Performance Estimation of Noisy Speech Recognition Considering the Accuracy of Acoustic Models

Takashi Takaoka, Takeshi Yamada, Shoji Makino and Nobuhiko Kitawaki
Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8573 Japan
takaoka@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 measure of speech distortion. However, there is the problem that the accuracy of acoustic models used for speech recognition affects the relationship between the recognition performance and the distortion value. To solve this problem, we propose a novel performance estimation method considering the accuracy of acoustic models. Experimental results confirmed that the proposed method gives accurate estimates of the recognition performance for different sets of acoustic models, when using the recognition error rate for clean speech as a measure of the accuracy of acoustic models.

Keywords: performance estimation, noisy speech recognition, acoustic model

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 using the data. However, this requires a skilled engineer and a lot of time. An alternative approach is to estimate the recognition performance based on a distortion value, which represents the spectral distortion between the 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 measure of speech distortion. In this method, an estimator, which is the function of the distortion value, is preliminarily obtained by approximating the relationship between the recognition performance and the distortion value. The recognition performance is then estimated by substituting the distortion value in the estimator. However, there is the problem that the accuracy of acoustic models used for speech recognition affects the relationship between the recognition performance and the distortion value. This means that each individual set of acoustic models requires the special estimator. It is, however, labor-intensive and time-consuming.

To solve this problem, we propose a novel performance estimation method considering the accuracy of acoustic models. In the proposed method, an estimator, which is the function of both the distortion value and the accuracy of acoustic models, is introduced. It can estimate the recognition performance by giving both the distortion value and the accuracy of acoustic models, and can provide the special estimator for each individual set of acoustic models by giving only the accuracy of acoustic models. We evaluate the effectiveness of the proposed method by an experiment using different sets of acoustic models.

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 expressed in the following form [3].

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

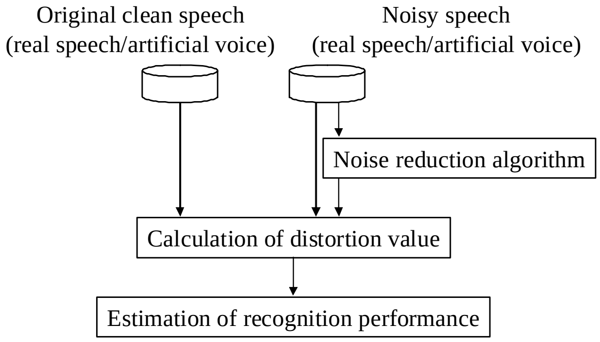
Original clean speech
(real speech/artificial voice)
Noisy speech
(real speech/artificial voice)
Noise reduction algorithm
Calculation of distortion value
Estimation of recognition performance

Figure 1: Overview of the recognition performance estimation.

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

In this paper, we adopt the recognition error rate for clean speech as a measure of the accuracy of acoustic models. Fig. 2 shows the relationship between the constant a in the special estimator and the recognition error rate for clean speech. Each point corresponds to one of the five sets of acoustic models.

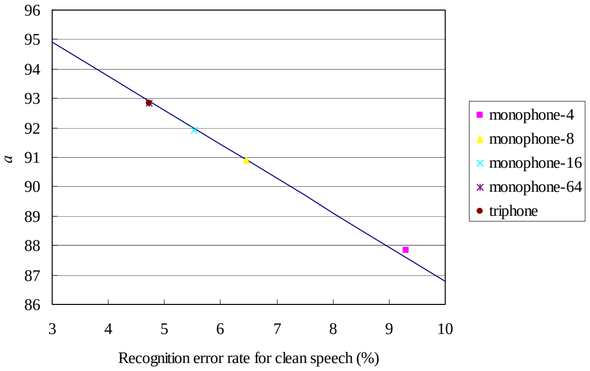

Figure 2: Relationship between the constant a in the special estimator and the recognition error rate for clean speech.

It can be seen that the constant a can be represented by the linear function of the recognition error rate for clean speech. The similar tendency was observed for the constants b and c. We therefore decided to use the following estimator.

$$y = f(x, \alpha) = \frac{p_1\alpha + q_1}{1 + e^{(-p_2\alpha - q_2)(x - p_3\alpha - q_3)}}, \tag{3}$$

where the constants $p_\cdot$ and $q_\cdot$ are determined by approximating the relationship between the recognition performance, the distortion value, and the recognition error rate for clean speech for various noise environments and sets of acoustic models.

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 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 conducted an isolated-word recognition experiment. We used the Tohoku University-Matsushita spoken word database [5], consisting of 3285 isolated words (railway station names). The dictionary size is 3285.

We prepared in-car noise, exhibition hall noise, train noise, and elevator hall noise included in the Denshikyo

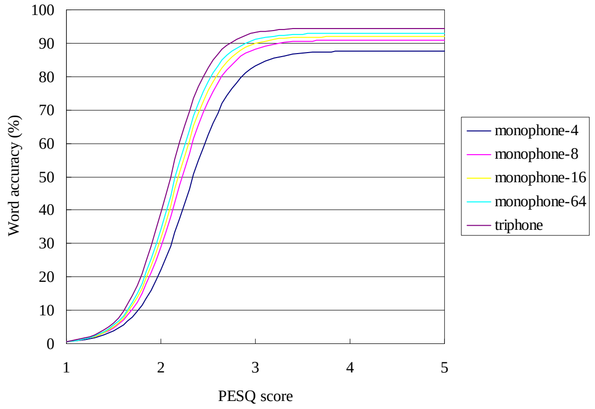

Figure 3: Special estimator for each individual set of acoustic models obtained by substituting only the corresponding the recognition error rate for clean speech in Eq. (3).

noise database [6] 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. We used five sets of acoustic models: four sets of monophone models with 4, 8, 16, and 64 Gaussians per state, respectively, and one set of triphone models [7]. The feature vector has 25 components consisting of 12 MFCCs, 12 delta MFCCs, and a delta log-power.

Using the word accuracy (the recognition rate), the PESQ score, and the recognition error rate for clean speech obtained for each individual set of acoustic models mentioned above, we determined the constants of the proposed estimator.

$$y = f(x, \alpha) = \frac{a(\alpha)}{1 + e^{-b(\alpha)(x - c(\alpha))}} \qquad (4)$$

$$a(\alpha) = -1.16(\alpha) + 98.41$$

$$b(\alpha) = -0.094(\alpha) + 4.89$$

$$c(\alpha) = 0.033(\alpha) + 1.96$$

In Eq. (4), α is the recognition error rate for clean speech. Fig. 3 illustrates the special estimator for each individual set of acoustic models obtained by substituting only the corresponding recognition error rate for clean speech in Eq. (4).

3.2. Evaluation

We then estimated the recognition performance by substituting both the PESQ score and the recognition error rate for clean speech in Eq. (4). This corresponds to a so-called closed test in the sense that the test data are the same as those used for determining Eq. (4). 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 2.8, respectively. We can see that the proposed estimator gives accurate estimates of the word accuracy.

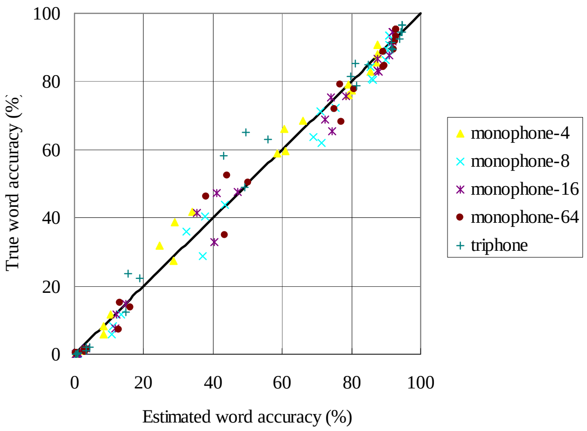

Figure 4: Relationship between the true word accuracy and the estimated word accuracy in the closed test.

An additional evaluation was conducted for an open test (a cross-validation test). The four sets of acoustic models are used for determining the estimator's constants and the remaining one set for testing. We estimated the recognition performance of each individual set of acoustic models in this manner. Fig. 5 shows the relationship between the true word accuracy and the estimated word accuracy. The coefficient of determination and the RMSE were 0.98 and 3.1, respectively. We can see again that the proposed estimator gives accurate estimates of the word accuracy.

4. Conclusion

Previously, we proposed a performance estimation method using a spectral distortion measure. However, there is the problem that the accuracy of acoustic models 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 accuracy of acoustic models. We conducted an experiment to evaluate the effectiveness of the proposed method. As a result, it was

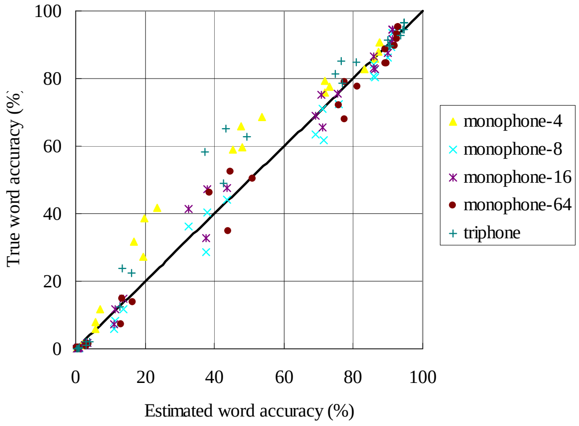

Figure 5: Relationship between the true word accuracy and the estimated word accuracy in the open test.

confirmed that the proposed method gives accurate estimates of the recognition performance for different sets of acoustic models, when using the recognition error rate for clean speech as a measure of the accuracy of acoustic models. As future work, we plan to estimate the recognition performance considering both the accuracy of acoustic models and recognition task complexity [8].

References

[1] M. Kondo, K. Takeda, and 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, and 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, and 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. 6pp. 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. Makino, N. Niyada, Y. Mafune, and 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..

[6] Denshikyo noise database, http://research.nii.ac.jp/src/list/detail

[7] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu,S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and 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.

[8] T. Yamada, T. Nakajima, N. Kitawaki, and S. Makino, “Performance estimation of noisy speech recognition considering recognition task complexity,” Proc. Interspeech 2010, pp. 2042-2045, Sep. 2010.