PREDICTING SPEECH RECOGNITION PERFORMANCE

Atsushi Nakamura
ATR Interpreting Telecommunications Research Laboratories.
2-2 Hikaridai, Seika-Cho, Soraku-Gun, Kyoto, 619-02 JAPAN
Tel. +81 774 95 1301, FAX: +81 774 95 1308, E-mail: atsushi@itl.atr.co.jp

ABSTRACT

Predicting speech recognition performance in place of expensive recognition experiments is a very useful approach for the research and development of speech recognition systems. In this paper, we propose a method to predict speech recognition performance when using new test data and/or a new acoustic model. Performance prediction tests showed that the proposed method can accurately predict recognition performance, thus saving a large amount of computer resources.

1. INTRODUCTION

In building a speech recognition system, many experiments are required to verify the effectiveness of new ideas. This, however, consumes too much time and a large amount of computing resources (CPU, memory). A performance prediction technique could save such time and resources and enable us to develop several systems more efficiently.

In [1], prediction of the word spotting performance when using a new test data of similar quality was investigated. This type of prediction is, however, not very useful because the assumption of similar quality is impractical for real-world speech.

In this paper, we address the issue of predicting the performance of a hidden Markov model (HMM) based continuous speech recognizer under two conditions:

- either new test data of unrestricted quality or a new acoustic model is given;
- both new test data and a new acoustic model are given.

In order to solve this problem, we must find a measure that can be calculated with much lighter processing than speech recognition and that has a strong correlation to recognition performance. Such a measure would also be useful in clarifying the mechanism of recognition error occurrence.

In Section 2, the basic ideas behind our prediction techniques are given. Section 3 shows the prediction procedure. In section 4, prediction tests carried out using spontaneously spoken dialogue speech data are explained, and the results show the accuracy of the proposed method.

2. BASIC IDEAS

2.1 Performance prediction for new test data and/or a new acoustic model

For each case, a new mismatch condition between test data and acoustic model is given. Therefore, if we can know beforehand how a recognizer performs for several degrees of the mismatch and if we can measure the degree of mismatch for each pair of new test data and acoustic model, we can approximate the performance of recognizer "X" for test data D and an acoustic model M as

$$\hat{\rho} = R\left(\delta(M,D)\,;\,X\right),$$

where

$R(\Delta;X)$: the performance of recognizer "X" for a degree of mismatch Δ, and

$\delta(M,D)$: the degree of mismatch between acoustic model M and test data D.

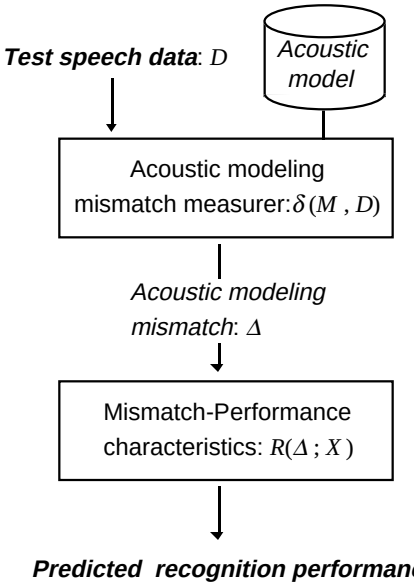

Fig. 1 Performance prediction for a new mismatch condition

Consequently, the performance prediction problem is reduced to the problem of how to define $R(\Delta ; X)$ and $\delta(M , D)$ (Fig. 1).

2.2 Estimating recognition performance for several degrees of acoustic modeling mismatch

In this paper, we try to estimate by simulation the performance of a recognizer for several degrees of acoustic modeling mismatch.

We can artificially generate a set of samples that ideally match the acoustic model by the Monte Carlo method; this is done by using output distributions of the acoustic model as generator distributions. We can also generate a set of samples with a certain degree of mismatch by substituting some generator distributions with alternative distributions during the Monte Carlo process. Then, by gradually varying the probability of the substitution, we can generate samples with gradually varying degrees of mismatch (Fig. 2). The Mismatch-Performance curve given by carrying out a recognition test using such artificial samples approximately shows the behavior of $R(\Delta ; X)$. Using a well-trained HMM and a phonetically rich text corpus (label files), the plotted curve can be generally used for any acoustic model and for any test data.

2.3 Measuring degree of mismatch in acoustic modeling

The degree of mismatch between test data and an acoustic model is measured by the frame-level error rate. The frame-level error rate is defined as the substitution error rate at the frame level, and can be easily calculated by applying Viterbi alignment using the label sequence for the test data (Fig. 3). That is:

$$\delta(M,D) = \frac{\sum_i E(P(o_i|g_i) - \max_{\gamma \in \Gamma} \{P(o_i|\gamma)\})}{N}$$

$$E(\varepsilon) = \begin{cases} 0 & (\varepsilon \geq 0) \\ 1 & (\varepsilon < 0) \end{cases},$$

where

o_i : the observation at frame #i in data D,

g_i : the output distribution assigned to frame #i by Viterbi alignment between data D and acoustic model M,

Γ : the set of output distributions in acoustic model M,

N : the number of frames in data D, and

$P(o|g)$: the local likelihood of observation o for output distribution g.

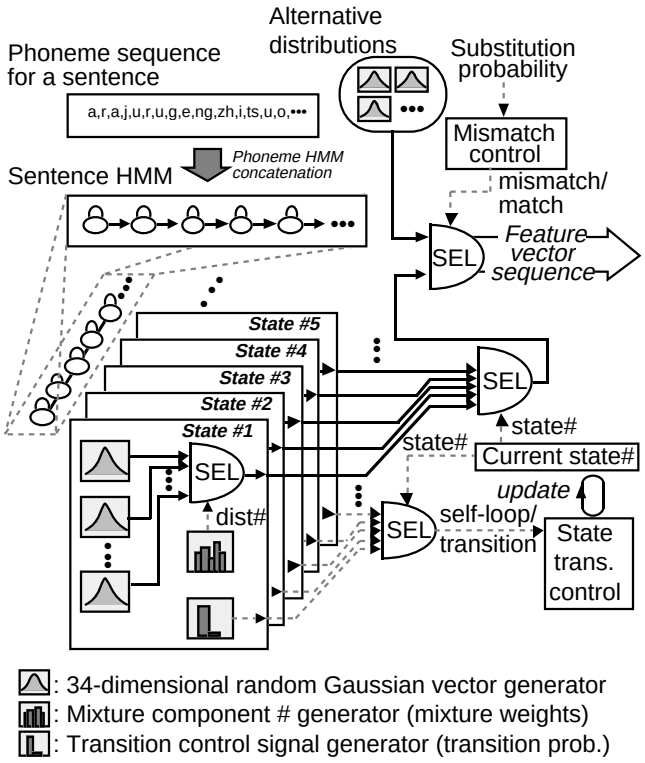

Fig. 2 Mismatched speech data generation

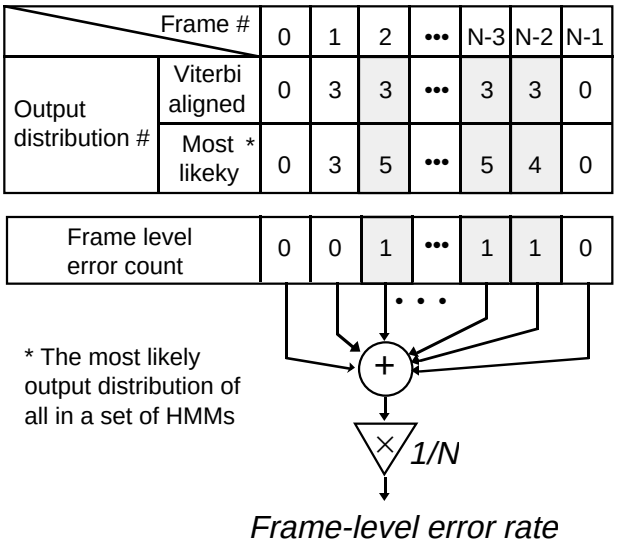

Fig. 3 Frame-level error rate calculation

3. PERFORMANCE PREDICTION PROCEDURE

The performance prediction procedure based on the above ideas is summarized as follows.

[Preliminary step]
Plot Mismatch-Performance curve for the target speech recognizer by the Monte Carlo method. Here, the mismatch is measured by the frame-level error rate.

[Step-1] Measure the frame-level error rate as the degree of mismatch between the given test data and acoustic model by Viterbi alignment.

[Step-2] Read the figure of performance for the degree of mismatch from the Mismatch-Performance curve.

In this procedure, The preliminary step can be executed beforehand and is required only once. For every new data-model pair, only Step-1 and Step-2 have to be executed. Because Viterbi alignment needs no lexicon and is associated with no growth of the search space, this method requires a much smaller amount of computer resources than speech recognition experiments. The effect of saving time is not so remarkable in this procedure since mismatch measuring based on Viterbi alignment takes some time. Nevertheless, it is still an advantage that the required time is generally constant because it does not depend on the growth of the search space.

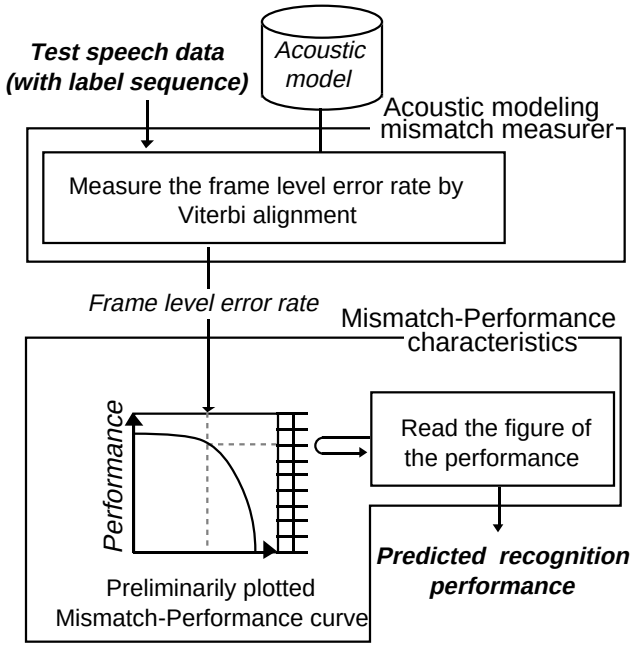

Fig. 4 Performance prediction procedure

4. EXPERIMENTS

4.1 Target speech recognizers

The performance of the following two types of speech recognizers were predicted.

- A Japanese continuous phoneme recognizer with syllabic rule constraints (Recog-1)
- A Japanese continuous speech recognizer featuring class bigram constraints and word graph outputs with a lexicon of 1,200 words [2] (Recog-2)

Here, the class bigram was generated by a variable-order N-gram procedure [3]. For both of the speech recognizers, state-shared, context dependent and speaker independent HMMs with diagonal covariances (HMnet [4]) were used as acoustic models. The details of recognition conditions are summarized in Table 1.

Table 1 Speech recognition conditions

Acoustic analysis	• Sampled at 12kHz with 16 bit • Preemphasized with $1-0.97z^{-1}$ • 20ms Hamming window • 10ms frame shift • 16th order LPC cepstra + Δcepstra + power + Δpower	
Acoustic model	• 401-state speaker independent HMnet 400 states for allophone HMMs 1 state for a silence HMM 10 mixture/state or 5 mixture/state	
Language model	Recog-1	• Japanese syllabic rules
	Recog-2	• 1,200 words • Variable order class N-gram Number of classes: 500

4.2 Plotting Mismatch-Performance curve

Mismatch-Performance curves were plotted using label sequences of 50 Japanese phonetically-balanced sentences with the mismatch measured by the frame-level error rate. The alternative distributions were chosen from other distributions in the HMnet, taking into account the frame-level error tendency. Even if no distributions were substituted in data generation process, there were some errors in generated data because of randomness. Therefore, the curves were plotted with the mismatch (error rate) from approximately 0.1.

4.3 Prediction tests

Performance prediction tests were carried out for Recog-1 and Recog-2, and the predicted performances were compared with the true performances for several combinations of test data and acoustic models. For these tests, all utterances were from a spontaneous speech dialogue corpus in the "Travel arrangement" (e.g., hotel reservation) task domain [5].

Fig. 5 and Fig. 6 show the results for Recog-1 (54 different speaker's speech) and Recog-2 (15 different speaker's speech), respectively. In both cases, the predicted performance by the proposed method well

fitted the upper-limit of the true performance at each degree of mismatch between the test data and acoustic model. From the viewpoint of system development, it is very useful to know beforehand the upper-limit of the recognition performance with much smaller amount of computer resources than speech recognition experiments. Our prediction method required thirty or forty times smaller amount of memories than the Recog-2 under a reasonable setting required.

We also found that, for test data with a low S/N ratio, the true performances tended to be lower than the predicted performances. This implies that the proposed method is more widely applicable when we improve the measure of mismatch (e.g., multidimensional measure).

Furthermore, the prediction results proved the strong correlation between the frame-level error rate and the recognition rate, which supports the hypothesis that the frame-level error is one of the major factors in recognition error occurrence.

5. CONCLUSIONS

A method of predicting the speech recognition performance when using new test data and/or a new acoustic model was proposed. Performance prediction tests showed that the proposed method can easily and accurately predict recognition performance. Since this method is based on mismatch measuring by Viterbi alignment, which needs no lexicon and is associated with no growth of the search space, we can save much of computing resources. And the strong correlation between the frame-level error rate which was defined in this paper and the recognition rate supports the hypothesis that the frame-level error is one of the major factors in recognition error occurrence.

As future work, we are planning to improve the mismatch measure to enhance accuracy.

REFERENCES

[1] Siu et al., "Predicting word spotting performance," Proc. of ICSLP '94, pp. 2195-2198 (1994)

[2] Shimizu et al., "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," Proc. of ICASSP '96, pp.145-148 (1996)

[3] Masataki et al., "Variable order N-gram generation by word class splitting and consecutive word grouping," Proc. of ICASSP '96, pp.188-191 (1996)

[4] Takami et al., "A successive state splitting algorithm for efficient allophone modeling," Proc. of ICASSP 92, pp. 573-576 (1992)

[5] Nakamura et al., "Japanese speech databases for robust speech recognition," Proc. of ICSLP 96, pp. 2199-2202 (1996)

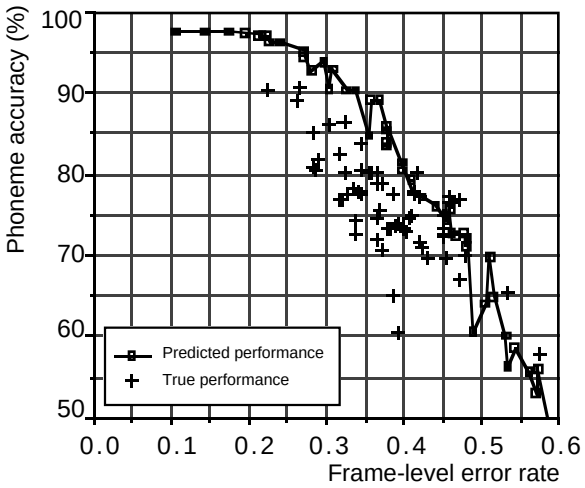

Fig. 5 Prediction result (Recog-1)

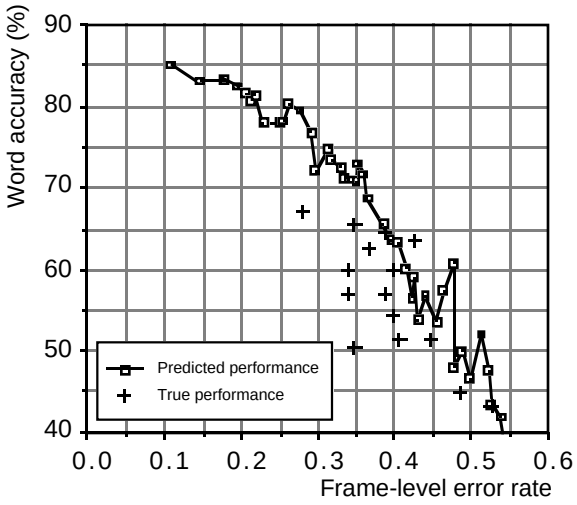

Fig. 6 Prediction result (Recog-2)