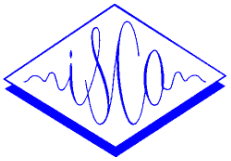

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

FRAME-PERIOD ADAPTATION FOR SPEAKING RATE ROBUST SPEECH RECOGNITION

Satoru TSUGE† *Toshiaki FUKADA*‡ *Kenji KITA*†

†Tokushima University
‡Canon Inc.
E-mail: {tsuge, kita}@is.tokushima-u.ac.jp, fukada@cis.canon.co.jp

ABSTRACT

This paper describes a frame-period adaptation method for speaking rate robust speech recognition. The proposed method determines an appropriate frame-period for each phrase by measuring its speaking rate or computing the acoustic likelihood with a set of frame-periods. Experimental results on spontaneous speech recognition show that the proposed method is effective for slower utterance. Actually, we can get about a 15% error reduction in error rate for slower utterance by using the likelihood based frame-period determination.

1. INTRODUCTION

All speech recognition systems include an initial signal processing front end that converts a noisy and/or degraded speech waveform into useful feature parameters for further processing. Specifically, the front end is required to extract important feature parameters from the speech waveform that are relatively insensitive to speaker, noise, channel, speaking rate variation[1][2]. These feature parameters provide a reasonable recognition performance if both the training and the testing environment are same. However, once mismatched conditions exist between training and testing data, due typically to background noise, channel distortion or speaking rate, the recognition performance drops severely.

We especially concert the speaking rate which is variant of speech signals. Not only inter-speaker variability but also intra speaker variability cause the difference in speaking rate, because speakers tend to vary the speaking rate in different situations. It is well known that the performance of the speech recognition systems degrades when the speaking rate is much different than the average speaking rate[3][4]. Several techniques have been proposed for relaxing speaking rate variability (e.g., adapting the acoustic model state-transition probability, language weight and insertion penalties[5], and reestimating the acoustic model using maximum likelihood estimation and maximum aposteriori estimation[6]). These techniques require to retrain and/or adapt acoustic models. Another method can be considered as a feature-based approach which speaking rate variability is relaxed in the acoustic analysis.

In this paper, we propose a frame-period adaptation method for speaking rate robust speech recognition and investigate the effectiveness of this method through the continuous speech recognition experiment.

In the following section, we explore a relationship between the speech recognition performance and the speaking rate. Section **3** describes a frame-period adaptation method for variable speaking rate. In Section **4**, we show the recognition result on a Japanese spontaneous speech, the effectiveness of the proposed method.

2. SPEAKING RATE AND RECOGNITION PERFORMANCE

First, we present a relationship between the speaking rate and the speech recognition performance. In this paper, the speaking rate, which is defined the number of morae per second (mora/sec), is computed upon each pause unit. Morae is a basic series of consonant-vowel syllables (CV-syllables) in Japanese. And a unit of mora/sec is widely used for the duration control in many Japanese text-to-speech systems.

To investigate the recognition performance influenced by speaking rate, we conduct continuous speech recognition experiments on the ATR spontaneous speech database[7]. The speech dialogues uttered by 86 male speakers are used for the evaluation. Average speaking rate of these dialogues is 9.54 mora/sec. The speech recognition is performed by a one-pass Viterbi algorithm under the phonotactic constraints of Japanese language expressed as phoneme-pair grammar[8]. Figure 1 shows the number of pause units with different speaking rate and their recognition results as phoneme accuracy.

It can be seen from this figure that the recognition performance degrades in compliance with difference of average speaking rate, i.e., it is difficult to recognize much slower and faster speech. Similar results were reported in [3] and [4]. Also, the fact that slower and faster speech samples are limited, indicates the difficulty of a construction of the speaking rate dependent acoustic models. Now, we have to develop a normalization or adaptation scheme for relaxing the speaking rate variability.

Note that even if the recognition performances of the slower and/or faster speech are improved, slight improvements in overall performance can be expected, because they are relatively much small compared to the medium speed speech. However, we believe that improving these bad performances is much important in real world applications than improving the average performance.

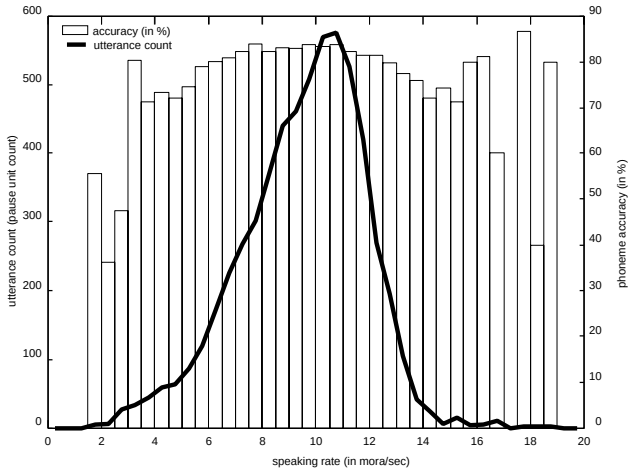

Figure 1: Speaking rate versus recognition performances and distribution of pause units.

3. FRAME-PERIOD ADAPTATION

In this section, we describe the speaking rate adaptation method on how to compensate the speaking rate mismatch between the utterance. This method is how to change a frame-period of the test utterance, namely faster utterance is changed to short frame-period, on the contrary, slower utterance, to longer frame-period. We call this method "frame-period adaptation". In frame-period adaptation, we have to determine a frame-period of each test utterance. We propose here the following two methods for frame-period determination:

1. frame-period adaptation based on *speaking rate*
2. frame-period adaptation based on *likelihood*.

These methods are described in the following sections.

3.1. frame-period adaptation based on speaking rate

We describe how to adapt a frame-period using speaking rate, we use mora/sec for the speaking rate in this paper.

For calculating the speaking rate for each utterance, we use transcription files. The transcription files include start and end time of the utterance, and the transcription of the utterance with Roman alphabet character. Using these transcription files, we can compute a mora/sec for each pause unit.

Using speaking rate computed from a transcription file, the adapted frame-period of each utterance ($\overline{FP}$) is calculated as follows:

$$\overline{FP} = FP * MS_t/MS_s, \tag{1}$$

where FP is the standard frame-period of training acoustic model, MS_t and MS_s are the average speaking rate (mora/sec) of the training set and the speaking rate of adapted utterance, respectively. In this paper, we use 10 msec ($FP = 10$) for the frame-period at training acoustic model.

When the frame-period is longer than analysis window length, adequate feature extraction can not be performed. On the contrary, when the frame-period is much smaller, computational cost will increase. Therefore, we restrict the frame-period ranges as

$$FP - 4 \leq \overline{FP} \leq FP + 4. \tag{2}$$

3.2. frame-period adaptation based on likelihood

Generally, the duration of each phoneme differ even if speaking rate is same. Therefore, we need to consider the kinds of phonemes for calculating the strict speaking rate. If we use duration model which are employed in text-to-speech systems, we can compute more strict speaking rate considering the kinds of phonemes. For simplicity, we here introduce a likelihood based method which determines the speaking rate implicitly.

This method uses an acoustic model trained with the standard frame-period. Using this acoustic model, the adapted frame-period and the recognition result are obtained by the following steps.

Step 1 Decode the utterance analyzed with the standard frame-period.

Step 2 Compute the acoustic likelihoods with a set of frame-period against the recognition result obtained by **Step 1**.

Step 3 Choose a frame-period which gives the highest acoustic likelihood as the adapted frame-period.

Step 4 Decode the utterance analyzed with the adapted frame-period again.

In **Step 2**, shorter frame-period generally gives higher likelihood because the number of frames is different among frame-periods. Therefore, we employ the following normalization method:

$$\overline{FP} = \underset{FP}{\operatorname{argmax}}(AM - GM), \tag{3}$$

where $\overline{FP}$ is the adapted frame-period, FP indicates a set of the selected frame-periods (i.e., 6, 7, ..., 13, 14 msec in this paper). AM and GM indicate the likelihood calculated from trained acoustic model and a generic model, respectively. As for a generic model, we use a GMM with 128 mixture components trained by the feature parameters with the standard frame-period.

4. EXPERIMENT

To investigate the effectiveness of the proposed method, we conducted the continuous phoneme recognition experiment using Japanese spontaneous speech database[7].

4.1. Conditions

A total of 86 male speakers' dialogues (about 144 minute) sampled at 16 kHz were used for the acoustic model training. For the test set, other 8 male speakers' dialogues (about 21 minute) were used. Average speaking rate of the training and the test set were 9.54 mora/sec and 7.84

Table 1: Preprocessing conditions.

sampling rate	16 kHz
preemphasis	0.98
frame length	20 msec
window type	Hamming
MFCC order	12
filter bank order	16

Table 2: Recognition results for the test set.

method	phoneme accuracy (%)
conventional	70.23
speaking rate	70.91
likelihood	70.75

mora/sec, respectively. 12-dimensional MFCCs, log power, and their first derivatives (i.e., 26 dimensions in total), which were analyzed on the condition described in Table 1, were used as feature vectors. The standard frame-period was set to 10 msec for the training. The frame-period for the test was determined by the proposed method among nine kinds of frame-period (i.e., 6, 7, . . . , 13, 14 msec).

Shared-state context dependent HMMs with five Gaussian mixture components per state were trained. The total number of states was set to 800. Using this acoustic model, we evaluated the proposed method on each pause units. As a first step of frame-period adaptation based on likelihood, for ignoring the influence of recognition error on standard frame-period, we used true sequence, i.e. contents of test utterance, instead of the recognition results of **Step 1** (see section 3.2).

Recognition for all conditions was performed using one-pass Viterbi algorithm with the phonotactic constraints of Japanese language expressed as phoneme-pair grammar[8]. Recognition results were given as phoneme accuracy. We used ATRSPREC[9] as acoustic modeling and recognition tools.

4.2. Results and discussion

The recognition results of the proposed method were shown in Table 2. For conventional method, we showed recognition result without proposed method, i.e., recognition result of standard frame-period, in this table. Unfortunately, we could get slight improvement comparing to the recognition performance of the conventional method.

In the following section, we would discuss about the speech recognition results in Table 2. We investigated the relationship between the recognition result of the proposed method and its speaking rate. Furthermore, we investigated whether the frame-period determined by the likelihood based method could capture its speaking rate in terms of Eq. (1).

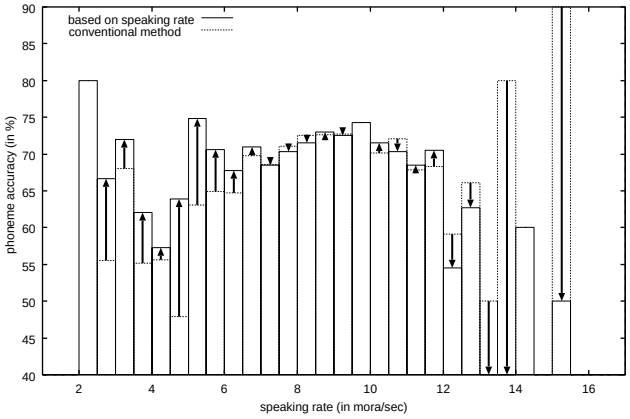

Figure 2: Comparison of recognition results (based on speaking rate v.s. conventional method). The arrow (→) indicates improvement or degradation of the recognition performance by the speaking rate based method.

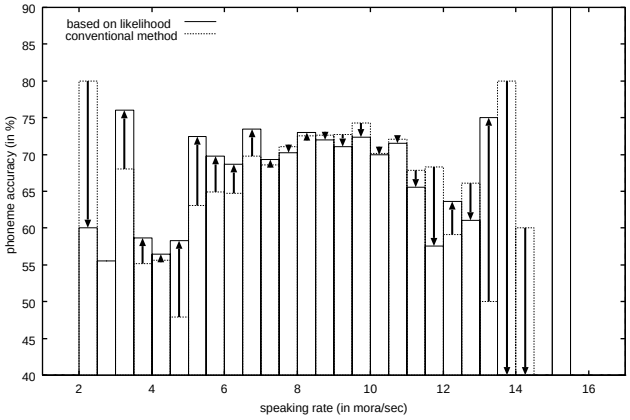

Figure 3: Comparison of recognition results (based on likelihood v.s. conventional method). The arrow (→) indicates improvement or degradation of the recognition performance by the likelihood based method.

4.2.1. relationship speaking rate and speech recognition result

Figures 2 and 3 represented the recognition performances of each speaking rate by the speaking rate based method and the likelihood based method, respectively. We computed the average recognition performances of each speaking rate of Table 2, these performances showed in these figures.

We could see from these figures that the proposed method was significantly improved the recognition performance than these of the conventional method at slower utterance. We investigated the recognition performances at slower utterance less than 7 mora/sec, these utterance were 264 pause units (about 32.2% of all test utterance). The recognition performance of these utterance was shown in

Table 3: Recognition results for the test set at speaking rate less than 7 mora/sec.

method	phoneme accuracy (%)
conventional	65.63
speaking rate	69.61
likelihood	70.76

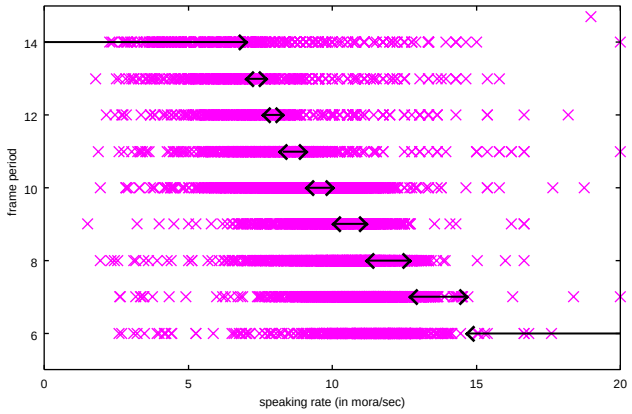

Figure 4: Relationships between selected frame-period and speaking rate. The times (×) indicates the frame-period selected by the likelihood based adaptation method. The arrow (↔) indicates a range of the frame-periods calculated by Eq. (1).

Table 3. We observed the relative improvements in the error rate by 11.57%(65.63%→69.61%) using the speaking rate based method and 14.93%(65.63→70.76%) using the likelihood based method, respectively.

4.2.2. relationship between the frame-periods determined by the proposed method

Here, we investigated whether the frame-period determined by the likelihood based method could capture its speaking rate in terms of Eq. (1). Using the likelihood based frame-period adaptation method described in section 3.2, we estimated frame-periods for the training set. The frame-periods selected by likelihood based method were shown in figure 4.

We could see from this figure that the selected frame-periods roughly captured their speaking rates. Consequently, we could expect the likelihood based method was useful for the implicit frame-period selection.

5. CONCLUSIONS

In this paper, we have proposed the frame-period adaptation method for speaking rate robust speech recognition. The adaptive frame-period is determined for each pause unit based on speaking rate or acoustic likelihood.

We could observe significant improvements in the phoneme accuracy at less than 7 mora/sec by using the proposed frame-period adaptation method. Actually, we could get the improvements of error rate 11.57% using adaptation based on speaking rate and 14.93% using adaption based on likelihood, respectively.

It is well known that people tend to speak slowly in real word environments when recognition systems are used. So, we believe that the proposed method is useful at real world environments.

In our experiments, we assume that the speaking rate is constant during the utterance. As a further study, we will have to develop more sophisticated techniques for capturing the speaking rate change during the utterance. Also, we will use the proposed adaptation method for acoustic modeling.

6. ACKNOWLEDGMENT

We considered basic idea of this paper from the time when we worked in ATR Interpreting Telecommunications Research Laboratories.

We would like to thank Dr. Seiichi Yamamoto, President of ATR Spoken Language Translation Research Laboratories, and Dr. Sagisaka, head of Department 2, for giving us the opportunity to carry out this study. We would also like to thank all members in Department 1 at ATR Interpreting Telecommunications Research Laboratories for their helpful discussions.

7. REFERENCES

[1] J. Picone. Signal modeling techniques in speech recognition. In *Proceedings of the IEEE*, volume 81, pages 1215–1247, 1993.

[2] Ch. Jankowski, H. Vo, and R. Lippmann. A comparison of signal processing front ends for automatic word recognition. In *IEEE Trans. Speech and Audio Processing*, volume 3, pages 286–293, July 1995.

[3] D. Pallet, J. Fiscus, W. Fisher, J. Garofolo, B. Lund, and M. Pryzbocki. 1993 benchmark tests for the ARPA spoken language program. In *Proc. ARPA workshop*, pages 51–73, 1993.

[4] M. Siegler and R. Stern. On the effects of speech rate in large vocabulary speech recognition system. In *Proc. ICASSP*, pages 612–615, 1995.

[5] F. Martinez, D. Tapias, J. Alvarez, and P. Leon. Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In *Proc. EuroSpeech*, pages 469–472, 1997.

[6] T. Pfau and G. Ruske. Creating Hidden Markov models for fast speech. *Proc. ICSLP*, 1998.

[7] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura, and Y. Sagisaka. Japanese speech databases for robust speech recognition. In *Proc. ICSLP*, pages 2199–2202, Philadelphia, 1996.

[8] H. Ohwaki, H. Singer, J. Takami, and A. Kurematsu. Phonetic typewriter using phonotactic constraints. *Technical report of IEICE*, SP93-113:71–78, December 1993. (in Japanese).

[9] H. Yamamoto, H. Singer, B. Reaves, and Y. Sagisaka. Control and structure of recognition subsystem in the ATR-MATRIX Japanese-English speech translation system. *In Proc. Acoust. Soc. Jap.*, pages 161–162, 1998. (in Japanese).