SPEAKING RATE COMPENSATION BASED ON LIKELIHOOD CRITERION IN ACOUSTIC MODEL TRAINING AND DECODING

Kozo Okuda[†] , *Tatsuya Kawahara*[†‡] *and Satoshi Nakamura*[†]

[†] ATR Spoken Language Translation Research Laboratories., Kyoto 619-0288, Japan.
[‡] School of Informatics, Kyoto University., Kyoto 606-8501, Japan.
[†]{kokuda,nakamura}@slt.atr.co.jp, [‡]kawahara@kuis.kyoto-u.ac.jp

ABSTRACT

In this paper, we propose a speaking rate compensation method using frame period and frame length adaptation. Our method decodes an input utterance using several sets of frame period and frame length parameters for speech analysis. Then, this method selects the best set with the highest score which consists of the acoustic likelihood normalized by frame period, language likelihood and insertion penalty. Furthermore, we apply this approach to the training of the acoustic model. We calculate the acoustic likelihood for each frame period and frame length using Viterbi alignment and select the best one for each training utterance. The proposed speaking rate compensation applied to both the acoustic model creation process and decoding process resulted in accuracy improvement of 2.9% (absolute) for spontaneous lecture speech recognition task.

1. INTRODUCTION

The performance of speech recognition system has been improved by using statistical approaches and large speech databases. However, the performance is not sufficient enough for automatic speech transcription and translation of real spontaneous speech. For the research of spontaneous speech recognition, a large scale spontaneous speech corpus collection was started under the Science and Technology Agency Priority Program in Japan in 1999[1]. This corpus is called "The Corpus of Spontaneous Japanese (CSJ)" and includes a large amount of lecture speech.

In lecture speech recognition, the influence of speaking rate variation to the performance is significant[2]. In fast utterances, the occurrence of deletion and substitution error increases and the performance degrades. Moreover, speaking rate changes not only between speakers but also between utterances. For this reason, speaking rate normalization or compensation is an important issue for spontaneous speech recognition.

In previous research, several speaking rate normalization or compensation methods have been proposed[3][4]. These methods estimate the phone boundaries or speaking rate and normalize the speaking rate by changing the analysis frame or acoustic model according to the estimated speaking rate. These methods, however, can not improve the performance sufficiently. One of the problem is that the speaking rate estimation and recognition process are separated, and the overall performance is highly depends on the precision of the former process.

In this paper, we propose a speaking rate compensation method using frame period and length adaptation. This method selects the best set of frame period and length for each utterance based on likelihood criterion in decoding process. Our method does not adopt the prior speaking rate estimation, and the frame period and length selection process is consistent with the decoding process. Moreover, we explore the application of the speaking rate compensation method to not only decoding phase but also acoustic model training phase, so that matched model is selectively applied in decoding.

2. TASK AND BASELINE SYSTEM

For test data, we use the standard test set of the CSJ[5] shown in Table 1. This set includes 10 male speakers.

For the baseline acoustic model, a 25-dimensional feature vector (12-dimensional mel-cepstral coefficients, 12 dimensional first-order derivatives of mel-cepstral coefficients and 1-dimensional first-order derivative of logarithmic power) is computed with a 10 msec frame period and 20 msec frame length. The number of phones is 26, and all phones are modeled with a left-to-right HMM with three states (no state-skip). We trained gender-dependent shared-state HMMs (1,400 states in total) with ten Gaussian mixture components per state[6]. The baseline model was trained using speech data of 200 lectures (about 34 hours) of the CSJ.

For the language model, we use a forward word bigram and backward word trigram created at Kyoto University in Japan and distributed with the CSJ. These models are trained with lecture transcription data of the CSJ. The size of the lexicon is 19 K words. For the decoder, we use the Julius[7].

3. RELATIONSHIP BETWEEN SPEAKING RATE AND ACCURACY

3.1. Recognition result using baseline acoustic model

To investigate a relationship between speaking rate and recognition accuracy, we conducted a recognition experiment with the baseline acoustic model. Figure 1 shows the word error rate and average speaking rate for each speaker. The average speaking rate, which is defined as the number of morae per second, is computed upon each pause unit by Viterbi alignment. Morae are basic units of consonant-vowel syllable (CV-syllables) in Japanese.

Table 1: *Test set of CSJ*

Speaker ID	time	#words
AS22	28min	6127
AS23	30min	4302
AS97	12min	2486
PS25	27min	5305
JL01	57min	9858
NL07	15min	2161
SG05	23min	4467
KK05	42min	6557
YG01	14min	2764
YG05	15min	2939

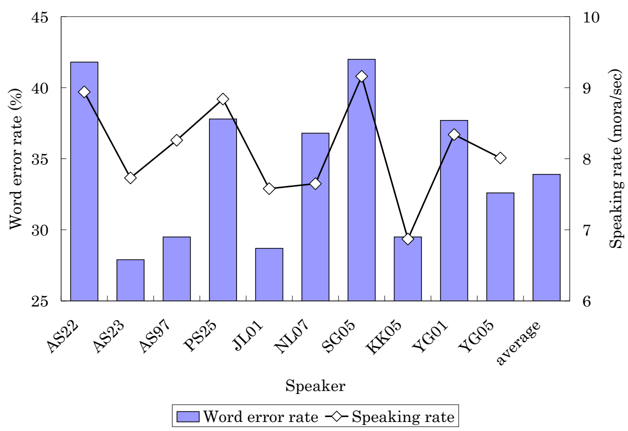

Figure 1: *Word error rate with baseline acoustic model and average speaking rate for each speaker*

The average word error rate is 33.7% and the range of word error rate varies highly among speakers. The recognition performance is low for high speaking-rate speakers. The correlation coefficient is 0.78 and there is a strong correlation between word error rate and speaking rate.

Figure 2 shows a distribution of word error rate and the occurrence of the phone segments with a duration of less than 30 msec for each speaker. The correlation coefficient is roughly 0.83 and the correlation between word error rate and speaking rate is stronger. From these results, it is confirmed that speaking-rate variation is a significant problem.

3.2. Effect of changing frame period and frame length

In fast uttered speech, use of a short frame period parameters for speech analysis reduces the mismatch of the acoustic model as well as the mismatch of the delta-parameter. In this work, we also change the frame length to adapt to the fast speech segments. From the preliminary experiment, we observed that the effect of the change of frame length depends on the frame period. Figure 3 shows the result of the recognition experiment using sets of frame period and length (10 msec, 20 msec) , (9 msec, 18 msec) and (8 msec, 16 msec). The figure shows that use of shorter frame period and length improves the performance for higher speaking-rate speakers and use of the longer frame period and length

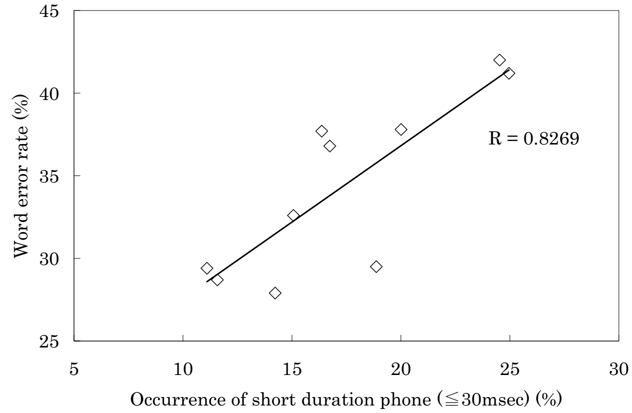

Figure 2: *Distribution of the word error rate and the occurrence of short phone (less than 30 msec)*

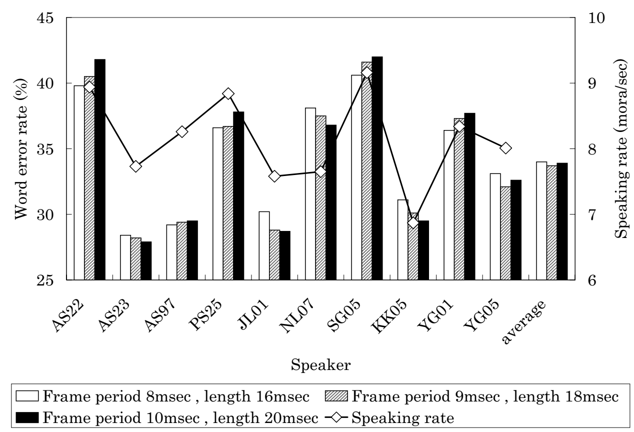

Figure 3: *Word error rate using each set of frame period and length*

is effective for slower speakers. From these results, changing the frame period and length achieves the effect of speaking rate compensation.

4. SELECTION OF FRAME PERIOD AND LENGTH BASED ON LIKELIHOOD CRITERION

In automatic speech recognition, speaking rate and the optimal set of frame period and length is not known. Although, automatic estimation of speaking rate has been studied[4], the estimation is far from perfect, and the precision significantly affects the following process of compensation and recognition. Thus, we propose a method to select the best set of frame period and length after decoding based on likelihood criterion.

Since the speaking rate changes not only among speakers but also among utterances of the same speaker, we perform the decoding and selection for each utterance.

First, we decode an input utterance using several sets of frame period and frame length parameters for speech analysis, then select the set with the highest acoustic likelihood.

Table 2: *Word error rates by frame period and length adaptation using acoustic likelihood criterion (%)*

	fixed frame period			adaptive
	8msec	9msec	10msec	frame period
AS22	39.8	40.5	41.8	**39.3**
AS23	28.4	28.2	**27.9**	28.5
AS97	29.2	29.4	29.5	**29.2**
PS25	36.6	36.7	37.8	**36.2**
JL01	30.2	28.8	**28.7**	29.5
NL07	38.1	37.5	**36.8**	37.8
SG05	40.6	41.6	42.0	**40.1**
KK05	31.1	30.1	**29.5**	30.2
YG01	36.4	37.3	37.7	**36.4**
YG05	33.1	32.1	32.6	**31.9**
average	34.0	33.7	33.9	**33.5**

Use of shorter frame period generally gives a lower acoustic likelihood because the number of frames is increased. Therefore, we normalize the acoustic likelihood by the frame period:

$$AM' = AM * \frac{frame\ period(msec)}{10} \cdots (1)$$

AM : $acoustic\ likelihood\ of\ each\ utterance$
AM' : $acoustic\ likelihood\ normalized\ by\ the\ frame\ period$

The denominator in equation (1) is 10 because we define 10 msec as the standard frame period.

The recognition results by the automatic selection based on acoustic likelihood criterion as well as the fixed frame period and length are listed in Table 2. The performance is improved for fast speakers. However, the method based on acoustic likelihood criterion does not work well for slow speakers. In LVCSR system, the recognition result is selected by using score that consists of acoustic likelihood, language likelihood and insertion penalty. Therefore, some results that have low language likelihood are chosen because of the high acoustic likelihood. Then, we select the set with the highest score that consists of acoustic likelihood, language likelihood and insertion penalty. The recognition results using acoustic and language likelihood criterion are listed in Table 3. The method improves the recognition rate by 0.8% from the baseline. With regard to the high speaking-rate speaker (AS22, PS25 and SG05), the average improvement is 2.0% (the improvements are 2.1%, 2.1% and 1.9%).

5. ACOUSTIC MODEL WITH SPEAKING RATE COMPENSATION

Next, we propose application of frame period and length adaptation to the training of the acoustic model.

In the training phase, we use only acoustic likelihood because the transcription of the training speech data is known. To select the best frame period and length, we use Viterbi alignment with the baseline acoustic model and calculate the normalized acoustic likelihood using equation (1) for each set of frame period and length. Figure 4 shows the amount of utterances for each selected frame period and length in the training data. Using the feature vectors extracted with the selected frame period and length, we created two types of acoustic models.

Table 3: *Word error rates by frame period and length adaptation using acoustic and language likelihood criterion (%)*

	fixed frame period			adaptive
	8msec	9msec	10msec	frame period
AS22	39.8	40.5	41.8	**39.6**
AS23	28.4	28.2	**27.9**	28.1
AS97	29.2	29.4	29.5	**28.8**
PS25	36.6	36.7	37.8	**35.7**
JL01	30.2	28.8	**28.7**	28.8
NL07	38.1	37.5	**36.8**	37.0
SG05	40.6	41.6	42.0	**40.1**
KK05	31.1	30.1	29.5	**29.3**
YG01	36.4	37.3	37.7	**35.8**
YG05	33.1	32.1	32.6	**31.5**
average	34.0	33.7	33.9	**33.1**

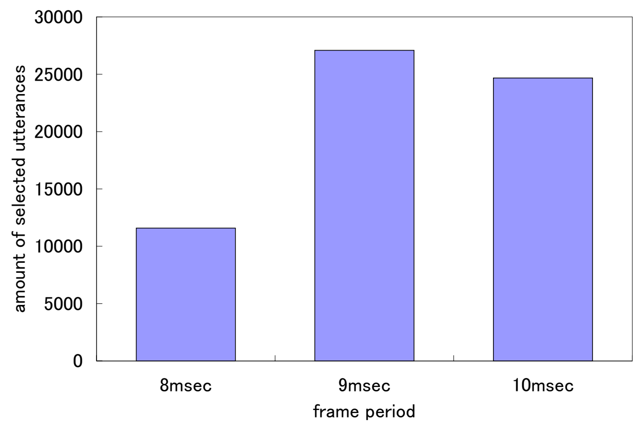

Figure 4: *Number of utterances for each frame period selected using the proposed method*

(1) Single acoustic model using all feature vectors of different sets of frame period and length (single model)

(2) Acoustic models according to a set of dedicated frame period and length (multiple models)

In the decoding process, the single model is applied by changing frame periods and lengths, while the multiple models are used for the matched frame period and length. Then, best hypothesis is chosen for either case.

Tables 4 and 5 show the results using the single model and multiple models, respectively. These results show that the proposed method effectively applies speaking rate compensation to the model creation. Moreover, frame period dependent acoustic models (multiple models) improve the performance more than single acoustic model.

In the single model, the difference in performance among the sets of frame period and length is small. Multiple mod-

Table 4: *Word error rates of each frame period using single model (%)*

	fixed frame period			adaptive
	8msec	9msec	10msec	frame period
AS22	40.5	41.9	43.1	**40.3**
AS23	28.2	28.4	28.7	**27.4**
AS97	29.0	29.2	30.9	**28.6**
PS25	**32.8**	33.7	34.7	32.8
JL01	29.5	28.9	28.9	**28.3**
NL07	34.6	34.3	34.3	**33.8**
SG05	**38.1**	39.9	41.8	38.2
KK05	30.2	29.3	29.2	**28.7**
YG01	35.7	36.4	37.1	**35.6**
YG05	33.7	**33.6**	33.7	33.9
average	32.9	33.2	33.8	**32.4**

Table 5: *Word error rates of each frame period using multiple models (%)*

	fixed frame period			adaptive
	8msec	9msec	10msec	frame period
AS22	40.3	41.3	44.0	**39.7**
AS23	28.8	27.8	31.0	**27.2**
AS97	29.8	28.2	31.7	**27.4**
PS25	36.4	31.9	35.2	**31.0**
JL01	31.6	27.6	35.5	**25.9**
NL07	37.0	33.7	38.8	**33.4**
SG05	41.0	38.6	41.4	**37.6**
KK05	34.8	30.6	30.6	**30.1**
YG01	37.3	36.5	39.5	**35.9**
YG05	37.1	33.1	33.5	**33.1**
average	35.2	32.6	36.1	**31.6**

els achieve larger improvement because they capture acoustic characteristics depending on the speaking rate. In this case (Table 5), adaptive frame period method gives better performance than any fixed frame period for all speakers.

Figure 5 shows the results using both the single model and multiple models simultaneously with the proposed decoding method. Using both models achieves the best performance of 31.0% and the improvement from the baseline is 2.9% absolute. Our proposed method assumes that the speaking rate does not change during the utterance. However, the speaking rate often changes within utterance and this change degrades the performance of multiple models. On the other hand, single model trained by all sets of frame period and length can better cope with such variation during the utterance. Therefore, using both models is most effective.

6. CONCLUSIONS

We have proposed a speaking rate compensation method using an adaptive frame period and length based on likelihood criterion. This method is effective for recognizing lecture speech in which the speaking rate is varied very much. Furthermore, we have applied the method to the training

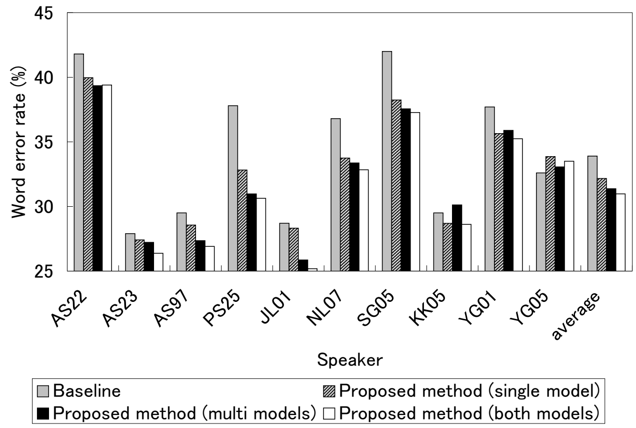

Figure 5: *Word error rates using both single model and multiple models*

phase of acoustic model and achieved more improvement. However, our proposed method does not handle the change of speaking rate during the utterance precisely, even when we use both the single model and multiple models. In addition, our proposed method compensates only for variations in the time domain. In future work, we will study handling the speaking rate variation in utterance and the change of acoustic characteristics in the spectral domain.

Acknowledgment: This research was supported in part by the Telecommunications Advancement Organization of Japan.

REFERENCES

[1] S.Furui, K.Maekawa and H.Isahara, "Toward the realization of spontaneous speech recognition – introduction of a Japanese priority program and preliminary results –," Proc.ICSLP2000, Vol.3, pp.518-521, Oct.2000.

[2] H.Nanjo, K.Kato and T.Kawahara, "Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition," Proc.EUROSPEECH2001, Vol.4, pp.2531-2534, Sep.2001.

[3] J.P.Nedel and R.M.Stern, "Duration normalization for improved recognition of spontaneous and read speech via missing feature methods," Proc.ICASSP2001,Vol.1, pp.313-316, 2001.

[4] S.Tsuge, T.Fukada and K.Kita, "Frame-period adaptation for speaking rate robust speech recognition," Proc.ICSLP2000, Vol.3, pp.718-721, 2000.

[5] T.Shinozaki, C.Hori and S.Furui, "Towards automatic transcription of spontaneous presentations," Proc.EUROSPEECH2001, Vol.1, pp.491-494. Sep.2001.

[6] M.Ostendorf and H.Singer, "HMM topology design using maximum likelihood successive state splitting," Computer Speech and Language, 11, 1, pp.17-41, 1997.

[7] A.Lee, T.Kawahara and S.Doshita, "An efficient two-pass search algorithm using word trellis index," Proc.ICSLP1998, Vol.5, pp.1831-1834, 1998.