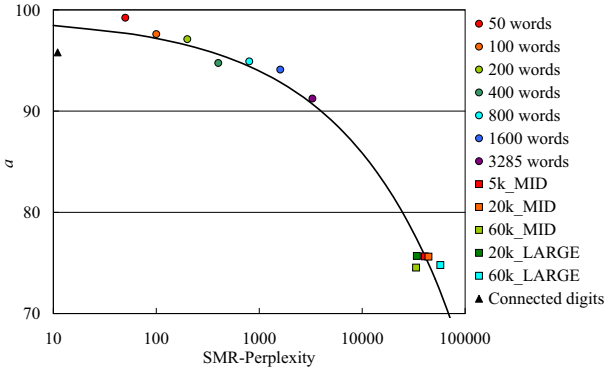

Figure 2: Relationship between the constant a in Eq. (1) and the SMR-Perplexity.

3. Evaluation

In this section, we evaluate the effectiveness of the proposed method. The noisy speech data generated by artificially adding noise data to speech data are used for determining the constants of the estimator. In contrast, the evaluation of the proposed method is done on the noisy speech data recorded in a real room. This is a realistic scenario in the sense that the recording cost can be reduced in determining the constants of the estimator.

In this experiment, we use the PESQ as a spectral distortion measure. The PESQ calculates the spectral distortion and outputs the value as the PESQ score ranging from -0.5 to 4.5. Note that the higher the PESQ score, the smaller the spectral distortion.

3.1. Determination of the estimator's constants

To determine the constants of the proposed estimator, we prepared the following recognition tasks and the clean speech data corresponding to each task.

- Grammar-based recognition: The speech data used are connected-digit utterances, which are the same as those in the AURORA-2J database [6], except that the sampling rate is 16 kHz. The grammar allows arbitrary repetitions of digits, a short pause, and a terminal silence.

- Isolated-word recognition: We used the Tohoku University-Matsushita spoken word database [7], consisting of 3,285 isolated words (railway station names). The dictionary size is set to 50, 100, 200, 400, 800, 1600, and 3285.

- LVCSR (Large Vocabulary Continuous Speech Recognition): We used two sets of sentence utterances by male speakers, in which the vocabulary size is set to 5k (MID) and 20k (LARGE), included in the ASJ-JNAS (Japanese Newspaper Article Sentences) database [8]. The language mod-

Table 1: SMR-Perplexity for each recognition task.

Recognition task		SMR-Perplexity
Grammar-based	Connected-digit	11
Isolated-word	50 words	50
	100 words	100
	200 words	200
	400 words	400
	800 words	800
	1,600 words	1,600
	3,285 words	3,285
LVCSR	5k_MID	40,588
	20k_MID	44,073
	60k_MID	33,381
	20k_LARGE	33,976
	60k_LARGE	57,424

els are word 3-gram models with 5k, 20k, and 60k words [9].

The SMR-Perplexity for each recognition task is summarized in Table 1.

We prepared the in-car noise, the exhibition hall noise, the train noise, and the elevator hall noise included in the Denshikyo noise database [10] as ambient noise. The noisy speech data were generated by artificially adding the noise data to the speech data at six different values of SNR (20, 15, 10, 5, 0,−5 dB). In this experiment, no noise reduction algorithm is used. The acoustic models are gender independent monophone models with 16 Gaussians per state [9]. The feature vector has 25 components consisting of 12 MFCCs, 12 delta MFCCs, and a delta log-power.

Using the word accuracy, the PESQ score, and the SMR-Perplexity obtained for each of the recognition tasks mentioned above, we determined the constants of the proposed estimator.

$$y = f(x,\alpha) = \frac{a(\alpha)}{1+e^{-b(\alpha)(x-c(\alpha))}}, \qquad (5)$$

where
$$a(\alpha) = -0.335(\alpha)^{0.401} + 99.31,$$
$$b(\alpha) = -1.66\cdot 10^{-8}\,(\alpha)^{1.67} + 4.44,$$
$$c(\alpha) = -13.42\,(\alpha)^{-0.00518} + 15.15,$$

where α is the SMR-Perplexity. Fig. 3 illustrates the special estimator for each individual recognition task obtained by substituting only the corresponding SMR-Perplexity in Eq. (5).

We then estimated the recognition performance by substituting both the PESQ score and the SMR-Perplexity in Eq. (5). This corresponds to a so-called closed test in the sense that the test data are the same as those used for determining Eq. (5). Fig. 4 shows the relationship between the true word accuracy and the estimated word accuracy. The coefficient of determination and the RMSE (Root Mean Square Error) were 0.99 and 3.4, respectively. We can see that the proposed estimator gives accurate estimates of the word accuracy.

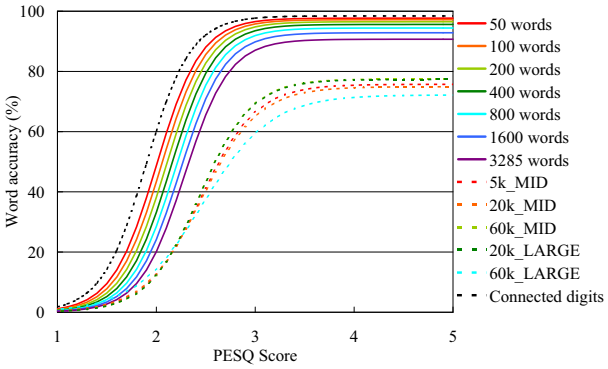

Figure 3: Special estimator for each individual recognition task obtained by substituting only the corresponding SMR-Perplexity in Eq. (5).

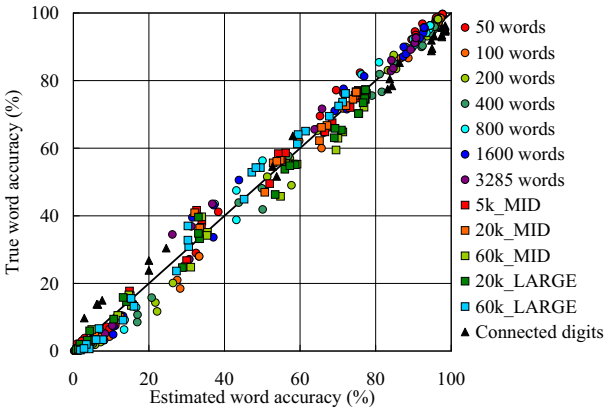

Figure 4: Relationship between the true word accuracy and the estimated word accuracy.

3.2. Evaluation using real data

We recorded noisy speech data in a real room. Fig. 5 outlines the recording setup. The room size is 7.2m × 7.7m. The room is surrounded by concrete walls. The reverberant time, T_{60}, is about 0.7 second. We used two loudspeakers instead of a human speaker and a noise source. The volume of the loudspeaker for noise is set to 3 steps. The close microphone and the remote microphone are placed at 10cm and 100cm from the loudspeaker for speech, respectively. The noise data are the exhibition hall noise and the factory noise included in the Denshikyo noise database [10]. The speech data are the same as those described in Section 3.1, expect that '800 words' and '1,600 words' are selected in the isolated-

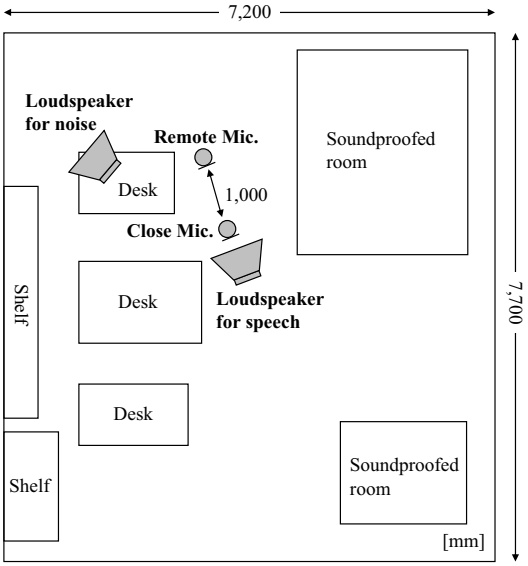

Figure 5: The recording setup.

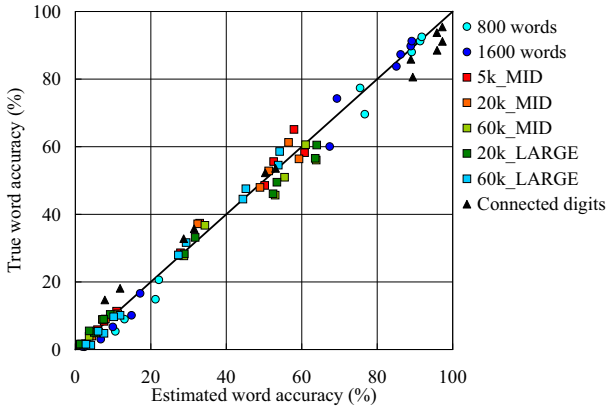

Figure 6: Relationship between the true word accuracy and the estimated word accuracy in the case of the real data.

word recognition task.

We estimated the recognition performance by substituting both the PESQ score calculated from the real data and the SMR-Perplexity in Eq. (5). Fig. 6 shows the relationship between the true word accuracy and the estimated word accuracy for both the close microphone and the remote microphone. The coefficient of determination and the RMSE are 0.98 and 3.2, respectively. We confirmed that the proposed estimator gives accurate estimates of the word accuracy. This result implies that we can use the artificially generated noisy speech data for determining the constants of the estimator.

4. Conclusions

Previously, we proposed a performance estimation method using a spectral distortion measure. However, there is the problem that recognition task complexity affects the relationship between the recognition performance and the distortion value. To solve this problem, this paper proposed a novel performance estimation method considering the recognition task complexity. We confirmed that the proposed method gives accurate estimates of the recognition performance for various recognition tasks by an experiment using noisy speech data recorded in a real room. As future work, we plan to estimate the recognition performance of unknown recognition tasks.

5. Acknowledgements

This work was supported in part by the Telecommunications Advancement Foundation.

6. References

[1] 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.

[2] 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. I, pp. 865–868, May 2004.

[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 narrow-band telephone networks and speech codecs,” Feb. 2001.

[5] S. Nakagawa, M. Ida, “A new measure of task complexity for continuous speech recognition,” The Transactions of the IEICE, Vol. J81-D-2, No. 7, pp. 1491–1500, July 1998. (in Japanese)

[6] 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.

[7] S. Makino, N. Niyada, Y. Mafune, K. Kido, “Tohoku University and Matsushita isolated spoken word database,” Journal of the Acoustical Society of Japan, Vol. 48, No. 12, pp. 899–905, 1992.

[8] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” The Journal of the Acoustical Society of Japan, Vol. 20, No. 3, pp. 199–206, May 1999.

[9] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, K. Shikano, “Free software toolkit for Japanese large vocabulary continuous speech recognition,” Proc. International Conference on Spoken Language Processing, ICSLP2000, pp. 476–479, Oct. 2000.

[10] Denshikyo noise database, http://research.nii.ac.jp/src/list/detail.html#JEIDA-NOISE.