PAPER

Performance estimation of noisy speech recognition using spectral distortion and recognition task complexity

Ling Guo*, Takeshi Yamada†, Shigeki Miyabe‡,
Shoji Makino§ and Nobuhiko Kitawaki¶

Graduate School of Systems and Information Engineering, University of Tsukuba,
1–1–1 Tennodai, Tsukuba, 305–8573 Japan

(*Received 2 December 2015, Accepted for publication 1 June 2016*)

Abstract: Previously, methods for estimating the performance of noisy speech recognition based on a spectral distortion measure have been proposed. Although they give an estimate of recognition performance without actually performing speech recognition, no consideration is given to any change in the components of a speech recognition system. To solve this problem, we propose a novel method for estimating the performance of noisy speech recognition, a major feature of which is the ability to accommodate the use of different noise reduction algorithms and recognition tasks by using two cepstral distances (CDs) and the square mean root perplexity (SMR-perplexity). First, we verified the effectiveness of the proposed distortion measure, i.e., the two CDs. The experimental results showed that the use of the proposed distortion measure achieves estimation accuracy equivalent to the use of the conventional distortion measures, the perceptual evaluation of speech quality (PESQ) and the signal-to-noise ratio (SNR) of noise-reduced speech, and has the advantage of being applicable to noise reduction algorithms that directly output the mel-frequency cepstral coefficient (MFCC) feature. We then evaluated the proposed method by performing a closed test and an open test (10-fold cross-validation test). The results confirmed that the proposed method gives better estimates without being dependent on the differences among the noise reduction algorithms or the recognition tasks.

Keywords: Performance estimation, Noisy speech recognition, Noise reduction, Spectral distortion, Recognition task complexity

PACS number: 43.60.Lq, 43.66.Jh [doi:10.1250/ast.37.286]

1. INTRODUCTION

Speech recognition services are becoming more prevalent with the spread of smartphones and tablets. However, current speech recognition systems still have a serious problem, namely, the recognition performance is degraded in noisy environments [1]. The degree of performance degradation depends on the nature of ambient noise. To ensure a satisfactory quality of experience (QoE) and to determine a system configuration suitable for each individual noise environment before starting a speech recognition service, it is essential to establish a method that can be used to efficiently investigate recognition performance in different noisy environments.

One typical approach is to prepare noisy speech data in the target noise environment and then perform a recognition experiment. However, this has high computational complexity and is time-consuming. An alternative approach to this problem is to estimate recognition performance based on the spectral distortion between noisy speech and its original clean version [2–4]. The original clean speech is available since we assume that the noisy speech is generated by recording the noise in different noisy environments and artificially adding it to the clean speech. This assumption is reasonable for both approaches from the viewpoint of reducing the recording cost. In the latter approach, a significant reduction of the amount of speech data to be processed is expected by using artificial voice signals with average speaker characteristics [3].

These methods give an estimate of recognition performance without actually performing speech recognition. Previously, we proposed a performance estimation method using the perceptual evaluation of speech quality (PESQ)

*e-mail: guoling@mmlab.cs.tsukuba.ac.jp
†e-mail: takeshi@cs.tsukuba.ac.jp
‡e-mail: miyabe@tara.tsukuba.ac.jp
§e-mail: maki@tara.tsukuba.ac.jp
¶e-mail: kitawaki.nobuhiko.gu@u.tsukuba.ac.jp

[5] as a distortion measure [3]. In this method, an estimator, which is a function of the PESQ score, is trained on the basis of the relationship between the recognition performance and the PESQ score. Fukumori *et al.* also proposed a method using the PESQ and room acoustic parameters to estimate recognition performance in noisy and reverberant environments [4].

Although these methods can give an accurate estimate of recognition performance, no consideration is given to any change in the components of a speech recognition system. For example, a noise reduction algorithm, which is used in the preprocessing stage, may be changed depending on the nature of the ambient noise in order to achieve better recognition performance. A recognition task, which decides what a speech recognition system is capable of recognizing, is often changed according to the content of a service. In general, a complex task makes speech recognition difficult. These changes require an estimator specialized for each individual noise reduction algorithm and recognition task. The training of such a specialized estimator is however labor intensive and time-consuming.

To solve this problem, we recently proposed a performance estimation method using the signal-to-noise ratio (SNR) of noise-reduced speech in addition to the PESQ as distortion measures [6]. This method can estimate recognition performance without being dependent on the differences among noise reduction algorithms. We also proposed a method using the PESQ as a distortion measure and square mean root perplexity (SMR-perplexity) [7] as a task complexity measure [8]. It allows recognition performance to be estimated for different recognition tasks. In this paper, we integrate these two methods to handle both different noise reduction algorithms and different recognition tasks, and propose a novel performance estimation method. The proposed method estimates recognition performance using two cepstral distances (CDs) and the SMR-perplexity. The estimator is defined as a function of these three variables and can be used with different noise reduction algorithms and for different recognition tasks without any additional training.

The rest of this paper is organized as follows. In Sect. 2, the proposed method is explained in detail. In Sect. 3, we evaluate the effectiveness of the proposed method. Section 4 summarizes the work.

2. PROPOSED METHOD

Figure 1 shows an overview of the proposed method. First, a distortion value, which represents the spectral distortion between the noisy/noise-reduced speech and its original clean version, is calculated. Then recognition performance is estimated from the distortion value and a task complexity value. In the proposed method, the distortion measure, the task complexity measure, and the estimator are all issues that need to be addressed and are therefore described in detail below.

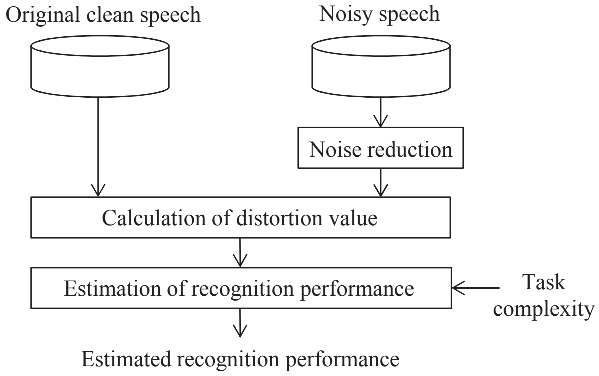

Original clean speech
Noisy speech
Noise reduction
Calculation of distortion value
Estimation of recognition performance
Task complexity
Estimated recognition performance

Fig. 1 Overview of the proposed method.

2.1. Distortion Measure

As mentioned above, we previously proposed a performance estimation method using the PESQ and the SNR of noise-reduced speech as distortion measures [6]. Although this method can estimate recognition performance without being dependent on the differences among noise reduction algorithms, one problem remains, namely, that the waveform of noise-reduced speech is required for the calculation of the PESQ score. Therefore, it is not easily applicable to noise reduction algorithms that directly output a speech feature for speech recognition.

To cope with this problem, we propose the use of two CDs instead of the PESQ and the SNR of noise-reduced speech. One is the CD calculated in the speech frames and the other is the CD calculated in the non-speech frames.

In this paper, the mel-frequency cepstral coefficient (MFCC) with the 0th-order coefficient is used for calculating the two CDs. The MFCC is widely used as a speech feature for speech recognition. There are many noise reduction algorithms that directly output not the waveform but the MFCC. The CD in the speech frames, $\mathrm{CD_s}$, is defined by

$$\mathrm{CD_s} = \frac{1}{N_{\mathbf{M_s}}} \sum_{m \in \mathbf{M_s}} \left\{ \frac{1}{K+1} \sum_{k=0}^{K} |c_\mathrm{d}(k;m) - c_\mathrm{r}(k;m)|^2 \right\}, \quad (1)$$

where m is the frame index, k is the cepstral index, K is the cepstral analysis order, and $c_\mathrm{r}(k;m)$ and $c_\mathrm{d}(k;m)$ are the MFCC of the reference speech (original clean speech) and the degraded speech (noisy speech or noise-reduced speech), respectively. $\mathbf{M_s}$ is a set of indexes of speech frames and $N_{\mathbf{M_s}}$ is the number of elements of $\mathbf{M_s}$. Although there are several distance measures including the Mahalanobis distance, we first used the simplest Euclidean distance. The CD in the non-speech frames, $\mathrm{CD_n}$, is calculated in the same manner. The PESQ is one of the objective quality measures for coded speech and it evaluates the speech quality on the basis of speech distortion. CD_s also represents the spectral distortion; thus, we adopted CD_s in place of the PESQ. The SNR of noise-reduced speech corresponds to the ratio of the power in the speech frames to the power in the non-speech frames. On the other hand, CD_n represents the spectral distortion of the noise, but this distortion value is strongly affected by the noise power since the 0th-order coefficient is used. Because the SNR of noise-reduced speech changes according to the noise power when the speech power is constant, we adopted CD_n in place of it. In the proposed method, a conventional power-based speech/non-speech detection algorithm, which is used in the PESQ [5], is applied to the original clean speech.

2.2. Task Complexity Measure

As a task complexity measure, the size of vocabulary, the perplexity, and its derivation, can be taken into account. Since it was confirmed that the SMR-perplexity is appropriate for estimating recognition performance for different recognition tasks [8], the SMR-perplexity is also adopted in the proposed method.

The SMR-perplexity is expressed in the following form:

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

where $P(\,|\,)$ is the word occurrence probability, $w_1, w_2, \cdots, w_n$ are the words, and n is the number of words. The symbol "$\cdot$" means the beginning or ending of a sentence. This measure overcomes a drawback that the conventional perplexity does not exactly represent the task complexity because the deviations of the sentence length and the number of branches are insufficiently normalized. The SMR-perplexity changes according to both the vocabulary size and the language model used. The larger the SMR-perplexity, the more difficult the recognition task is.

2.3. Estimator

In the previous method [3], the estimator was expressed in the following form:

$$y = f(x) = \frac{a}{1 + e^{-bx+c}}, \tag{3}$$

where y and x represent the estimated word accuracy and the PESQ score, respectively. The constants a, b, and c correspond to the word accuracy for clean speech, the rate of performance degradation, and robustness against spectral distortion, respectively. These constants are determined by data fitting using the iterative least-squares method, that is, by approximating the relationship between the word accuracy and the PESQ score obtained by using a single noise reduction algorithm in different noisy environments.

The estimator in Eq. (3) was then expanded to deal with two types of distortion values: the PESQ score and the SNR of noise-reduced speech [6],

$$y = f(x_1, x_2) = \frac{a}{1 + e^{-b_1 x_1 - b_2 x_2 + c}}, \tag{4}$$

where x_1 and x_2 indicate the PESQ score and the SNR of the noise-reduced speech, respectively. These constants are determined by approximating the relationship between the word accuracy, the PESQ score, and the SNR of the noise-reduced speech obtained by using different noise reduction algorithms in different noisy environments. The estimator in Eq. (3) was also modified as follows to introduce the SMR-perplexity as a task complexity measure [8]:

$$y = f(x, \alpha) = \frac{g_a(\alpha)}{1 + e^{-g_b(\alpha)x + g_c(\alpha)}}, \tag{5}$$

where x and α indicate the PESQ score and the SMR-perplexity, respectively. These constants are determined by approximating the relationship between the word accuracy, the PESQ score, and the SMR-perplexity obtained by using a single noise reduction algorithm in different noisy environments and for different recognition tasks. In Eq. (5), each constant in Eq. (3) is replaced by a function of the SMR-perplexity α. This was motivated by the fact that the task complexity considerably affects the constants in Eq. (3) as shown in the experiment below. In this paper, we newly integrate Eqs. (4) and (5) to handle both different noise reduction algorithms and different recognition tasks. The estimator in the proposed method is expressed by

$$y = f(x_1, x_2, \alpha) = \frac{g_a(\alpha)}{1 + e^{-g_{b_1}(\alpha)x_1 - g_{b_2}(\alpha)x_2 + g_c(\alpha)}}, \tag{6}$$

where x_1 and x_2 represent the two CDs.

To decide the form of the function of the SMR-perplexity α in Eq. (6), we first trained the estimator defined by Eq. (4) for each of different recognition tasks with different SMR-perplexity values. Here, note that two CDs are used, x_1 and x_2 in Eq. (4). In the training, four different noises, seven values of SNR (including clean speech), and five different noise reduction algorithms are used. The details of the recognition tasks and the other conditions are described in Sect. 3. We then investigated the relationship between each of the constants a, b_1, b_2, and c in Eq. (4) and the SMR-perplexity. Figure 2 shows the relationship between each constant and the SMR-perplexity. In this figure, each point represents the SMR-perplexity calculated for one of the recognition tasks and the value of

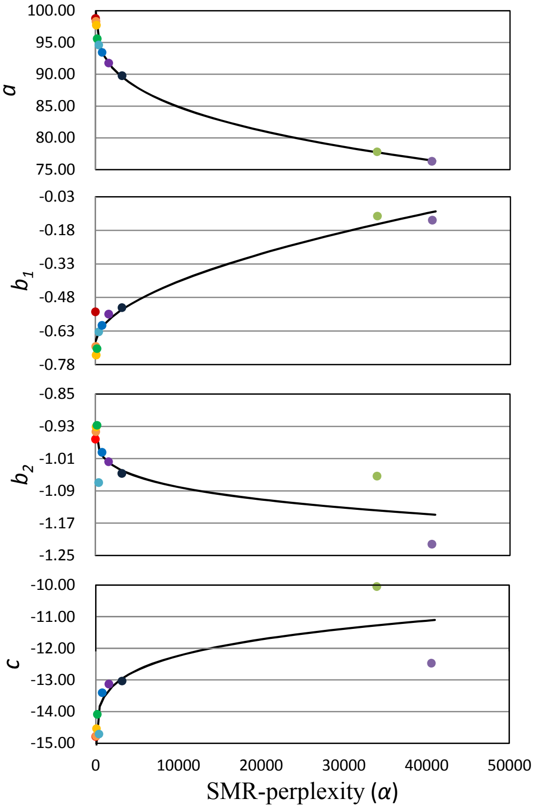


Fig. 2 Relationship between each of the constants a, b_1, b_2, c in Eq. (4) and the SMR-perplexity α.

one of the constants in the estimator specialized for that recognition task. It can be seen that each constant can be represented by an exponential function of the SMR-perplexity, shown as the solid line in the figure. The form of the function was then decided as follows.

$$\begin{aligned} g_a(\alpha) &= p_1 \cdot \alpha^{q_1} + r_1 \\ g_{b_1}(\alpha) &= p_2 \cdot \alpha^{q_2} + r_2 \\ g_{b_2}(\alpha) &= p_3 \cdot \alpha^{q_3} + r_3 \\ g_c(\alpha) &= p_4 \cdot \alpha^{q_4} + r_4 \end{aligned} \tag{7}$$

The constants $p_\cdot$, $q_\cdot$, and $r_\cdot$ are determined by approximating the relationships among the word accuracy, the two CDs, and the SMR-perplexity obtained for different noisy environments, noise reduction algorithms, and recognition tasks.

3. EVALUATION

In this section, we first explain the experimental conditions and then verify the effectiveness of the proposed distortion measure, i.e., the two CDs. Finally, we evaluate the proposed method by performing a closed test and an open test (10-fold cross-validation test).

3.1. Experimental Conditions

We used the following four noise reduction algorithms, in addition to the reference case where no algorithm was used.

- noise reduction is not used (NON)
- minimum mean square error short-time spectral amplitude estimator (MMSE) [9]
- Wiener filtering (WF) [10]
- advanced front-end of ETSI ES 202 050 (AFE) [11]
- stereo-based piecewise linear compensation for environments (SPLICE) [12]

NON, MMSE, and WF output the waveform of noise-reduced speech, and AFE and SPLICE output the MFCC feature of noise-reduced speech. In AFE, Wiener filtering is applied twice and blind equalization is performed in the feature domain. These methods are commonly used as a baseline method when developing a noise reduction algorithm to improve recognition performance and are the basis of many derivations.

We prepared the following recognition tasks and clean-speech data corresponding to each task. The sampling rate of all the speech data described below is 16 kHz. In this paper, we assume that out of vocabulary (OOV) words do not exist (or the rate of the OOV words is small).

- Isolated word recognition: We used the Tohoku University-Matsushita spoken word database [13], consisting of 3,285 isolated words (railway station names) uttered by 12 male and female speakers. The dictionary size was set to 50, 100, 200, 400, 800, 1,600, and 3,200. The number of speech data for each speaker was the same as the dictionary size and OOV words did not exist.
- Grammar-based recognition: The speech data used were 4,004 connected-digit utterances by 52 male and female speakers, which were the same as those in the AURORA-2J database [14]. The grammar allows arbitrary repetitions of digits, a short pause, and a terminal silence.
- Large-vocabulary continuous speech recognition: We used two sets of 100 sentence utterances by 23 male speakers included in the Japanese Newspaper Article Sentences (ASJ-JNAS) database [15]. The vocabulary size was set to 5,000 words and 20,000 words, and the rate of OOV words for each set was 0.14% and 0.03%, respectively. The language models were word 3-gram models with 5,000 and 20,000 words [16], which were trained with the Balanced Corpus of Contemporary Written Japanese [17].

The SMR-perplexity for each recognition task is summarized in Table 1. The recognition experiments and the training of each estimator were performed using all the speech data in each task.

As ambient noise, we used eight forms of noise data, car1, hall1, train2, lift2, factory1, road2, crowd, and lift1, included in the Denshikyo noise database [18]. The former

Table 1 SMR-perplexity α for each task.

Task		α
Isolated word recognition	50 words	50
	100 words	100
	200 words	200
	400 words	400
	800 words	800
	1,600 words	1,600
	3,200 words	3,200
Grammar-based recognition	Connected digits	11
Large-vocabulary continuous speech recognition	5,000 words	40,588
	20,000 words	33,975

four forms of noise data were used for training and the latter were used for testing. 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.

The acoustic models are gender-independent monophone models with 16 Gaussians per state [16], which are trained with the clean-speech data from the ASJ-JNAS database. In order to prevent the difference in the acoustic models used from affecting the recognition performance, we used common monophone models. We used the Julius speech recognizer [19] (rev. 4.3.1) to perform the recognition experiment for all the recognition tasks. The feature vector in Julius has 25 components consisting of 12 MFCCs, 12 delta MFCCs, and a delta log-power. The recognition performance is represented by the word accuracy (*%Acc*), which is defined by

$$\%Acc = \frac{H - I}{N}, \tag{8}$$

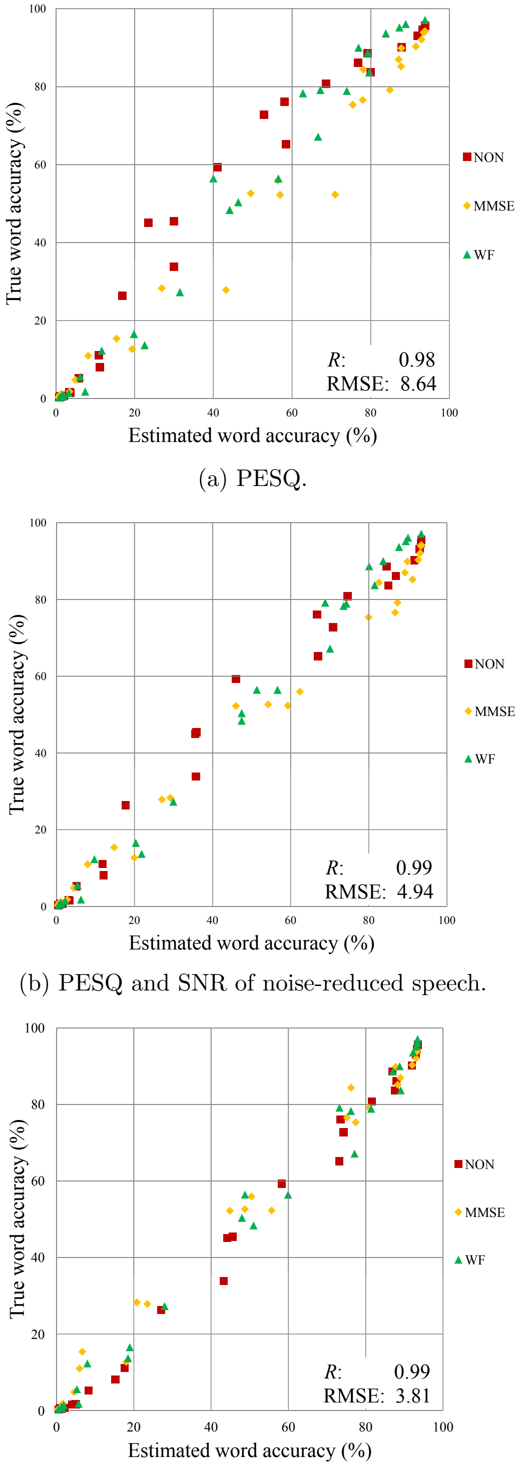
where H, I, and N indicate the number of correct words, the number of erroneously inserted words, and the total number of words, respectively.

3.2. Verification of the Proposed Distortion Measure

In this subsection, we verify the effectiveness of the two CDs proposed. For this purpose, we first compare the following distortion measures.

- PESQ: The estimator defined by Eq. (3) is used.
- PESQ and SNR of noise-reduced speech [6]: The estimator defined by Eq. (4) is used.
- Two CDs: The estimator defined by Eq. (4) is also used, but the two CDs are used as x_1 and x_2 in Eq. (4).

In this experiment, only isolated word recognition with 800 words is used as a recognition task. This is so that only the variation in recognition performance caused by different noise reduction algorithms is observed. The noise reduction algorithms used for training and testing of each estimator are NON, MMSE, and WF, since the PESQ requires the waveform of the noise-reduced speech.



(a) PESQ.

(b) PESQ and SNR of noise-reduced speech.

(c) Two CDs.

Fig. 3 Relationship between the true word accuracy and the word accuracy estimated using each of the distortion measures.

Figure 3 shows the relationship between the true word accuracy and the word accuracy estimated using each of the distortion measures, with the correlation coefficient R and the root mean square error (RMSE). In these figures,

Table 2 RMSE for each of the ten recognition tasks.

SMR-perplexity Method	11	50	100	200	400	800	1,600	3,200	33,975	40,588	average
PESQ	8.93	8.85	8.88	8.41	8.59	8.64	9.96	11.27	13.62	13.87	10.10
PESQ & output SNR	7.61	6.69	6.65	5.82	4.67	4.94	5.15	6.25	8.77	9.43	6.59
Two CDs	5.17	5.43	4.16	4.34	3.70	3.81	4.26	5.18	5.64	5.62	4.73

each point represents the result obtained using one of the noise reduction algorithms for one of the 25 noise conditions, that is, 4 (noise data) × 6 (SNRs) conditions plus one clean case. From Fig. 3(a), it can be seen that there is significant variation. This is caused by the difference in the noise reduction algorithms. On the other hand, the use of the PESQ and the SNR of noise-reduced speech gives good estimates without being dependent on the difference in the algorithms. Figure 3(c) also shows that the use of the two CDs achieves estimation accuracy equivalent to the use of the PESQ and the SNR of noise-reduced speech. Furthermore, we conducted the same experiments for each of the remaining recognition tasks. The results are shown in Table 2. From the table, we can again confirm the validity of the use of the two CDs.

Next, we verify the effectiveness of the two CDs using all the noise reduction algorithms. The noise reduction algorithms used for training and testing are AFE and SPLICE, which directly output the MFCC feature for speech recognition, in addition to NON, MMSE, and WF. The recognition task is isolated word recognition with 800 words. The estimator defined by Eq. (4) is used, but the two CDs are used as x_1 and x_2 in Eq. (4).

Figure 4 shows the relationship between the true word accuracy and the word accuracy estimated using the two CDs. As can be seen from this figure, the use of the two CDs again achieves estimation accuracy equivalent to that in Fig. 3(c), in spite of the increased number of noise reduction algorithms.

These results show that the use of the two proposed CDs is effective compared with the use of the PESQ and the SNR of noise-reduced speech, since they are applicable to the noise reduction algorithms that directly output the MFCC feature.

3.3. Verification of the Proposed Method

In this subsection, we verify the effectiveness of the proposed method for different noise reduction algorithms and recognition tasks. For this purpose, we first compare the proposed method with the use of the two CDs, using the method described in Sect. 3.2. The efficacy of the introduction of the SMR-perplexity is investigated by this comparison.

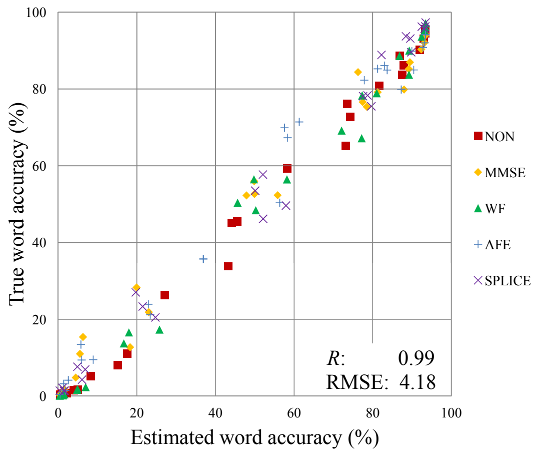


Fig. 4 Relationship between the true word accuracy and the word accuracy estimated using the two CDs.

- Two CDs: The estimator defined by Eq. (4) is used, and the two CDs are used as x_1 and x_2 in Eq. (4).
- Proposed method: The estimator defined by Eqs. (6) and (7) is used.

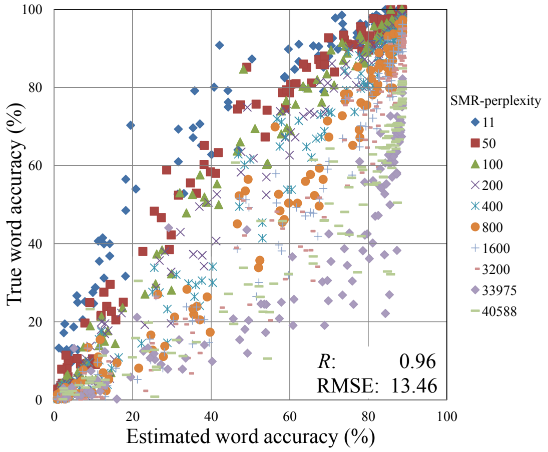
In this experiment, all the recognition tasks and the noise reduction algorithms described in Sect. 3.1 are used for the training and testing of each estimator. The estimator of the use of the two CDs was decided as follows:

$$y = \frac{88.71}{1 + e^{0.77x_1 + 0.71x_2 - 12.12}}. \tag{9}$$

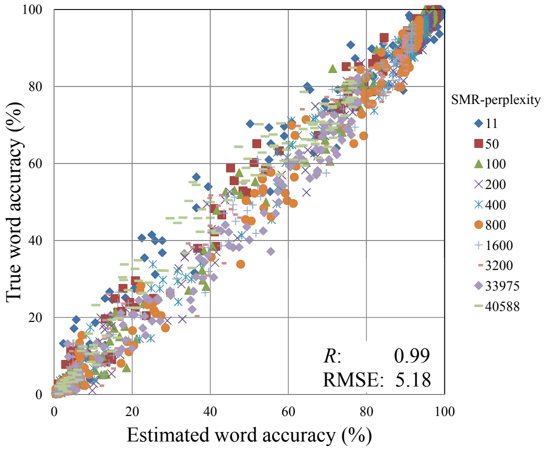
The estimator of the proposed method was also trained as follows:

$$\begin{aligned} y &= \frac{g_a(\alpha)}{1 + e^{-g_{b_1}(\alpha)x_1 - g_{b_2}(\alpha)x_2 + g_c(\alpha)}}, \\ g_a(\alpha) &= -7.08 \cdot \alpha^{0.13} + 110.22 \\ g_{b_1}(\alpha) &= 0.03 \cdot \alpha^{0.28} - 0.83 \\ g_{b_2}(\alpha) &= 0.48 \cdot \alpha^{-0.12} - 1.23 \\ g_c(\alpha) &= 0.62 \cdot \alpha^{0.20} - 16.17 \end{aligned} \tag{10}$$

Figure 5 shows the relationship between the true word accuracy and the word accuracy estimated by each method. In these figures, each point represents the result obtained using one of the noise reduction algorithms for one of the noise conditions and one of the recognition tasks. From



(a) Two CDs.



(b) Proposed method.

Fig. 5 Relationship between the true word accuracy and the word accuracy estimated by each method.

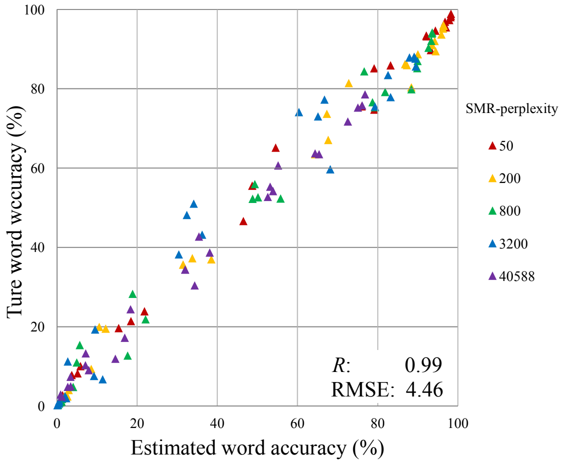


Fig. 6 Relationship between the true word accuracy and the word accuracy estimated by the proposed method.

Fig. 5(a), it can be seen that there is significant variation. This means that the estimator in the use of the two CDs cannot eliminate dependence on the recognition tasks. On the other hand, we can confirm that the proposed method gives better estimates without being dependent on the differences in the recognition tasks. The RMSE of 5.18 for the proposed method is comparable to the average RMSE of 4.73 for the use of the two CDs shown in Table 2, which was obtained under the easier condition that only a single recognition task is considered.

Finally we evaluate the effectiveness of the proposed method using an open test (*noise reduction algorithm and recognition task complexity open test*). For this purpose, we perform a 10-fold cross-validation test. A testing set comprises one of the five noise reduction algorithms and five of the ten recognition tasks. The training set for this test set comprises the remaining four noise reduction algorithms and five recognition tasks. Note here that the recognition tasks are divided into two groups: one has tasks where $\alpha = 11$, 100, 400, 1,600, and 33,975 and the other has tasks where $\alpha = 50$, 200, 800, 3,200, and 40,588. The number of pairs in the test set and the training set therefore becomes ten. In this experiment, the estimator defined by Eqs. (6) and (7) is used.

Figure 6 shows the relationship between the true word accuracy and the word accuracy estimated by the proposed method. This figure is an estimation result for the testing set with the MMSE and the recognition tasks with $\alpha = 50$, 200, 800, 3,200, and 40,588. In this figure, each point represents the result obtained using the MMSE for one of the noise conditions and one of the recognition tasks. It can be seen that the proposed method achieves estimation accuracy equivalent to that in Fig. 5(b), corresponding to a closed test (*noise reduction algorithm and recognition task complexity closed test*). The correlation coefficient R and the RMSE for each of the ten testing sets are summarized in Table 3. The left column and the top row indicate the recognition tasks and the noise reduction algorithm in each testing set, respectively. From Table 3, we can see that there is little difference among the RMSEs of the ten test sets. The correlation coefficient R and the RMSE averaged over the ten testing sets are 0.98 and 5.71, respectively. This result is comparable to the RMSE of 5.18 for the closed test.

From these results, it is confirmed that the proposed method gives better estimates without being dependent on the differences among the recognition tasks or the noise reduction algorithms.

4. CONCLUSION

We proposed a performance estimation method for noisy speech recognition, in which the major feature is the ability to accommodate the use of different noise reduction algorithms and recognition tasks by using two CDs and the SMR-perplexity. First, we verified the effectiveness of the

Table 3 Correlation coefficient R (left) and RMSE (right) for each of the ten testing sets.

Noise reduction algorithm / SMR-perplexity	NON	MMSE	WF	AFE	SPLICE
$\alpha = 11, 100, 400, 1{,}600, 33{,}975$	0.98/6.20	0.99/5.18	0.98/6.16	0.98/5.51	0.98/6.06
$\alpha = 50, 200, 800, 3{,}200, 40{,}588$	0.99/6.16	0.99/4.46	0.98/7.10	0.99/5.07	0.98/5.18

proposed distortion measure, i.e., the two CDs. The experimental results showed that the use of the proposed distortion measure achieves estimation accuracy equivalent to the use of the conventional distortion measures of the PESQ and the SNR of noise-reduced speech, and has the advantage of being applicable to noise reduction algorithms that directly output the MFCC feature. We then evaluated the proposed method by performing a closed test and an open test (10-fold cross-validation test). The results confirm that the proposed method gives better estimates without being dependent on the differences among the noise reduction algorithms or the recognition tasks.

REFERENCES

[1] J. Li, L. Deng, Y. Gong and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **22**, 745–777 (2014).

[2] H. Sun, L. Shue and J. Chen, "Investigations into the relationship between measurable speech quality and speech recognition rate for telephone speech," *Proc. ICASSP 2004*, Vol. 1, pp. 865–868 (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 Trans. Audio Speech Lang. Process.*, **14**, 2006–2013 (2006).

[4] T. Fukumori, M. Nakayama, T. Nishiura and Y. Yamashita, "Estimation of speech recognition performance in noisy and reverberant environments using PESQ score and acoustic parameters," *Proc. APSIPA ASC 2013*, Paper ID: 144 (2013).

[5] 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" (2001).

[6] L. Guo, T. Yamada, S. Makino and N. Kitawaki, "Performance estimation of noisy speech recognition using spectral distortion and SNR of noise-reduced speech," *Proc. TENCON 2013*, Paper ID: 540 (2013).

[7] S. Nakagawa and M. Ida, "A new measure of task complexity for continuous speech recognition," *Trans. IEICE*, **J81-D-2**, 1491–1500 (1998) (in Japanese).

[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 (2010).

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Audio Speech Lang. Process.*, **32**, 1109–1121 (1984).

[10] P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, 2013).

[11] ETSI ES 202 050 v1.1.5, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms" (2007).

[12] L. Deng, A. Acero, M. Plumpe and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," *Proc. INTERSPEECH 2000*, pp. 806–809 (2000).

[13] S. Makino, K. Niyada, Y. Mafune and K. Kido, "Tohoku University and Matsushita isolated spoken word database," *J. Acoust. Soc. Jpn.* (*J*), **48**, 899–905 (1992) (in Japanese).

[14] S. Nakamura, K. Takeda, K. Yamamoto, T. Tamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima and M. Fujimoto, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," *IEICE Trans. Inf. Syst.*, **E88-D**, 535–544 (2005).

[15] 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," *J. Acoust. Soc. Jpn.* (*E*), **20**, 199–206 (1999).

[16] 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 tool kit for Japanese large vocabulary continuous speech recognition," *Proc. ICSLP 2000*, pp. 476–479 (2000).

[17] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka and Y. Den, "Balanced corpus of contemporary written Japanese," *Lang. Resour. Eval.*, **48**, 345–371 (2004).

[18] Denshikyo noise database, http://research.nii.ac.jp/src/list/detail.html#JEIDA-NOISE (accessed 2015-12-01).

[19] Julius, http://julius.osdn.jp/ (accessed 2015-12-01).

Ling Guo received her B.S. degree from Jiaxing University, China, in 2010, and her M. Eng. degree from University of Tsukuba, Japan, in 2014. She is currently working toward her Ph.D. Her research interests includes the performance estimation of noisy speech recognition. She is a student member of the ASJ.

Takeshi Yamada received his B. Eng. degree from Osaka City University, Japan, in 1994, and his M. Eng. and Dr. Eng. degrees from Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. He is presently an associate professor with Faculty of Engineering, Information and Systems, University of Tsukuba, Japan. His research interests include speech recognition, sound scene understanding, multichannel signal processing, media quality assessment, and e-learning. He is a member of the IEEE, the IEICE, the IPSJ, the ASJ, and the JLTA.

Shigeki Miyabe received his B.E. degree from Kobe University, Japan, in 2003, and his M.E. and Ph.D. degrees from Nara Institute of Science and Technology, Japan, in 2005 and 2007, respectively. From 2008 to 2009, he was a visiting scholar with Georgia Institute of Technology, USA. In 2009, he joined the Graduate School of Information Science and Technology, University of Tokyo, Japan, as a researcher, and became an assistant professor in 2010. He is currently an assistant professor of Life Science Center of Tsukuba Advanced Research Alliance, University of Tsukuba, Japan. He is a member of the IEEE and the ASJ.

Shoji Makino received his B.E., M.E., and Ph.D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981. He is now a professor at University of Tsukuba, Japan. His research interests include adaptive filtering technologies, the realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech, and acoustic signal processing for speech and audio applications. He is an IEEE Fellow, an IEICE Fellow, a council member of the ASJ, and a member of the EURASIP.

Nobuhiko Kitawaki received his B. Eng., M. Eng. and Dr. Eng. degrees from the Tohoku University, Japan, in 1969, 1971, and 1981, respectively. From 1971 to 1997 he was engaged in research on speech and acoustics information processing at NTT Laboratories. From 1997 to 2015, he served as a professor of the Graduate School of Systems and Information Engineering, University of Tsukuba. Emeritus Prof. Kitawaki is an IEEE Fellow, an IEICE Fellow, and a member of the Acoustical Society of Japan.