Performance Estimation of Spontaneous Speech Recognition Using Non-Reference Acoustic Features

Ling Guo, Takeshi Yamada and Shoji Makino
University of Tsukuba, Ibaraki, Japan
E-mail: guoling@mmlab.cs.tsukuba.ac.jp

***Abstract*—To ensure a satisfactory QoE (Quality of Experience), it is essential to establish a method that can be used to efficiently investigate recognition performance for spontaneous speech. By using this method, it is allowed to monitor the recognition performance in providing speech recognition services. It can be also used as a reliability measure in speech dialogue systems. Previously, methods for estimating the performance of noisy speech recognition based on spectral distortion measures have been proposed. Although they give an estimate of recognition performance without actually performing speech recognition, the methods cannot be applied to spontaneous speech because they require the reference speech to obtain the distortion values. To solve this problem, we propose a novel method for estimating the recognition performance of spontaneous speech with various speaking styles. The main feature is to use non-reference acoustic features that do not require the reference speech. The proposed method extracts non-reference features by openSMILE (open-Source Media Interpretation by Large feature-space Extraction) and then estimates the recognition performance by using SVR (Support Vector Regression). We confirmed the effectiveness of the proposed method by experiments using spontaneous speech data from the OGVC (On-line Gaming Voice Chat) corpus.**

I. 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 for spontaneous speech. Therefore, to ensure a satisfactory QoE (Quality of Experience), it is essential to establish a method that can be used to efficiently investigate recognition performance for spontaneous speech. By using this method, it is allowed to monitor the recognition performance in providing speech recognition services. It can also be used as a reliability measure in speech dialogue systems.

One approach is to estimate the recognition performance by using a confidence measure[1], which is originally used to evaluate reliability of recognition results. However, it requires to perform speech recognition with high computational complexity. Another approach is to estimate the recognition performance from a distortion value without actually performing speech recognition[2]–[4]. For example, we previously proposed a method[5] that estimates the recognition performance using the PESQ (Perceptual Evaluation of Speech Quality)[6] and the output SNR (SNR of noise-reduced speech) as distortion measures. Although they give an estimate of recognition performance without actually performing speech recognition, the methods cannot be applied to spontaneous speech because they require the reference speech to obtain the distortion values.

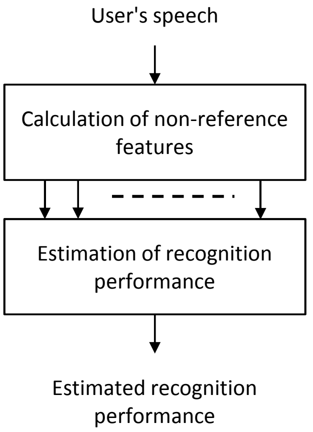

User's speech
Calculation of non-reference features
Estimation of recognition performance
Estimated recognition performance

Fig. 1. Estimation of the recognition performance from non-reference acoustic features.

To solve this problem, we propose a novel recognition performance estimation method for spontaneous speech. The main feature is to use non-reference acoustic features that do not require the reference speech. First, it extracts non-reference features by openSMILE (open-Source Media Interpretation by Large feature-space Extraction) and then estimates the recognition performance by using SVR (Support Vector Regression)[7]. We confirmed the effectiveness of the proposed method by experiments using spontaneous speech data from the OGVC (On-line Gaming Voice Chat) corpus[8].

II. Proposed method

A. Overview

Figure 1 shows an overview of the proposed method. In this method, first the method extracts non-reference acoustic features only from the user's speech. Then, it estimates the recognition performance by SVR.

TABLE I
LOW-LEVEL DESCRIPTORS FOR THE FEATURE SET.

Low-level descriptor	Description
RMSenergy	Root-mean-square signal frame energy
MFCC	Mel-frequency cepstral coefficients 1-12
pcm_zcr	Zero-crossing rate of time signal (frame-based)
voiceProb	Voicing probability computed from the ACF
F0	Fundamental frequency computed from the cepstrum
(a first-order delta coefficient of the above features)	

TABLE II
FUNCTIONALS FOR THE FEATURE SET.

Functional	Description
max	Maximum value of the contour
min	Minimum value of the contour
range	max-min
maxPos	Absolute position of the maximum value (in frames)
minPos	Absolute position of the minimum value (in frames)
amean	Arithmetic mean of the contour
linregc1	Slope (m) of a linear approximation of the contour
linregc2	Offset (t) of a linear approximation of the contour
linregerrQ	Quadratic error computed as the difference between the linear approximation and the actual contour
stddev	Standard deviation of the values in the contour
skewness	Skewness (third-order moment)
kurtosis	Kurtosis (fourth-order moment)

In the proposed method, the definition of recognition performance and non-reference acoustic features used are all issues that need to be addressed and are therefore described in detail below.

B. Definition of recognition performance

Here, we describe a definition of recognition performance. To apply the proposed method to performance monitoring in speech recognition services and the use as a reliability measure in dialogue systems as noted above, it is useful and reasonable to estimate the recognition performance for each individual utterance. We therefore decided to use the word accuracy (%Acc) in one utterance, which is defined by

$$\%Acc = \frac{H}{N}, \tag{1}$$

where H and N are the number of morphemes that are correctly recognized and the total number of morphemes, respectively. However, this definition has the disadvantage that the resolution of the word accuracy becomes low when N decreases. For example, in the case of utterances with only one morpheme, the word accuracy can only be 0 or 100. It is possible that the estimation of the recognition performance will be affected by N.

C. Non-reference acoustic features

As mentioned in Sect. I, to solve the problem that no reference speech exists when dealing with spontaneous speech, we propose the use of non-reference acoustic features. In this paper, we focus on acoustic features for emotion recognition. Since emotional speech can be considered as typical spontaneous speech, it is expected that these features represent the characteristics of spontaneous speech. We therefore decided to adopt the acoustic features of openSMILE (open-Source Media Interpretation by Large feature-space Extraction)[9] with the configuration for the INTERSPEECH 2009 Emotion Challenge feature set[10]. This feature set is widely used for emotion recognition. We expect that these non-reference acoustic features will give a good estimate of the performance of spontaneous speech recognition.

This feature set contains 384 acoustic features, which are the statistical functions of low-level descriptors. The names of the 16 low-level descriptors are shown in Table I. The names of the 12 functionals of the feature set are shown in Table II. These features represent prosodic, spectral, and voice quality features[10].

III. EVALUATION

A. Experimental conditions

To confirm the effectiveness of the proposed method, we prepared an experiment to estimate the performance of spontaneous speech recognition.

We used the Julius Japanese Dictation-kit[11] to perform the recognition experiment. The feature vector has 25 components consisting of 12 MFCCs, 12 delta MFCCs, and a delta log-power. The acoustic models are gender independent triphone models with DNN-HMM (Deep Neural Network-Hidden Markov Model), which are trained with clean speech data from the ASJ-JNAS database. The language models are word 3-gram models which are trained with the Balanced Corpus of Contemporary Written Japanese.

In this paper, we estimate the recognition performance by using SVR. The learning of SVR with the radial basis kernel function is performed using a pair data comprising the feature set and the speech recognition performance for each of the spontaneous speech data.

B. Spontaneous speech data used

We used acting speech data in the OGVC corpus[8] as spontaneous speech data, consisting of 2656 utterances by four speakers (two females and two males). The emotion and strength to be expressed were specified for each utterance. There were eight emotions (fear, surprise, sadness, disgust, anger, anticipation, joy, and acceptance) and four levels of strength (neutral, weak, medium, strong) for each emotion. The sampling rate of the speech data is 16 kHz. The transcription of each speech data is prepared that was used to calculate the word accuracy. The ratio of OOV (Out-Of-Vocabulary) words was 0.03%. In this paper, we do not consider the effect of OOV words on the estimation.

The histogram of the word accuracy in the speech data used is shown in Figure 2. In this figure, the horizontal axis shows a range of the word accuracy and the vertical axis shows the number of utterances. From this figure, we can confirm that the word accuracy is totally low. The average value of the word accuracy was 35.42%.

The histogram of the number of morphemes in the speech data used is shown in Figure 3. In this figure, the horizontal

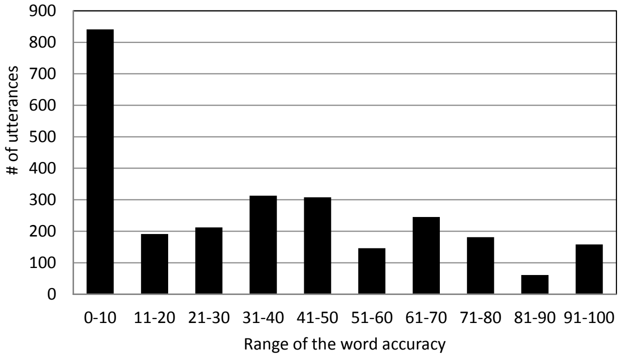


Fig. 2. Histogram of the word accuracy in the speech data used.

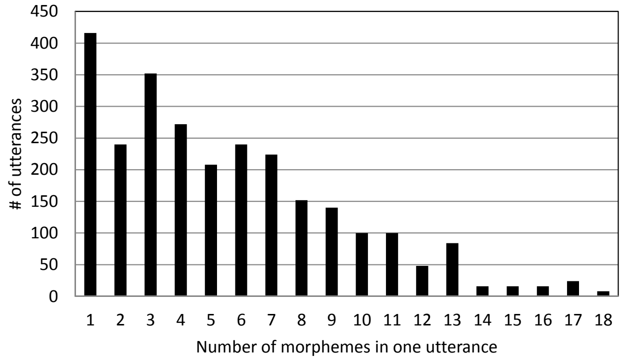


Fig. 3. Histogram of the number of morphemes in the speech data used.

axis shows the number of morphemes in one speech data and the vertical axis shows the number of utterances. From this figure, we can see that many speech data contain only a few morphemes.

C. Verification of the proposed method

In this section, we verify the effectiveness of the proposed method. We prepared 10 sets of speech data. Each set consists of the speech data with n or more morphemes ($n = 1, 2, ..., 10$). This is to investigate the effect of the number of morphemes on the estimation of recognition performance as mentioned in Sect. II.A. We perform a four-fold cross-validation test. Each of the ten speech data sets is divided into four groups. Each group contains the speech data of one speaker. A testing set comprise one of the four speakers, and the training set for this test set comprises the remaining three speakers.

The correlation coefficient R and RMSE (Root Mean Square Error) for each speech data set are shown in Table III. The correlation coefficient and RMSE in the table are the average of those calculated from each cross-validation test set. The correlation coefficient and RMSE are defined as follows:

$$\mathrm{R} = 1 - \frac{\sum(\mathrm{True\%Acc} - \mathrm{Estimated\%Acc})}{\sum(\mathrm{True\%Acc} - \overline{\mathrm{True\%Acc}})}, \qquad (2)$$

TABLE III

CORRELATION COEFFICIENT R AND RMSE FOR EACH SPEECH DATA SET IN SPEAKER-OPEN TEST

Minimum number of morphemes in speech data set	R	RMSE
1	0.59	24.54
2	0.52	24.56
3	0.54	23.38
4	0.59	21.11
5	0.58	20.19
6	0.60	19.12
7	0.62	18.04
8	0.71	15.76
9	0.77	13.77
10	0.84	11.55

and

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum(\mathrm{True\%Acc} - \mathrm{Estimated\%Acc})^2}. \qquad (3)$$

From the table, we can confirm that when the minimum number of morphemes n is small, the RMSE tends to be high. This would be due to the low resolution of the word accuracy. On the other hand, when n is large, the RMSE becomes smaller.

The estimation result for the speech data set with ten or more morphemes shown in Table III are plotted in Figure 4. These figures show the relationship between the true word accuracy and the word accuracy estimated by the proposed method. Each figure is the estimation result for one of the four testing sets of the four-fold cross-validation test. From these results, it is confirmed that the proposed method gives good estimates almost independent on the speakers. The correlation coefficient and RMSE were 0.83–0.85 and 10.90–11.88, respectively.

IV. CONCLUSIONS

In this paper, we proposed a novel method for estimating the recognition performance of spontaneous speech with various speaking styles. The main feature is to use non-reference acoustic features that do not require the reference speech. We confirmed the effectiveness of the proposed method by experiments using spontaneous speech data from the OGVC (On-line Gaming Voice Chat) corpus. The proposed method gives a good estimate of recognition performance when the number of morphemes in one utterance is large, namely, the utterance has a sufficient length.

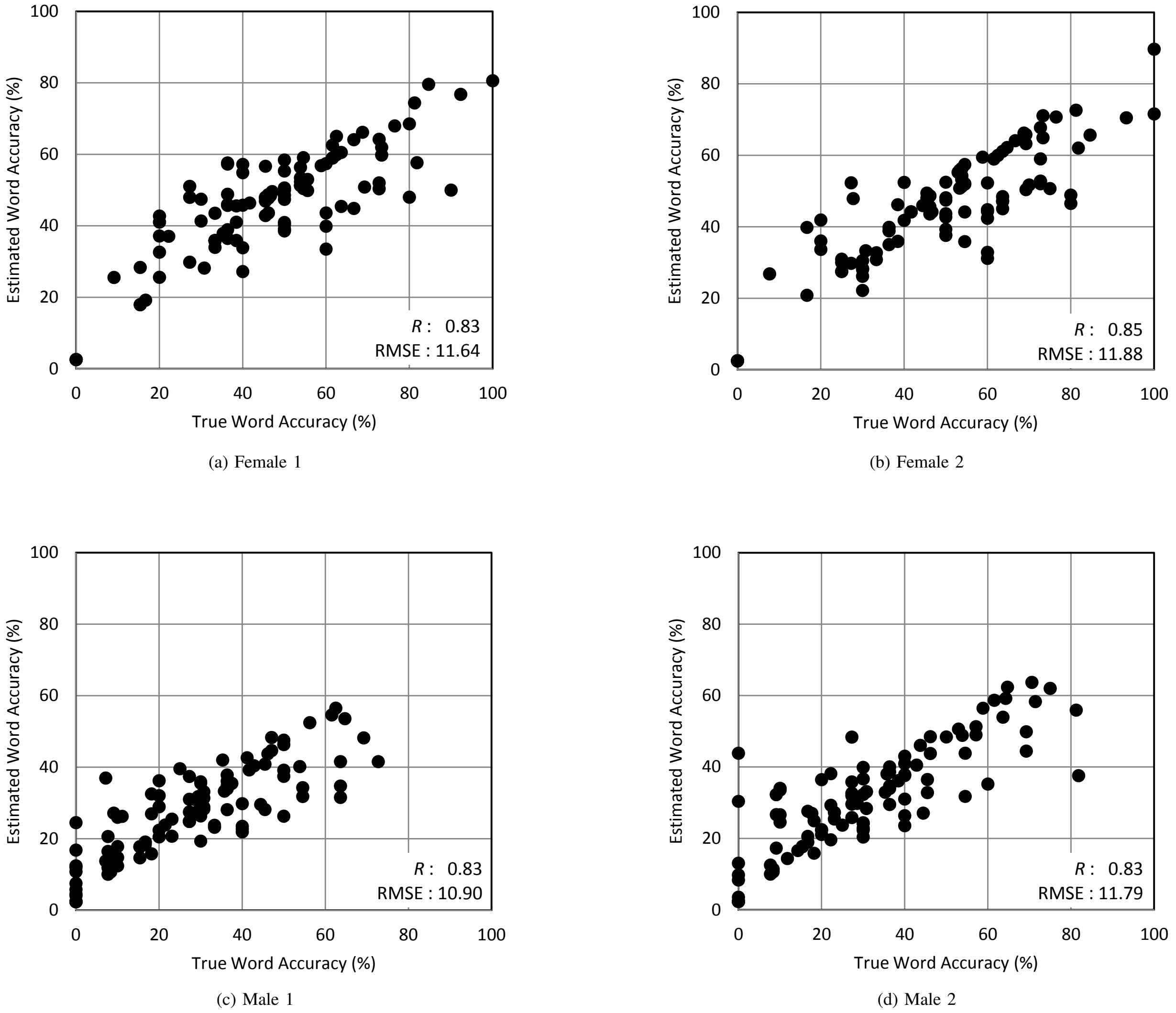
As future work, we plan to adopt DNN in place of SVR in the proposed method.

REFERENCES

[1] J. Hui, "Confidence measures for speech recognition: A survey," Speech communication, Vol. 45, No. 4, pp. 455–470, 2005.

[2] H. Sun, L. Shue, J. Chen, "Investigations into the relationship between measurable speech quality and speech recognition rate for telephone speech," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2004, Vol. 1, 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.



(a) Female 1

(b) Female 2

(c) Male 1

(d) Male 2

Fig. 4. Relationship between the true word accuracy and the estimated word accuracy for each speaker.

[4] T. Fukumori, M. Nakayama, T. Nishiura, Y. Yamashita, “Estimation of speech recognition performance in noisy and reverberant environments using PESQ score and acoustic parameters,” Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2013, Paper ID: 144, Oct. 2013.

[5] L. Guo, T. Yamada, S. Makino, N. Kitawaki, “Performance estimation of noisy speech recognition using spectral distortion and SNR of noise-reduced speech,” Proc. TENCON 2013, Paper ID: 540, Oct. 2013.

[6] ITU-T Rec. P.862, “Perceptual evalation of speech quality(PESQ):An objective method for end-to-end speech quality assesment of narrow-band telephone networks and speech codecs,” Feb. 2001.

[7] C.-C. Chang, C.-J. Lin, “LIBSVM: a library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, 2:27:127:27, 2011.

[8] Online gaming voice chat corpus with emotional label (OGVC), http://research.nii.ac.jp/src/en/OGVC.html

[9] F. Eyben, M. Wollmer, B. Schuller, “openSMILE The Munich Versatile and Fast Open-Source Audio Feature Extractor,” ACM Multimedia Conference MM, pp. 1459–1462 (2010).

[10] B. Schuller, S. Steidl, A. Batliner, “The INTERSPEECH 2009 Emotion Challenge,” Proc. Of INTERSPEECH 2009, pp. 312–315 (2009).

[11] Julius, http://julius.osdn.jp/