SPEAKING-RATE DEPENDENT DECODING AND ADAPTATION FOR SPONTANEOUS LECTURE SPEECH RECOGNITION

Hiroaki Nanjo and Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{nanjo,kawahara}@kuis.kyoto-u.ac.jp

ABSTRACT

This paper addresses the problem of speaking rate in large vocabulary spontaneous speech recognition. In spontaneous lecture speech, the speaking rate is generally fast and may vary a lot within a talk. We also observed different error tendencies for fast and slow speech segments. Therefore, we first present a speaking-rate dependent decoding strategy that applies the most adequate acoustic analysis, phone models and decoding parameters according to the speaking rate. Several methods are investigated and their selective application leads to accuracy improvement. We also propose to make use of speaking-rate information in speaker adaptation, in which the different adapted models are set up for fast and slow utterances. It is confirmed that the method is more effective than normal adaptation.

1. INTRODUCTION

Under the Science and Technology Agency Priority Program in Japan (1999-2004)[1], a large scale spontaneous speech corpus is being collected and we have started extensive studies on large vocabulary spontaneous speech recognition. Our main goal is the automatic transcription of live lectures such as oral presentations in conferences.

In acoustic modeling of spontaneous speech, the speaking rate, especially fast speech, is considered as one of the most significant causes of degradation. Fast speaking often causes incomplete articulation, thus poor acoustic matching. The spectral patterns change and moreover the phone itself may disappear. There have been studies that consider the factor of speaking rate in acoustic modeling[2][3][4]. We also explored the approach[5].

On the other hand, it has been observed that there are frequent changes of speaking rate in a single lecture presentation. These changes cause significant problems when decoding with uniform models and parameters. Actually, the tendency of recognition errors is different for fast utterances and slow utterances. It is also regarded that spectral variation due to the fast speaking rate is dependent on speakers. Most of the previous studies deal with the speaking rate in speaker-independent acoustic modeling. In this paper, we present a decoding strategy depending on the current speaking rate and also a model adaptation scheme for both speakers and speaking rate.

Table 1. Test-set of lectures

	#words	duration (min.)	WER (%)
A01M0035 (AS22)	6294	28	41.1
A01M0007 (AS23)	4391	30	27.6
A01M0074 (AS97)	2508	12	27.5
A05M0031 (PS25)	5372	27	35.3
A02M0117 (JL01)	9833	57	37.3
KK99DEC005 (KK05)	6527	42	35.3
A03M0100 (NL07)	2644	15	32.0
A06M0134 (SG05)	4460	23	41.4
YG99JUN001 (YG01)	2759	14	38.5
YG99MAY005 (YG05)	3108	15	32.8
total	47896	263	35.8

2. DATABASE AND TASK

The Corpus of Spontaneous Japanese (CSJ) currently developed by the project consists of a variety of oral presentations at technical conferences and informal monologue talks on given topics.

For language model training, all transcribed data (as of June 2001) are used. There are 612 presentations and talks by distinct speakers. The text size in total is 1.48M words (=Japanese morphemes). As for acoustic model training, only male speakers are used in this work. We use 224 presentations that amount to 37.9 hour speech.

The test-set for evaluation consists of ten lecture presentations specified in Table 1. Many of them are invited lectures at technical meetings, thus relatively longer than simple paper presentations. They were given by experienced lecturers who did not prepare drafts.

0-7803-7402-9/02/$17.00 ©2002 IEEE

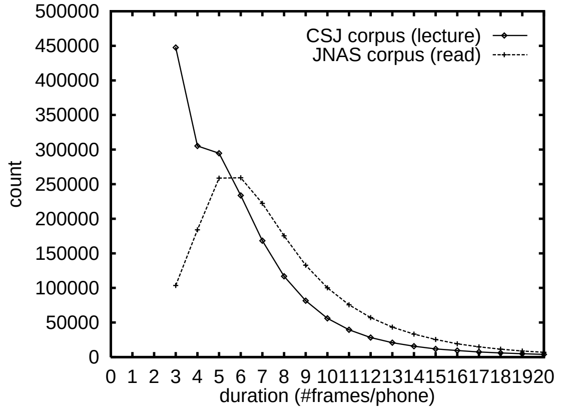

Fig. 1. Phone duration distribution of CSJ and JNAS corpus

3. BASELINE SYSTEM

Acoustic models are based on continuous density Gaussian-mixture HMM. Speech analysis is performed every 10 msec and a 25-dimensional parameter is computed (12 MFCC + 12 Δ MFCC + Δ Power).

The number of phones used is 43, and all of them are modeled with left-to-right HMM of three states and no state-skipping transitions. We trained context-dependent triphone models. Decision-tree clustering was performed to set up 2000 shared-states. We also adopt PTM (phonetic tied-mixture) modeling[6], where triphone states of the same phone share Gaussians but have different weights. Here, 129 codebooks of 128 mixture components are used.

We built a lexicon of 19158 words from the training corpus, and then made a trigram language model. It realizes coverage of 97% and test-set perplexity of 135. We use the large vocabulary speech recognition decoder Julius rev.3.1 that was developed at our laboratory[7].

The average word error rate with the baseline system is 35.8%. The rate for each speaker is listed in Table 1.

4. ANALYSIS ON SPEAKING RATE

4.1. Phone Duration Distribution

Distribution of phone duration in lecture speech (CSJ - lecture corpus: 35 hours) and read speech (JNAS - newspaper corpus: 40 hours) is plotted in Figure 1. Phone duration is estimated with Viterbi alignment. As we use three-state phone HMMs without state-skipping, the minimum duration is three frames (=30 msec). Many segments in CSJ data may have shorter duration, but are forcedly aligned with three frames. This may have caused a serious mis-match. Moreover, fast speaking rate suggests that these segments are poorly articulated and cause problems in recognition.

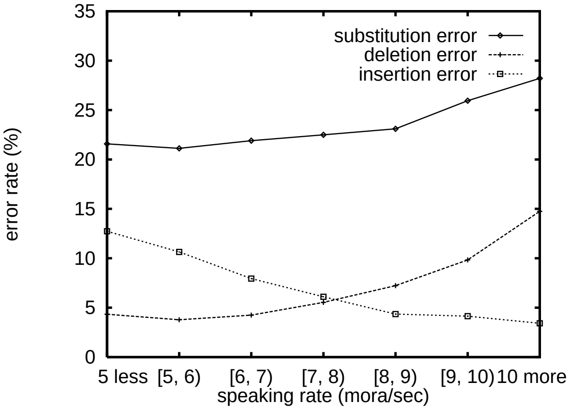

Fig. 2. Ratio of substitution, deletion and insertion errors for each speaking rate

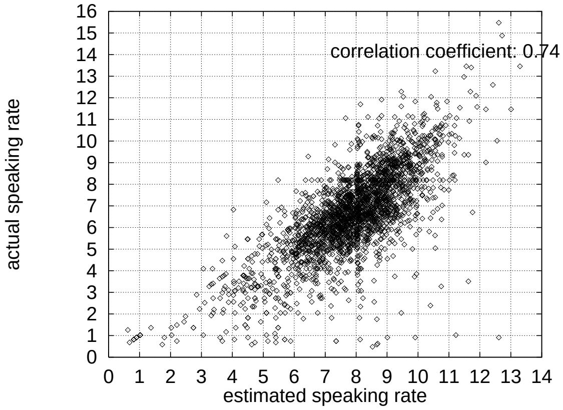

Fig. 3. Relation of actual and estimated speaking rate

4.2. Relation with Recognition Errors

The relationship between the word error rate and speaking rate is plotted for the test-set. Speaking rate is defined as the mora count divided by the utterance duration (sec). The utterances are automatically segmented from the recorded materials based on pauses in pre-processing, thus they do not necessarily match the linguistic sentences. The total number of utterances in the test-set is 2517.

In Figure 2, the breakdown of recognition errors is shown for each speaking rate. It is confirmed that faster utterances are generally harder for recognition. Moreover, we observe different tendencies in the errors according to the speaking rate. In fast utterances, substitution errors are increased as well as deletion errors. On the other hand, there are many insertion errors in slow segments.

4.3. Automatic Estimation of Speaking Rate

We also implement automatic estimation of the speaking rate. Decoding with a phonotactic syllable constraint and

Table 2. Word error rate with different decoding according to speaking rate (%)

actual speaking rate	-5	5-6	6-7	7-8	8-9	9-10	10-	average
(#utterances)	(433)	(434)	(596)	(435)	(343)	(161)	(115)	(2517)
baseline	38.7	35.5	34.1	34.1	34.7	39.9	46.4	35.8
1. analysis frame	39.7	34.5	33.5	33.1	**32.8**	**38.3**	**43.9**	34.7
2. skipping transition	37.7	34.0	**33.4**	**32.8**	34.2	39.3	45.2	34.8
3. syllable model	40.4	35.4	33.8	34.1	33.4	38.9	43.8	35.3
1.+2.	41.0	36.0	34.9	34.7	34.7	39.6	44.0	36.2
1.+3.	44.0	38.2	35.6	34.5	34.0	37.6	33.5	36.5
2.+3.	39.5	35.5	33.7	33.7	33.9	37.2	43.0	35.1
1.+2.+3.	45.7	39.3	36.6	35.1	33.8	38.0	42.1	37.1
4. insertion penalty	**35.7**	**32.7**	33.6	35.3	37.2	44.2	49.9	36.3
best one selected [oracle]	35.7	32.7	33.4	32.8	32.8	38.3	43.9	34.1
selected with estimated speaking rate	37.4	33.6	33.3	33.1	33.6	39.4	44.2	34.6

phone models is performed for mora counting. Figure 3 plots the relation between the actual and estimated speaking rate. There is high correlation between the two: the correlation coefficient is 0.74. The result verifies the feasibility of speaking rate estimation.

5. SPEAKING-RATE DEPENDENT DECODING

Based on these analyses, we propose applying different decoding methods according to the speaking rate within the multiple-pass search framework. The speaking rate in the current speech segment is estimated in the first pass. Then, the most adequate acoustic analysis, phone models and decoding parameters are applied.

Specifically, the following processings are investigated. The first three are intended for fast speech and the last one is for slow speech.

(1) Shorter frame length and shift

To cope with fast speech segments, where spectral pattern changes rapidly, the frame length and shift for spectral analysis are shortened. After preliminary experiments, we set the frame length of 20ms and the shift of 8ms from the baseline of 25ms and 10ms.

(2) State-skipping transitions in phone models

Another way to cope with fast speech is to add state-skipping transitions in phone models. It allows flexible matching with less than three frames.

(3) Use of syllable models

Since not a few phone segments may disappear, we model them with syllables of a phone sequence. We select syllables by considering both their duration and training data amount[5].

(4) Use of different insertion penalty

For slow speech segments, a larger word insertion penalty is used in order to suppress insertion errors.

These techniques and their combinations are evaluated on the test-set. They are compared with the baseline system that adopts uniform decoding. The recognition results are listed in Table 2.

For fast speech segments, all proposed methods (1,2,3) are shown to be effective and improve the overall accuracy. Combinations of them have effect on the very fast speech (9 mora/sec or faster), but result in the increase of errors in slow speech, which cancel this effect. For slow utterances, the use of a severe insertion penalty reduces errors as expected.

Then, a selective application of these methods according to the speaking rate is implemented, as specified with bold font in Table 2. The speaking rate is classified into three categories based on the experimental result. If the speaking rate is known and the best techniques are chosen accordingly (oracle case), the overall accuracy could be improved by 1.7% absolute. In actual, we estimate the speaking rate with a syllable constraint and apply the dedicated decoding methods in the second pass. This strategy achieves improvement of 1.2% absolute (last row of Table 2).

6. SPEAKING-RATE DEPENDENT ADAPTATION

Next, we introduce a speaker adaptation technique based on MLLR[8]. Since lecture speech has long duration (large data) per speaker, the unsupervised adaptation scheme works very well. First, we make phone transcriptions for the test utterances using recognition results with the baseline speaker-independent model. Using these labels, MLLR adaptation of Gaussian means of the acoustic model is applied and a speaker-adapted model is generated (**adap-all-1**). Using the new recognition results with the adapted model, this process is iterated (**adap-all-2**). The first adaptation reduced the error rate from 35.8 to 31.8%. The second iteration brought further improvement of 0.6%. For reference, we could get an error rate of 29.6% with the super-

Table 3. Word error rate with speaker and speaking-rate adapted models (%)

actual speaking rate	-5	5-6	6-7	7-8	8-9	9-10	10-	average
(#utterances)	(433)	(434)	(596)	(435)	(343)	(161)	(115)	(2517)
baseline	38.7	35.5	34.1	34.1	34.7	39.9	46.4	35.8
adap-all-1	33.3	31.3	30.0	30.5	30.8	36.1	44.8	31.8
adap-all-2	31.4	30.3	29.0	29.9	30.6	36.5	45.4	31.2
adap-fast	31.9	30.3	29.3	**29.7**	**30.2**	**35.5**	**43.1**	31.0
adap-slow	**31.0**	**30.1**	**28.9**	30.6	31.0	35.5	45.0	31.2
adap-fast + adap-slow [oracle]	31.0	30.1	28.9	29.7	30.2	35.5	43.1	30.8
selected with estimated speaking rate	31.4	29.9	29.1	29.9	30.1	35.5	43.1	30.9

vised (oracle) adaptation using the correct transcription of test utterances.

In this process, we take the factor of speaking rate into account, since acoustic patterns in fast segments and slow segments are different even for the same speaker. Specifically, we perform MLLR adaptation separately for fast segments and slow segments. The adaptation scheme will ease the problem of data sparseness that speaking-rate dependent modeling often encounters. In [5], we showed that simple (speaker-independent) speaking-rate dependent modeling lowered the accuracy due to insufficient training data for the respective models.

After preliminary experiments, we set two categories of fast and slow utterances with the boundary of 7 mora/sec (≈ mean of speaking rate). The MLLR adaptation is applied twice from the speaker independent model as in the normal speaker adaptation. As a result, we get models adapted to fast and slow speech (**adap-fast** and **adap-slow**), respectively, which are applied to the corresponding utterances.

The recognition results are listed in Table 3. The adaptation considering the speaking rate brought slight improvement over the normal speaker adaptation (0.3% absolute). It is more effective on faster speech and comparable on slow speech.

It is also noticed that there is a large difference in the speaking rate among the test-set speakers and there are some speakers who tend to speak slow and have only a few fast utterances. For these speakers, this speaking-rate dependent adaptation method does not work properly. Actually, fast utterances are mainly made by half (=five) of the test-set speakers. By looking into these five test speakers who have both fast and slow segments in a sufficient amount, we confirmed accuracy improvement for both slow and fast utterances.

There is no degradation due to estimation errors of the speaking rate this time. Although the improvement is not large, there is little extra computation by the method. The method runs even faster because the adaptation is done separately for the two categories with fewer data for each.

7. CONCLUSIONS

We have presented methods that deal with the speaking rate in the decoding and adaptation techniques.

The speaking-rate dependent decoding strategy applies the most adequate acoustic analysis, phone models and decoding parameters according to the estimated speaking rate. We investigated several methods and demonstrated that the selective application is effective. We have also proposed the use of speaking rate information in speaker adaptation. It is confirmed that the separate adaptation based on the speaking rate works reasonably.

Acknowledgment: The authors are grateful to Prof. Sadaoki Furui and other members of the Science and Technology Agency Priority Program on "Spontaneous Speech: Corpus and Processing Technology".

8. REFERENCES

[1] S. Furui, K. Maekawa, and H. Isahara, "Toward the realization of spontaneous speech recognition – introducing of a japanese priority program and preliminary results –," in *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, 2000, vol. 3.

[2] J.Zheng, H.Franco, and F.Weng, "Word-level rate of speech modeling using rate-specific phones and pronunciations," in *Proc. IEEE-ICASSP*, 2000, pp. 1775–1778.

[3] C.Fugen and I.Rogina, "Integrating dynamic speech modalities into context decision trees," in *Proc. IEEE-ICASSP*, 2000, pp. 1277–1280.

[4] J.Nedel and R.Stern, "Duration normalization for improved recognition of spontaneous and read speech via missing feature methods," in *Proc. ICASSP*, 2001, vol. 1, pp. 313–316.

[5] H.Nanjo, K.Kato, and T.Kawahara, "Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition," in *Proc. EUROSPEECH*, 2001, pp. 2531–2534.

[6] A.Lee, T.Kawahara, K.Takeda, and K.Shikano, "A new phonetic tied-mixture model for efficient decoding," in *Proc. ICASSP*, 2000, pp. 1269–1272.

[7] A.Lee, T.Kawahara, and K.Shikano, "Julius – an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.

[8] C.J.Leggetter and P.C.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.