Speaking Rate Dependent Acoustic Modeling for Spontaneous Lecture Speech Recognition

Hiroaki Nanjo, Kazuomi Kato and Tatsuya Kawahara

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

Abstract

The paper addresses large vocabulary spontaneous speech recognition focusing on acoustic modeling that considers the speaking rate. Using the real lecture speech corpus collected under the priority research project in Japan, we have made baseline acoustic model, and evaluated on the automatic transcription of oral presentations by experienced speakers and obtained word accuracy of 58.2%. Compared with read speech, we have observed significant difference in the speaking rate. To handle fast and poorly articulated phone segments, several extensions of the modeling are explored. Specifically, we introduce state-skipping modeling, speech rate-dependent model, and syllable sub-word modeling. As a result, we reduced the word error rate by absolute 0.8%-2.0%. We also address a language modeling especially on effective use of various large text corpora.

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 target is automatic transcription of live lectures such as oral presentations in conferences.

In acoustic modeling of spontaneous speech, the speaking rate, especially fast speech segments, is considered as one of significant causes of degrading the performance of speech recognition[2] [3] [4]. The articulation is influenced by speaking fast and this causes poor matching The spectral patterns are changed a lot, moreover the phone itself may disappear. Thus, we focus on the modeling of fast speech segments and investigate the speaking rate-dependent model.

2. Database and Test-Set

The lecture speech corpus is being collected under the project called “The Corpus of Spontaneous Japanese (CSJ)”, and it consists of live recordings of oral presentations in technical conferences and studio recordings of monologue speech on given topics such as hobbies and travels. Speech data are recorded by a head-set microphone. As of Oct. 2000, initial portion of them are available, which are listed in Table 1. The data-set CSJ1 amounts to 195 lectures (35.3h) which do not include studio recordings to match the training data-set to the the test-set mentioned below. On the other hand, the data-set CSJ2 includes all of them and amounts to 299 lectures (45.3h). They are all given by male speakers.

The test-set specification is shown in Table 2. The lectures are live presentations in technical conferences by four males. All of them are experienced speakers, and gave lectures without any drafts. The symbol such as “AS” and “PS” in Table 2, indicates the name of a technical conference, for instance “AS” means Acoustical Society of Japan. We can estimate the speaking rate of each presentation from its duration and the number of words. Although *AS99SEP022* and *AS99SEP023* were presented at the same conference, their speaking rates are quite different. The ratio of fillers also varies according to the individual speakers. All speech materials are given without segmentation, thus automatic segmentation is performed based on pause models.

Table 1: Corpus of Spontaneous Japanese (CSJ) (available Oct. 2000)

	Conference (#lectures)	amount
CSJ1	AS(102) + SP(11) + NL(45) + JL(9) + PS(17) + KK(6) + YG(5)	35.3h
CSJ2	CSJ1 + *IG(78)+ *ST(26)	45.3h

*: studio recording monologue

Table 2: Test-Set

	lecture spec.		rate of fillers	
	time	#words	interjections	repairs
AS99SEP022	28min	6305	9.0%	2.9%
AS99SEP023	30min	4391	7.5%	2.2%
AS99SEP097	13min	2508	5.7%	1.1%
PS99SEP025	27min	5372	11.9%	1.2%

As for language model, we use a statistical word trigram model which is trained with the transcriptions of the same corpus (CSJ) and other lecture texts available on World Wide Web. The vocabulary size is 11k. The decoder is Julius-3.1[5], which was developed at our laboratory.

3. Baseline Acoustic Model

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

The number of phones is 43, and all of them are modeled with left-to-right HMM of three states (no state-skipping). We trained context-dependent triphone models. Decision-tree clustering is performed to set up shared-state triphone models of 1000, 2000 and 3000 states, respectively. Each state has 16 mixture components. We also made PTM (phonetic tied-mixture)

Table 3: Evaluation of several acoustic models (%word accuracy)

	training set	
models	CSJ1	CSJ2
monophone 129x32	50.8	50.3
monophone 129x64	52.1	51.6
triphone 1000x16	58.2	58.0
triphone 2000x16	58.2	57.7
triphone 3000x16	56.5	56.4
PTM 129x64 (s1000)	57.8	57.9
PTM 129x64 (s2000)	58.4	58.0
PTM 129x64 (s3000)	58.1	57.4

model[6], where tripohnes of the same phone share Gaussians but have different weights. PTM is usually modeled with a larger number of Gaussians per state, but the total number of Gaussians is mush smaller than conventional shared-state triphone models. Here, 129 codebooks of 64 mixture components are used.

3.1. Evaluation of Baseline Models

Evaluation of baseline acoustic models on the test-set is summarized in Table 3. We got word accuracy of about 58% by both triphone and PTM models.

When we compare the training set of CSJ1 and CSJ2, no significant difference was observed by the addition of data. The fact suggests that there is little effect of adding studio recording data, which were quite different from live presentations. In fact, in the studio recording, as reported in [7], speakers were more relaxed and spoke more casually and slowly.

The performance of the triphone model of 3000 states is degraded, because it is not fully trained with the given training data size. Thus, we use the triphone 2000x16 as the baseline in the following section.

The overall accuracy was quite poor, compared with the newspaper corpus task (20k vocabulary), where we adopted the same modeling with the comparable size of training speech data and achieved accuracy of 90-95%[8].

4. Speaking Rate-Dependent Modeling

4.1. Motivation

The speaking rate is considered to significantly affect the performance of speech recognition[3] [4]. In [3], speaking rate-dependent phone models are prepared for fast, normal and slow phone segments. In [4], they introduced speaking rate information in decision-tree clustering.

We first investigate the duration distributions of phone segments in live presentations (CSJ1: 35h) and read speech (Japanese Newspaper Article Sentences corpus: 40h) in Fig. 1. It is obvious that the distributions are quite different and the live lectures are spoken faster than the read speech. In particular, there are a lot of phone segments that have duration of three frames (30msec) in CSJ. Since each phone segment is modeled with three state HMM without state-skipping, minimum duration of Viterbi aligned phone segments is three frames. These segments may have fewer durations, but are forcedly assigned to three frames. This causes a significant problem because the correct acoustic score is not evaluated. Moreover, spectral patterns are changed in the segments because of fast articulation. In order to solve this problem, we modify the HMM structure so that the speaking rate is taken into account. Specifically, we present the following three methods.

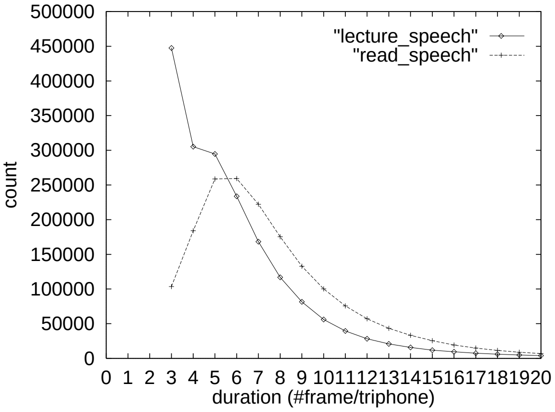

Figure 1: Duration distributions of read and lecture speech

Table 4: Effect of state-skipping transitions (%word accuracy)

	state-skipping	baseline
AS99SEP022	57.6%	56.5%
AS99SEP023	61.9%	59.2%
AS99SEP097	70.3%	67.3%
PS99SEP025	59.1%	57.1%
average	60.7%	58.7%

triphone (s2000x16mix)

- add state-skipping transitions.
- make new HMM with one or two states for typical triphones that are articulated fast.
- set up syllable models that are frequently articulated fast instead of phones.

Details are described in the following sections.

4.2. State-Skipping Transitions

As the simplest modification to cope with the fast speech, we add a transition arc from the first state to the third state. Gaussian distributions and state transition probabilities are re-estimated.

Table 4 lists word accuracy of this model compared with the baseline. Both of them are modeled with triphone (s2000x16mix)[1]. The state-skipping transition has good effect on the short phone segments, and we reduced word error rate by 2% absolutely.

4.3. Speaking Rate-Dependent Phone Model

The model of the previous section does not model changes of spectral patterns caused by fast articulation. Therefore, we introduce speaking-rate dependent phone modeling that have dedicated models for normal speech and fast speech

To handle fast segments shorter than three frames, we prepare one state or two state model for fast phones. Phone segments are labeled with normal/fast based on the duration according to Viterbi algorithm. There is a tendency that some phones are more likely to be articulated fast. We select such

[1] Changing decoding parameters, baseline word accuracy is different from that of Table 3.

Table 5: phone level speaking rate-dependent model (%word accuracy)

	rate-ph1	rate-ph2	baseline
AS99SEP022	54.8%	53.3%	56.5%
AS99SEP023	60.0%	59.9%	59.2%
AS99SEP097	67.6%	66.9%	67.3%
PS99SEP025	57.9%	54.9%	57.1%
average	58.6%	57.1%	58.7%

rate-ph1:1state-HMM, rate-ph2:2states-HMM

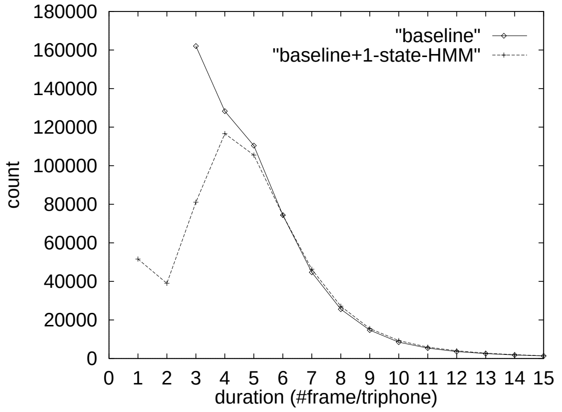

Figure 2: Duration distribution with speaking-rate dependent model

phones (triphone contexts) that have enough training data for both normal and fast HMM. The number of selected triphones is 154. Many of these triphones are concerned with functional words and auxiliary verbs that may not be clearly articulated, for example "-Nde" and "-desu". By assuming that the changes of the speaking rate occur at word boundaries, we prepare both normal and fast baseform for each word in the lexicon if it contains the selected 154 triphones. For example, the word "DESUGA" that is transcribed as /d e s u g a/ for a normal baseform and contains the selected triphone /d+e/,/d-e+s/,/e-s+u/,/s-u+g/,/u-g+a/ is transcribed as /d_f e_f s_f u_f g_f a/ for a fast baseform.

We trained triphone models (s2000x16mix) and evaluated with the test-set. The results are shown in Table 5. No improvement was observed. The case using one-state model for the fast segments has better performance than that using two-state model. The one-state model achieved comparable performance to the baseline. The decrese is caused by the shortage of training data, because the modeling demands different sets for fast and normal segments.

We investigated the distribution of phone segments with the baseline and one-state models and plot it in Fig. 2. Fast models are obviously used and there are a lot of segments of only one frame, which may be actually missing. In [3], they introduced the zero-length-phone model for such segments.

4.4. Speaking Rate-Dependent Syllable Model

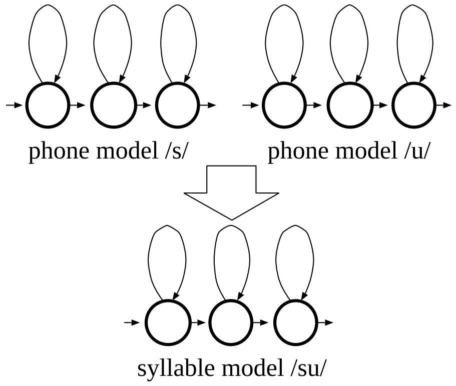

Figure 3: Speaking Rate-Dependent Syllable Model

Table 6: syllable level speaking rate-dependent model (%word accuracy)

	rate-syl	baseline
AS99SEP022	55.7%	56.5%
AS99SEP023	61.4%	59.2%
AS99SEP097	69.6%	67.3%
PS99SEP025	58.0%	57.1%
average	59.5%	58.7%

Since not a few phone segments may disappear, we modeled them with syllables of phone sequence as illustrated in Fig. 3. In Japanese, every syllable is made up of a consonant followed by a vowel. We select syllables considering both training data amount and the speaking rate. The following statistic is defined as a criterion for selection.

$$V_s = \sum_i P^{Duration(s_i)}$$

where s_i is a sample *i* of syllable *s*, *P* is an average probability of self-looping transition ($= 0.56$) and $Duration(s_i)$ is a number of frames with which s_i is aligned. The more fast segments occur, the value of V_s gets larger.

We selected 30 syllables based on the criterion. They are all concerned with functional words. We modeled these syllables with three state HMM and evaluated with the test-set. The result is shown in Table 6. Comparing the results of Table 5 and Table 6, the syllable model is more effective than the phone model for fast speech segments.

5. Improvement in Language Modeling

5.1. Incorporation of Other Text Corpus

We have also made several improvements in language modeling. For training language model of spontaneous speech, it is necessary to collect a corpus of accurate transcriptions. The text size is essentially much smaller than written text corpus such as newspaper articles and broadcast drafts, since recording and manually transcribing spontaneous speech costs a lot.

Thus, we explore effective use of various text corpora. Specifically, texts of lecture notes available via World Wide Web are collected. A topic-independent vocabulary selection based on mutual information criterion is performed[9]. The text size amounts to 1692K words in total, which is four times larger than the CSJ corpus built so far. These texts are not actual transcription of lectures, but manual editing process is performed for readability. It is not matched for language modeling of spontaneous lecture speech recognition.

Therefore, adaptation or weighted combination of text corpora is introduced. Suppose the occurrence count of word se-

Table 7: coverage and perplexity

	CSJ	WEB	CSJ+WEB (simple → optimized)
data amount	0.47M	1.7M	2.2M
vocabulary	10K	8K	13K
AS99SEP022	94.5% 141.2	87.0% 210.6	95.1% 159.0 → 149.8
AS99SEP023	95.7% 127.3	83.8% 221.8	96.1% 157.3 → 146.3
AS99SEP097	95.9% 140.6	85.5% 151.0	96.3% 162.5 → 153.7
PS99SEP025	95.7% 193.0	81.9% 320.3	96.1% 253.8 → 223.8

Table 8: Effect of corpus combination (word accuracy)

			CSJ+WEB	
	CSJ	WEB	simple	optimized
AS99SEP022	55.5%	51.2%	56.4%	57.9%
AS99SEP023	68.1%	49.0%	66.5%	67.9%
AS99SEP097	67.8%	61.0%	68.6%	70.4%
PS99SEP025	60.3%	45.8%	58.8%	61.4%
average	61.5%	50.4%	61.1%	63.0%

quence W in the matched corpus (=CSJ) is $C_1(W)$ and that in the un-matched large corpus (=Web) is $C_0(W)$, then these corpora are combined by the following formula.

$$\lambda_0 \cdot C_0(W) + \lambda_1 \cdot C_1(W) \quad (1)$$

We adopt the weighted combination of word count level, rather than probability level, because it is more straightforward in computing back-off coefficients.

5.2. Optimization of Weights with Deleted-Interpolation Method

Many previous works do not address automatic optimization of the weight parameters λ. In the previous section of evaluation of acoustic modeling, we use simple concatenation of two text corpora, that means $\lambda_0 = \lambda_1 = 1$. Here, estimation of the weights without using the test-set is done with the deleted-interpolation method. We split the matched corpus (=CSJ) into M (=7) portions, and estimate parameter λ_m that minimizes perplexity of each $1/M$ portion using the other $(M-1)$ portions combined with large corpus (=Web). We repeat this process M times and calculate an average value , which is set as an estimated weight parameter λ_0 and λ_1.

As a result, we have derived λ_1=0.93 and λ_0=0.16.

5.3. Evaluation of Language Model

We compared language models with different training sets in terms of coverage and perplexity as shown in Table 7. As the vocabulary sizes are different among models, we cannot compare perplexity values directly. From the result, we verified the mis-match of Web lecture notes with the actual speech transcription caused by the post-processing process.

The recognition results on the test-set using PTM triphone model (s2000) are listed in Table 8. [2]

A simple concatenation of two text corpora actually gave no improvement of the performance, but the proposed automatic optimization method improves both perplexity and the recognition accuracy. It reduced the error rate by absolute 1.5%.

6. Conclusion

We have studied acoustic modeling that considers the speaking rate. We have made baseline acoustic model using the real lecture speech corpus (CSJ), and evaluated on the automatic transcription of oral presentations by experienced speakers. Comparing spontaneous speech with read speech, significant difference in the speaking rate has been observed. Since fast segments in spontaneous speech are not clearly articulated and poorly modeled, several extensions of the modeling are explored. The state-skipping transition is effective. With speaking rate-dependent phone model, no improvement was observed, we confirmed that not a few phone segments disappear in fast segments. So we modeled them with syllables of phone sequence, and achieved some improvement. As a result, we reduced the word error rate of 0.8%-2.0%.

We also investigated language modeling especially on effective use of large text corpora. Specifically, weighted combination of Web texts and automatic estimation of the weights are introduced. The method improves the perplexity and the recognition accuracy by 1.5%.

Acknowledgment: The work is supported by the priority research program of the science and technology agency "spontaneous speech: corpus and processing technology". The authors express their thanks to Professor Frui at Tokyo Institute of Technology and other members of the project for their comments and cooperation.

7. 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] Thilo Pfau and Guenther Ruske, "Creating Hidden Markov Models for Fast Speech," in *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, 1998, vol. 2.

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

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

[5] A.Lee, T.Kawahara, and S.Doshita, "An efficient two-pass search algorithm using word trellis index," in *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, 1998, pp. 1831–1834.

[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] T Kagomiya, H. Kikuchi, H. Koiso, and K. Maekawa, "Variety of speech style in large-scale corpus of spontaneous speech –analysys of transcription –(in Japanese)," in *THE 2000 AUTUMN MEETING OF THE ACOUSTICAL SOCIETY OF JAPAN*, Sept. 2000, 2-Q-9.

[8] T.Kawahara et al, "Free software toolkit for Japanese large vocabulary continuous speech recognition," in *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, 2000, vol. 4, pp. 476–479.

[9] K.Kato, H.Nanjo, and T.Kawahara, "Automatic transcription of lecture speech using topic-independent language modeling," in *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, 2000, vol. 1, pp. 162–165.

[2] As we fixed the lexical entries and modified handling of pauses in the language model, the baseline word accuracy is improved.