AUTOMATIC TRANSCRIPTION OF SPONTANEOUS LECTURE SPEECH

Tatsuya Kawahara *Hiroaki Nanjo*

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

Sadaoki Furui

Tokyo Institute of Technology
Meguro-ku, Tokyo 152-8552, Japan

ABSTRACT

We introduce our extensive projects on spontaneous speech processing and current trials of lecture speech recognition. A large corpus of lecture presentations and talks is being collected in the project. We have trained initial baseline models and confirmed significant difference of real lectures and written notes. In spontaneous lecture speech, the speaking rate is generally faster and changes a lot, which makes it harder to apply fixed segmentation and decoding settings. Therefore, we propose sequential decoding and speaking-rate dependent decoding strategies. The sequential decoder simultaneously performs automatic segmentation and decoding of input utterances. Then, the most adequate acoustic analysis, phone models and decoding parameters are applied according to the current speaking rate. These strategies achieve improvement on automatic transcription of real lecture speech.

1. INTRODUCTION

Automatic speech recognition of read speech has been successful in achieving accuracy of over 90% and realizing dictation systems. The system, however, assumes that users clearly utter grammatically correct sentences with orthodox pronunciation as the human-to-machine interfaces. On the other hand, recognition of human-to-human spontaneous speech, which would realize applications of automatic transcription or translation of lectures and meetings, is very poor and needs more extensive studies.

From this perspective, we started the project of "Spontaneous Speech Corpus and Processing Technology" sponsored by the Science and Technology Agency Priority Program in Japan. The project is conducted over 5 years (1999-2004) in pursuit of the following three major targets[1].

(1) Building a large-scale spontaneous speech corpus.

The designed *Corpus of Spontaneous Speech (CSJ)* consists of roughly 7M words or 700 hours. Manily recorded are monologues such as lectures, presentations and news commentaries. They are manually given orthographic and phonetic transcription. One-tenth of the corpus ("Core") will be tagged manually with morphological and paralinguistic information for linguistic analysis.

(2) Acoustic and linguistic modeling for spontaneous speech recognition, understanding and summarization.

(3) Constructing a prototype of a spontaneous speech summarization system.

In this paper, we report initial studies on speech recognition of lecture presentations using the corpus compiled at present. Lecture speech can be regarded as in-between of broadcast news and telephone conversation, both of which are widely dealt with so far. The speaker is not professional, nor reading a draft material as in broadcast news[1]. But the speaking style is not so casual as in telephone conversation. One of the prominent characteristics in spontaneous monologue speech is the speaking rate is generally faster and changes a lot. Since speakers do not necessarily utter sentence by sentence, there are many segments of very long duration. We address acoustic and language modeling as well as decoding strategies considering these factors.

2. DATABASE AND TASK

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. The speech data are recorded via a head-set microphone to digital audio tapes, and digitized at 16 kHz and 16 bit sampling. They are not segmented at all, i.e. one large file corresponds to a lecture.

For language model training, all transcribed data (as of June 2001) are used. There are 612 presentations and talks by distinct speakers basically. 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. It is observed that there is much difference in speaking rate among these speakers.

[1]Some presenters are reading a draft, but we do not include such kind of speech in the test-set.

Table 1: Test-set of lectures

	#words	duration (min.)
A01M0035 (AS22)	6294	28
A01M0007 (AS23)	4391	30
A01M0074 (AS97)	2508	12
A05M0031 (PS25)	5372	27
A02M0117 (JL01)	9833	57
KK99DEC005 (KK05)	6527	42
A03M0100 (NL07)	2644	15
A06M0134 (SG05)	4460	23
YG99JUN001 (YG01)	2759	14
YG99MAY005 (YG05)	3108	15

3. BASELINE MODEL

3.1. Acoustic Modeling

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

The number of phones 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 3000 shared-states. We also adopt PTM (phonetic tied-mixture) modeling[2], where triphone states of the same phone share Gaussians but have different weights. Here, 129 codebooks of 64 mixture components are used.

3.2. Language Modeling

We built a lexicon of 19158 words from the training corpus, and then a trigram language model. It realizes coverage of 97% and test-set perplexity of 135. The perplexity is very large since the Japanese morpheme unit is shorter than English words and that of the newspaper task is about 50-80. Main reason is the amount of training data is not sufficient while the topics and domains of lectures are of a wide variety. The training data for spontaneous speech is essentially much smaller than written text corpora such as newspaper articles, since recording and manually transcribing spontaneous speech costs a lot.

Therefore, 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[3]. The text size amounts to 1.69M words in total, which is larger than the CSJ corpus built so far. These texts are not actual transcription of lectures, but manual editing process is performed for readability.

Then, weighted combination of text corpora is performed. Suppose the occurrence count of word sequence 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.

$$C(W) = \lambda_0 \cdot C_0(W) + \lambda_1 \cdot C_1(W)$$

Here, estimation of the weights is done with the deleted-interpolation method by splitting the matched corpus (=CSJ) into M portions. As a result, we derived values of λ_1=0.95 and λ_0=0.066.

Preliminary evaluation with the four test lectures shows that combination of texts improved accuracy to 65.6% compared with the baseline (=CSJ only) of 65.1%. The weight of the Web text is very small and only a little improvement is observed. The result suggests that even lecture notes are much different and not good for language modeling of real lectures [2]. But we use the combined model in the following experiments.

We also tried to incorporate the newspaper corpus with a sentence selection mechanism, but only got performance degradation.

4. SEQUENTIAL DECODING WITH MULTIPLE-PASS SEARCH

In spontaneous speech, utterances do not necessarily match the linguistic sentences, because people put pauses on arbitrary timing. In giving lectures, speakers often utter many sentences without a break and sometimes put many filled and un-filled pauses. When we cut recorded material based on pauses in pre-processing, there are a lot of very long utterances as well as many short segments of only fillers.

Too long inputs are hard to deal with for the conventional decoders, especially two-pass decoders including our Julius[4] which keeps a huge number of candidates before re-scoring. Moreover, getting N-best string lists is almost meaningless when the input is too long. Automatic segmentation of spontaneous speech itself is not so easy because existence of weakly articulated portions and change of the speaking rate make it harder to use a fixed threshold.

To solve the problem, we revise our two-pass forward-backward decoding algorithm so that it does not need prior segmentation of speech. It simultaneously performs recognition and segmentation with model-based detection of short pauses. We have a specific model for short pauses both in acoustic and language modeling. When the short pause model is ranked first in the hypotheses beam for consecutive frames, the decoder suspends the forward search and performs backward search. Then, the decoding is resumed using the fixed word history.

The method makes full use of acoustic and language model in detecting short pauses, thus it is more reliable than

[2]Use of Web lecture notes was more effective when the size CSJ corpus was smaller.

Table 2: Word accuracy using sequential decoding (%)

	conventional decoder	sequential decoder
AS22	58.9	60.1
AS23	72.4	71.9
AS97	72.5	73.8
PS25	64.7	65.2
JL01	62.7	64.8
KK05	64.7	66.8
NL07	68.0	69.0
SG05	58.6	57.4
YG01	61.5	63.3
YG05	67.2	68.0
average	64.2	65.3

the prior segmentation using conventional end-point detection algorithms. Moreover, on-line adaptation of the short pause model is also possible.

This sequential decoding algorithm is evaluated on the whole test-set. Word accuracy is listed in Table 2 in comparison with the conventional decoder that takes prior segmentation. The proposed sequential decoder successfully handles whole lecture speech (of 12-57 minutes) and achieves better accuracy.

5. SPEAKING-RATE DEPENDENT DECODING

Distribution of phone duration in lecture speech (CSJ corpus: 35 hours) and read speech (JNAS 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 corpus may have fewer durations, but are forcedly alinged with three frames. This may have caused a serious mis-match. Moreover, fast speaking rate suggests that these segments are poorly articulated and big problem in recognition [5][6][7].

Relationship between the word accuracy and speaking rate is plotted for the test-set. Speaking rate is defined as the mora counts divided by the utterance duration (sec). It is confirmed that faster utterances are harder for recognition. In Figure 2, breakdown of recognition errors is shown for each 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.

Based on these facts, we propose applying different decoding methods according to the speaking rate within the multiple-pass search framework. 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 applied. The

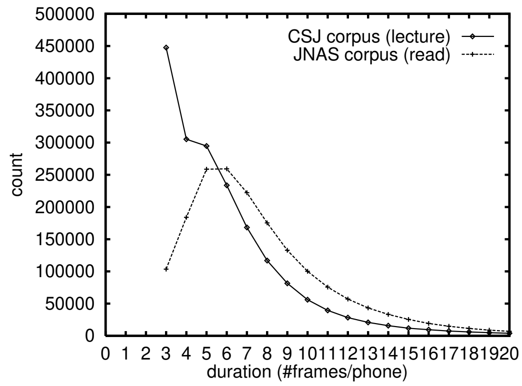

Figure 1: Phone duration distribution of CSJ and JNAS corpus

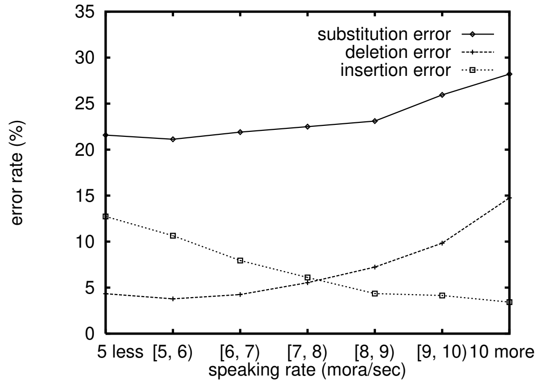

Figure 2: Ratio of substitution, deletion and insertion errors for each speaking rate

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 phone sequence. We select syllables by considering both their duration and training data amount[8].

(4) Change insertion penalty

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

Table 3: Accuracy with different decoding according to speaking rate (%)

actual speaking rate (#utterances)	-5 (433)	5-6 (434)	6-7 (596)	7-8 (435)	8-9 (343)	9-10 (161)	10- (115)	average (2517)
baseline	61.3	64.5	65.9	65.9	65.3	60.1	53.6	64.2
1. analysis frame	60.3	65.5	66.5	66.9	**67.2**	**61.7**	**56.1**	65.3
2. skipping transition	62.3	66.0	**66.6**	**67.2**	65.8	60.7	54.8	65.2
3. syllable model	59.6	64.6	66.2	65.9	66.6	61.1	56.2	64.7
1.+2.	59.0	64.0	65.1	65.3	65.3	60.4	56.0	63.8
1.+3.	56.0	61.8	64.4	65.5	66.0	62.4	56.5	63.5
2.+3.	60.5	64.5	66.3	66.3	66.1	62.8	57.0	64.9
1.+2.+3.	54.3	60.7	63.4	64.9	66.2	62.0	57.9	62.9
4. insertion penalty	**64.3**	**67.3**	66.4	64.7	62.8	55.8	50.1	63.7
best one selected [oracle]	64.3	67.3	66.6	67.2	67.2	61.6	56.1	65.9
selected with estimated speaking rate	62.6	66.4	66.7	66.9	66.4	60.6	55.8	65.4

These techniques and their combinations are evaluated on the test-set. They are compared with the baseline decoding that first segments speech into utterances for convenience of experiments. Utterances are labeled with the speaking rate (mora/sec). The results are listed in Table 3.

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 the effect. For slow utterances, the use of severe insertion penalty reduces errors as expected.

Then, selective application of these methods according to the speaking rate is implemented, as specified with bold font in Table 3. If the speaking rate is known and 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 phone models and syllable constraint and apply the different decoding methods in the second pass. This strategy achieves improvement of 1.2% absolute (last row).

6. CONCLUSIONS

We have set a task of automatic lecture transcription for spontaneous speech recognition and understanding. In the first half of the five-year project, we have collected the largest corpus for this purpose and made clear the problems of spontaneous speech.

In this paper, we mainly address decoding strategies dedicated for spontaneous lecture speech. One is sequential decoding and the other is speaking-rate dependent decoding. Both of them are shown to be effective and their combination is ongoing. We plan to further pursue speaking-rate normalization approach combined with speaker adaptation techniques.

Acknowledgement: The authors are grateful to Dr. Maekawa of NIJLA and Dr. Isahara of CRL and other members of the project for fruitful collaboration and discussions.

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 –. In *Proc. ICSLP*, volume 3, pages 518–521, 2000.

[2] A.Lee, T.Kawahara, K.Takeda, and K.Shikano. A new phonetic tied-mixture model for efficient decoding. In *Proc. IEEE-ICASSP*, pages 1269–1272, 2000.

[3] K.Kato, H.Nanjo, and T.Kawahara. Automatic transcription of lecture speech using topic-independent language modeling. In *Proc. ICSLP*, volume 1, pages 162–165, 2000.

[4] A.Lee, T.Kawahara, and K.Shikano. Julius – an open source real-time large vocabulary recognition engine. In *Proc. EUROSPEECH*, (to appear), 2001.

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

[6] C.Fugen and I.Rogina. Integrating dynamic speech modalities into context decision trees. In *Proc. IEEE-ICASSP*, pages 1277–1280, 2000.

[7] J.Nedel and R.Stern. Duration normalization for improved recognition of spontaneous and read speech via missing feature methods. In *Proc. ICASSP*, volume 1, pages 313–316, 2001.

[8] H.Nanjo, K.Kato, and T.Kawahara. Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition. In *Proc. EUROSPEECH*, (to appear), 2001.