Research Report

Estimating Speech-Recognizer Performance Based on Log-Likelihood Difference Distribution of Word-Pairs

Ryuta Terashima

Abstract

This paper describes an efficient method of estimating word recognition rates without speech data. The method is based on the minimum value of the word-pair recognition rate, which correspons to the word recognition rate. The estimated word-pair recognition rate can be calculated by the measured log-likelihood difference distribution that can be obtained by phoneme recognition, and it is assumed that the distribution is approximated by a normal distribution. To illustrate the effectiveness of our method, we evaluated the performance of the proposed method by actual recognition experiments using 3000 word-pairs. The correlation coefficient value between the estimated and the measured recognition rates was 0.87 when the phoneme lengths of the word-pairs were equal. Furthermore, we also evaluated a 95% confidence interval for the measured recognition rates. The percentage of estimated words that fell within the confidence interval was 94.8%.

Keywords: Recognition performance evaluation method, Log-likelihood difference, Word-pair recognition rate, Word recognition rate, Likelihood distribution

1. Introduction

Recently, to improve safety and enhance driver convenience, human-machine interfaces based on speech recognition have been widely adopted for in-vehicle information equipment such as car navigation systems. Mostly, these systems can accept only isolated words or simple sentences described with network grammar, while a wide range of devices can be manipulated by the human voice.

In general, a degradation in the recognition rate causes an increase in a total task time needed to complete the system operations. Shimizu et al. showed that the total task time could be predicted based on the recognition rate.[1] Furthermore, generally, the distribution of the recognition rates is biased. This means that there are some words that tend to cause an increase in the total task time, as their recognition rates are much lower than those of others. Thus, it is important to be able to predict such words.

However, a car navigation system has to have a vocabulary of hundreds of thousands of words, making it difficult to collect evaluation data for all the words and also measure their recognition rates. Therefore, there is a need for a method of estimating word recognition rates without speech data.

It has always been thought that word recognition performance could be determined based on a "word distance" or "word similarity". This could be calculated by the probabilistic distance for HMM,[2] the acoustic similarity between sub-phonemic segments,[3] and so on. However, these would not always correspond to the word recognition rate. To overcome this, several methods for predicting the recognition performance have been proposed.[4-7] Of particular note, Abe et al. proposed a method of predicting the word recognition rate based on the distribution of a recognition score. For this method, however, an assumption of the distribution was not practical and recognition experiments on an actual decoder had not been carried out. In this paper, we propose a method of estimating those words with a low recognition rate for an actual decoder without speech data.

The remainder of this paper consists of the following sections. **Section 2** provides an overview of the basic concept. **Section 3** gives details of the method based on the log-likelihood difference distribution of a word-pair. **Section 4** describes the recognition experiments that we used to evaluate our proposed method, together with the results we obtained. Finally, Section 5 concludes the paper.

2. Outline

2. 1 Overview of proposed method

We estimated the word recognition rate from the word-pair recognition rates. Word-pair recognition is of a recognition type that is used by a dictionary defining only two words. The basic concept of the proposed method is as follows:

(1) Estimation of the word-pair recognition rate for all combinations in the word dictionary.

(2) The minimum value of the word-pair recognition rates for one word $rmin_{wi} = \min(r_{wi,1}, r_{wi,2}, \cdots, r_{wi,M})$ is used as a representative value for the recognition performance value of the word.

Here, M is the number of words in the dictionary. $r_{wi,j}$ denotes the word-pair recognition rate of word w_i for word-pair (w_i, w_j).

$rmin_{wi}$ can be regarded as the upper limit on the $_{wi}$ recognition rate without considering the effect of search errors on the decoders. Therefore, we used it as a representative value for the word recognition performance. We can use the approximation to detect words with low recognition rates.

2. 2 Relationship between word-pair recognition rate and word recognition rate

A word-pair recognition experiment was performed by HTK[8] to clarify the relationship between the word-pair recognition rate and word recognition rate. Four dictionaries (vocabulary size: 100, 200, 400 and 800 words) were prepared, with each smaller dictionary being a subset of the larger ones, and fifty words were used as a test set, all of which were included in all of the dictionaries. The word recognition rates and the minimum word-pair recognition rates were measured for each dictionary.

Figure 1 shows the results of our experiments. The minimum word-pair recognition rate gives not only the upper limit on the word recognition rate but also high correlation with the word recognition rate. Therefore, those words with low recognition rates

can be detected by predicting the word-pair recognition rates.

3. Proposed method

3. 1 Log-likelihood difference

Let word w be composed of phonemes $p_{w1}, p_{w2}, \cdots, p_{wN}$. N is the phoneme length of w. Let lp_{wi} denote the likelihood of the *ith* phoneme segmentation of word w, as derived from the Viterbi forced alignment.

Consider the log-likelihood difference $ld_{w1, w2} = l_{m1} - l_{w2}$ of the word-pair (w_1, w_2). The sign of corresponds to the result of the word-pair recognition; if positive, the recognition result would be w_1. Here, suppose that the phoneme lengths of both words are equal and that the segmentations of each phoneme are identical (**Fig. 2**), therefore, the log-likelihood difference of a word-pair is given by

$$l_{w1} - l_{w2} = \sum_{i=1}^{N} (l_{p_{w1_i}} - l_{p_{w2_i}}) \quad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

3. 2 Distribution of log-likelihood difference

Generally, log-likelihood differences are distributed by the variance in the utterances. If we assume that the distributions of each phoneme are independently determined by the other phonemes, then the distribution of the log-likelihood difference of a word-pair can be calculated from the log-likelihood difference distributions of the phoneme pairs. Furthermore, the difference in the distribution of a word-pair can be independently determined by its phoneme composition.

Our method can estimate the log-likelihood difference distribution of a word-pair, as follows:

(1) Conventionally, by measuring the log-likelihood difference distributions of phoneme pairs for all combinations of phonemes by phoneme recognition.

(2) Translating the words into phoneme sequences.

(3) Calculating the sum of the distributions for all different phoneme pairs.

For example, **Fig. 3** shows that the distribution of ("home", "phone") can be calculated by the sum of the distributions of (/h/, /f/) and (/m/, /n/).

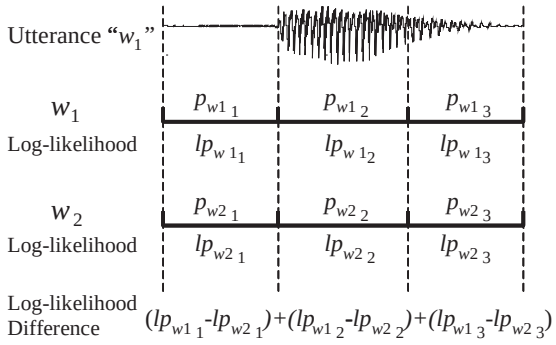

Fig. 2 Log-likelihood difference of word-pair (w_1, w_2).

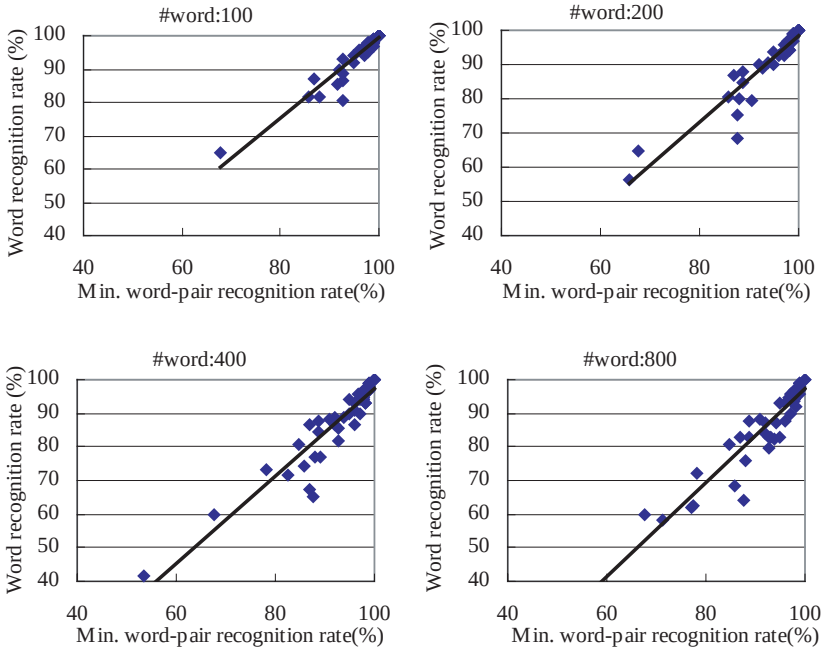

Fig. 1 Relationship between measured minimum word-pair recognition rate and word recognition rate.

Here, we assume that the log-likelihood difference distribution is approximated by a normal distribution. Then, the word recognition rate *rw* is given, as follows:

$$r_w = \int_0^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \quad \cdots\cdots\cdots\cdots\cdots (2)$$

where

$$\mu = \sum_i^N m_{p_{w1_i}, p_{w2_i}}, \quad \sigma^2 = \sum_i^N s^2_{p_{w1_i}, p_{w2_i}} \quad \cdots\cdots\cdots\cdots\cdots (3)$$

$m_{p,\,q}$ and $s_{p,\,q}$ denote the mean and variance of the distribution of (p, q), respectively.

4. Experiments and results

4. 1 Conditions

A speaker-independent isolated word recognition experiment was carried out to evaluate our proposed method. We used the NTT Japanese word database, which consists of control commands ("Open," "Close," "Copy," and so on) and Japanese city names. From these databases, we arbitrarily chose about 3000 word-pairs for use as the test set. HTK was used as the speech recognizer, together with the 1997 IPA standard acoustic model.[9] The detailed conditions are shown in **Table 1**.

The test set was divided into two subsets. Subset 1 consisted of word-pairs for which the phoneme lengths of both words were equal, while subset 2 consisted of the remainder of the test set.

The log-likelihood difference distributions of the phoneme pair were approximated by a normal distribution. The distribution was calculated from the Viterbi-segmented speech data for the database. Thus, the speech data for measuring the actual recognition rate and predicting the rate were identical. Therefore, the following results were obtained by close tests.

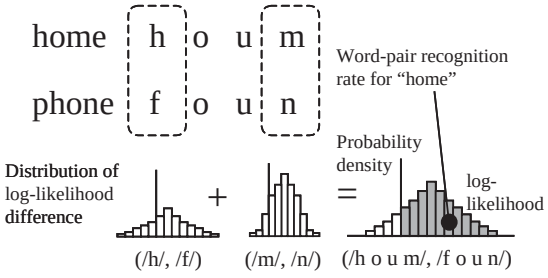

Fig. 3 Prediction of distribution of log-likelihood difference for ("home", "phone") and word recognition rate for "home".

4. 2 Results of experiments with subset 1

The relationship between the predicted and measured recognition rates for subset 1 is plotted in **Fig. 4**. The correlation coefficient between the predicted and the measured recognition rate was 0.87. Furthermore, we also estimated a 95% confidence interval for the measured recognition rates. We assumed that the distributions of the measured word-pair recognition rate could be approximated by a binominal distribution, and that the number of test data used for measuring the actual recognition rate was about 100 per word. Under these conditions, the percentage of the words fell within the confidence interval was 94.8%. These results imply that our method can estimate the

Table 1 Experimental conditions.

Acoustic feature	Sampling frequency: 16kHz Frame shift: 10ms Frame length: 25ms(Hamming) Feature vector: 12th order MFCC + 12th order ΔMFCC + ΔPow
Acoustic model	#state: 3, left-to-right #mixture: 4 #phoneme: 43 monophone, male
Speech data	#speakers: 113 age: 10-60, male

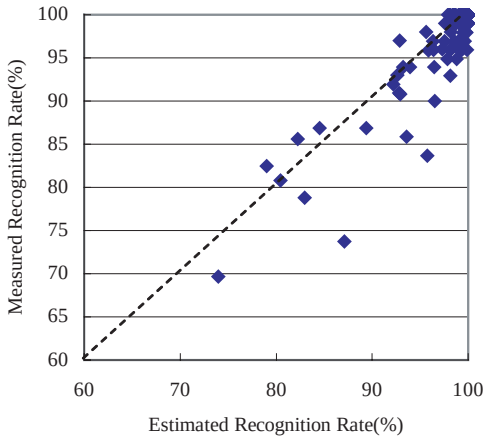

Fig. 4 Comparison of predicted and measured word-pair recognition rate (subset 1).

recognition rate of a word-pair when the word lengths are equal.

4. 3 Results of experiments with subset 2

The results obtained with subset 2 are shown in **Fig. 5**. Since it has been postulated that word lengths are equal in our method, they were corrected by means of DP matching prior to applying our method. The DP path is given by

$$g(\mu_{i,j}, \sigma^2_{i,j}) = \min\{ g(m_{p_{w1_i}, p_{w2_j}} + \mu_{i,j-1}, s^2_{p_{w1_i}, p_{w2_j}} + \sigma^2_{i,j-1}), g(m_{p_{w1_i}, p_{w2_j}} + \mu_{i-1,j}, s^2_{p_{w1_i}, p_{w2_j}} + \sigma^2_{i-1,j}), g(m_{p_{w1_i}, p_{w2_j}} + \mu_{i-1,j-1}, s^2_{p_{w1_i}, p_{w2_j}} + \sigma^2_{i-1,j-1}) \} \quad \cdots\cdots (4)$$

where

$$g(\mu', \sigma') = \int_0^{\infty} \frac{1}{\sqrt{2}\mu\sigma'} \exp(-\frac{(x-\mu')^2}{2\sigma'^2}) dx \quad \cdots\cdots (5)$$

$(\mu_{i,j}, \sigma^2_{i,j})$ denotes the cumulation of the mean and the variance from p_{w11} and p_{w21} to p_{w1i} and p_{w2i} following the DP path.

The correlation coefficient between the predicted and the measured recognition rate was 0.50, and the percentage of the words contained within the confidence interval was 97.0%. The reason why the percentage of the words fell within the confidence interval was high is that the recognition rates of most of the words constituting subset 2 were above 99.0%, and the prediction performance for these words was almost perfect. However, the percentage of words fell within the confidence interval was 42.0% for those words for which the recognition rate was less than 99.0%. This result means that the estimation precision is low under this condition. In this experiment, we tried to apply the simple method of DP matching to different phoneme length word-pairs. Therefore, further improvement is needed.

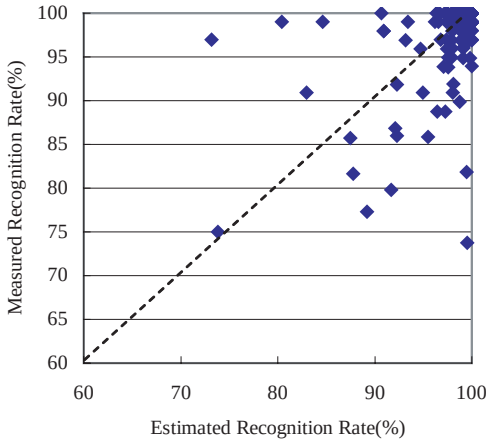

Fig. 5 Comparison of predicted and measured word-pair recognition rate (subset 2).

5. Summary

This paper has described a method of estimating word recognition performance without speech data.

First, we showed that word-pair recognition is an efficient way of evaluating the word recognition performance by word-pair recognition experiments. Second, an estimation method was proposed, which was based on the log-likelihood difference distribution. The distribution was calculated by the measured log-likelihood difference distribution of the phoneme pairs, after which the word-pair recognition rate could be calculated. Finally, experiments showed that the correlation coefficient between the predicted and the measured recognition rate was 0.87 when the word lengths were equal.

Henceforth, we will evaluate the estimation performance for open speech data. Furthermore, more research is needed to improve the predicted precision when the phoneme lengths are different.

References

1) Shimizu, T., Kojima, S., Wakita, T. and Hongo, T. : "Evaluation of Spoken Dialog Systems for a Vehicle", Trans. of IPSJ, 2000-SLP-32-16(2000), 87-92 (in Japanese)
2) Juang, B. -H. and Rabiner, L. R. : "A Probabilistic Distance Measure for Hidden Markov Models", AT&T Tech. J., **64**-2(1985), 391-408
3) Tanaka, K. and Kojima, H. : "Estimation of a Degree of Speech Recognition Difficulty for Word Sets - An Application of Between-Word Distance Calculation in a Symbolic Domain - ", J. Acoust. Soc. Jpn. (E), **19**-5(1998), 339-347
4) Abe, K., Hatano, K. and Fukumura, T. : "Performance Evaluation of Character Recognition System with Dictionary", IECEJ Trans., **52**-C-6(1969), 305-312, (in Japanese)
5) Nakagawa, S. : "Relationship between Phoneme Recognition Performance and Word Recognition Rate", IPSJ Trans., **22**-5(1981), 488-496 (in Japanese)
6) Nakamura, A. : "Predicting Speech Recognition Performance", Proc. of Eurospeech'97, (1997), 1567-1570
7) Yamashita, Y. : "Prediction of Keyword Spotting

Accuracy Based on Simulation", Proc. of Eurospeech'99, (1999), 1235-1238
8) Young, S., Jansen, J., Odell, J., Ollason, D. and Woodland, P. : The htk book (for htk version 2. 2), Entropic, (1999)
10) Kawahara, T., Lee, A., et al. : "Japanese Dictation Toolkit - 1997 version - ", J. Acout. Soc. Jpn. (E), **20**-3(1999), 233-239

(Report received on Dec. 23, 2003)

Ryuta Terashima
Year of birth : 1968
Division : ITS Lab. II
Research fields : Spoken Dialogue System, Speech Recognition
Academic society : Inf. Process. Soc. Jpn., Acoust. Soc. Jpn.