Performance Estimation of Noisy Speech Recognition Based on Short-Term Noise Characteristics

Eri Morishita, 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
morishita@mmlab.cs.tsukuba.ac.jp

Abstract: General users, who are unfamiliar with speech recognition, can't judge whether speech recognition works well or not in different noise environments. It is therefore important to constantly show a user the recognition performance expected at the next moment by an intuitive expression. This enables a user to select the other input device without wasting time. However, this issue has not been considered in the research area of noisy speech recognition. To improve the usability of speech recognition in noise environments, this paper proposes a method for estimating the recognition performance expected at the next moment from the characteristics of the noise observed previously. We conducted an experiment to evaluate the effectiveness of the proposed method. As a result, it was confirmed that the proposed method gives relatively accurate estimates of the recognition performance.

Keywords: performance estimation, noisy speech recognition, short-term noise characteristics

1. Introduction

In recent years, the technology of speech recognition has considerably been improved; however, it has not come into wide use. One of the reasons is that the recognition performance is degraded in the presence of ambient noise [1]. Here, the critical issue is that general users, who are unfamiliar with speech recognition, can't judge whether speech recognition works well or not in different noise environments. As a result, the users are annoyed by the recognition errors that occur frequently. It is therefore important to constantly show a user the recognition performance expected at the next moment by an intuitive expression, just like the antenna icon that most cell phones have on the display. This enables a user to select the other input device without wasting time. However, this issue has not been considered in the research area of noisy speech recognition.

To improve the usability of speech recognition in noise environments, this paper proposes a method for estimating the recognition performance expected at the next moment from the characteristics of the noise observed previously. It is well-known that ambient noise is generally non-stationary, time-variant, and unpredictable. This means that the long-term characteristics of the noise are unsuitable for estimating the future performance [2]. The proposed method therefore adopts the short-term characteristics of the noise, under the assumption that ambient noise is stationary and time-invariant in a short-term period of several-hundred msec, and estimates the performance of phoneme recognition. We evaluate the effectiveness of the proposed method by an experiment using different noise signals.

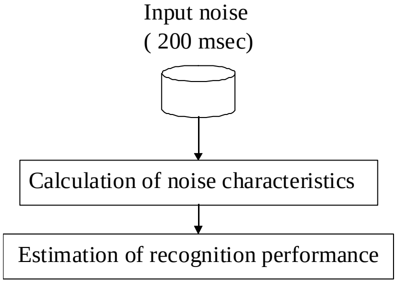
Input noise
(200 msec)
Calculation of noise characteristics
Estimation of recognition performance

Figure 1: Overview of the proposed method.

2. Proposed method

Fig. 1 illustrates the overview of the proposed method. The proposed method constantly observes the ambient noise and shows a user the estimate of the recognition performance expected at the next moment.

2.1. Calculation of the noise characteristics

The short-term characteristics of the input noise are calculated from the segment of 200 msec. We investigate the following two types of characteristics.

- **Log power**

  The log power of the noise segment is calculated by the following equation.

$$P = \log \sum_{i=0}^{N-1} n(i)^2 \; , \qquad (1)$$

where *n(i)* and *N* are the noise signal and the length of the segment, respectively. This is based on the fact that the recognition performance is degraded as the noise power becomes larger.

- **Mel-scaled log power spectrum**

  The mel-scaled log power spectrum of the noise segment is obtained by a simple Fourier transform based 24-dimensional filterbank on a mel-scale. The mel-scale is defined by

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) , \quad (2)$$

  where f is the frequency [3]. This is motivated by the hypothesis that the shape of the noise spectrum affects the degree of the recognition performance degradation.

2.2. Estimation of the recognition performance

Previously, we developed a performance estimation method using the speech distortion measure, PESQ (Perceptual Evaluation of Speech Quality) [2][4]. The proposed method is similar to this method. The recognition performance is estimated by using the estimator expressed in the following form.

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

where y, x_i, and n represent the estimated recognition performance, the *i*-th component of the noise characteristics, and the number of the components of the noise characteristics, respectively. The constants a, b_i, and c correspond to the recognition performance for clean speech, the slope of the performance degradation for each component, and the robustness against the noise, respectively. These constants are determined by approximating the relationship between the recognition performance and the noise characteristics for various noise environments. The proposed method gives the estimate of the performance of phoneme recognition for a speech segment of 200 msec, since the length of the noise segment is short.

3. Evaluation

In this section, we evaluate the effectiveness of the proposed method by an experiment using different noise signals.

Table 1: Experimental conditions.

Speech data	100 sentence utterances
Noise data	in-car, exhibition hall, train, elevator hall
SNR	20, 18, ···, 0 dB
Frame length	25 msec
Frame period	10 msec
Feature vector	12 MFCCs
HMM	3states, 16 Gaussians per state

3.1. Experimental conditions

The experimental conditions are briefly summarized in Table 1. As speech data, 100 sentence utterances included in the ASJ-JNAS (Japanese Newspaper Article Sentences) database [5] are used. The first speech segment of 200 msec is cut out from each utterance. As noise data, in-car noise, exhibition hall noise, train noise, and elevator hall noise, which are included in the Denshikyo noise database [6], are used. The eight segments of 200 msec are randomly cut out from each noise data. The noisy speech data are generated by artificially adding the noise data to the speech data at eleven different values of SNR (20, 18, ···, 0 dB). As a result, the total number of the noisy speech data is 35,200, that is, 100 (utterances) x 4 (noise types) × 8 (noise segments) × 11 (SNRs).

The noisy speech data sampled at 16 kHz is windowed by a 25 msec Hamming window every 10 msec. The feature vector has 12 components consisting of MFCCs (Mel-Frequency Cepstral Coefficients). The acoustic models used for phoneme recognition are 3-state HMMs with 16 Gaussians per state. The set of HMMs is trained with the ASJ database and the ASJ-JNAS database. The recognition grammar generates arbitrary repetitions of Japanese syllables [7]. Note that a single phoneme can appear at the end of the speech segment.

3.2. Results

We determined the constants of the proposed estimator, Eq. (3), for each of the two types of noise characteristics mentioned above, and then estimated the recognition performance by using each estimator.

Fig. 2 shows the relationship between the true phoneme accuracy and the estimated phoneme accuracy when the log power was used as noise characteristics. In this figure, each point corresponds to one of the 352 noise conditions. The

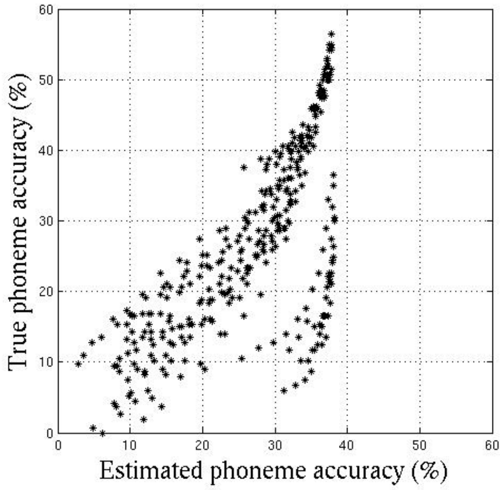

Figure 2: Relationship between the true phoneme accuracy and the estimated phoneme accuracy (log power).

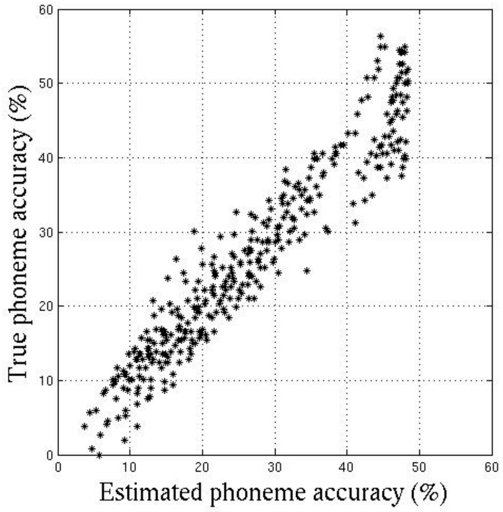

Figure 3: Relationship between the true phoneme accuracy and the estimated phoneme accuracy (mel-scaled log power spectrum).

coefficient of determination, R^2, and the RMSE (Root Mean Square Error), which are defined by

$$R^2 = 1 - \frac{\sum(True\ Accuracy - Estimated\ Accuracy)^2}{\sum\left(True\ Accuracy - \overline{True\ Accuracy}\right)^2} \qquad (4)$$

and

$$RMSE = \sqrt{\frac{\sum(True\ Accuracy - Estimated\ Accuracy)^2}{N}}\ , \quad (5)$$

were 0.49 and 9.66, respectively. We can see that the log power gives poor estimates of the phoneme accuracy.

Fig. 3 also represents the relationship between the true phoneme accuracy and the estimated phoneme accuracy when the mel-scaled log power spectrum was used as noise characteristics. The coefficient of determination and the RMSE (Root Mean Square Error) were 0.92 and 3.89, respectively. It is confirmed that the proposed method gives relatively accurate estimates of the phoneme accuracy when the mel-scaled log power spectrum was used.

4. Conclusion

To improve the usability of speech recognition in noise environments, this paper proposed a method for estimating the recognition performance expected at the next moment from the characteristics of the noise observed previously. We conducted an experiment to evaluate the effectiveness of the proposed method. As a result, it was confirmed that the proposed method gives relatively accurate estimates of the recognition performance when the mel-scaled log power spectrum was used. As future work, we plan to investigate the optimal length of the noise segment and the suitable representation of the noise characteristics.

References

[1] J.-C. Junqua and J.-P. Haton, “Robustness in automatic speech recognition: fundamentals and applications,” Kluwer Academic Pub., 1996.

[2] 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. 6, pp. 2006-2013, Nov. 2006.

[3] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, “The HTK Book (for HTK Version 3.3),” Cambridge Univ. April 2005.

[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] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “JNAS Japanese speech corpus for large vocabulary continuous speech recognition research,” The Journal

of the Acoustical Society of Japan (E), Vol. 20, No.3, pp. 199-206, May 1999.

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

[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.