EFFICIENT ACCESS TO LECTURE AUDIO ARCHIVES THROUGH SPOKEN LANGUAGE PROCESSING

Tatsuya Kawahara Tasuku Kitade Kazuya Shitaoka Hiroaki Nanjo

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
kawahara@i.kyoto-u.ac.jp
http://www.ar.media.kyoto-u.ac.jp/

ABSTRACT

The paper firstly addresses the current state of speech recognition using the "Corpus of Spontaneous Japanese (CSJ)". It is shown that the large-scale corpus had strong impact in training acoustic and language models considering morphological and pronunciation variations which are characteristic to spontaneous Japanese. Unsupervised adaptation of these models and the speaking rate is also effective, and we obtained word accuracy of 78.0%. Then, an intelligent archiving system of lectures based on automatic transcription and indexing is introduced. Transcriptions are automatically edited for improving readability, and key sentences are indexed based on statistically-derived discourse markers and topic words. Thus, we realize efficient browsing of lecture audio archives.

1. INTRODUCTION

Recent progress of large-volume storage devices and high-speed networks has enabled digital archiving and streaming of audio and video materials. In academic societies and universities, multi-media archives of lectures will be technically feasible. Such archives would help students audit lectures at their convenient time and places with their own paces. In these kinds of audio archives, appropriate indices are necessary for efficient browsing and searching portions of specific topics or speakers. Spoken language technologies will be useful for automating the indexing process which would cost a lot if manually done.

Toward this application, we have studied following issues of spoken language processing.

(1) Automatic transcription of spontaneous speech

While automatic speech recognition (ASR) of read speech has achieved accuracy exceeding 90%, spontaneous speech recognition faces difficult problems of acoustic and linguistic variations which are yet to be solved. We have taken part in the project of "Spontaneous Speech Corpus and Processing Technology" sponsored by the Science and Technology Agency Priority Program in Japan[1][2]. The *Corpus of Spontaneous Speech (CSJ)*[3] developed by the project consists of roughly 7M words or 500 hours, which is the largest in scale. The corpus has been an infrastructure of our studies presented in this paper.

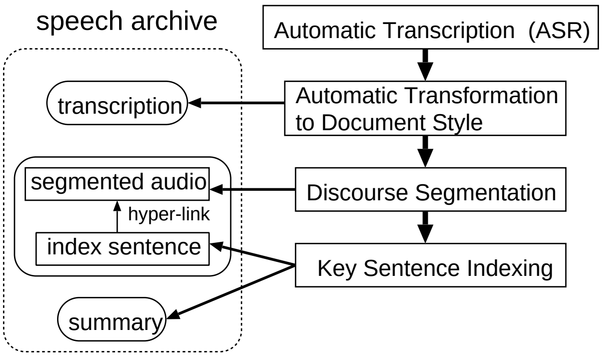

Fig. 1. System overview

(2) Automatic segmentation of a lecture into sentences and sections

Baseline indexing for quick browsing of lecture audio is done by sentence segmentation. We realize the process in a framework of transforming (or translating) the raw transcription into document style. Moreover, we have proposed segmentation of sections by assuming a lecture-style discourse structure[4]. The method is based on presumed discourse markers that are derived in an unsupervised manner.

(3) Automatic indexing of key sentences

More elaborate indexing for efficient browsing is realized by extracting key sentences, as they concisely express topics of the portions. We present a statistical measure of importance of sentences by focusing on both discourse markers and topic words.

We are developing an intelligent lecture archiving system based on these approaches, which are addressed in this paper. The overview of system is depicted in Figure 1.

2. AUTOMATIC TRANSCRIPTION SYSTEM

Oral presentations are 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. But the speaking style is not so casual as in telephone conversation.

As many previous studies point out, various factors in spontaneous speech affect ASR performance. They include acoustic variation caused by fast speaking and imperfect articulation, and linguistic variation such as colloquial expressions and disfluencies. Thus, the problems should be addressed from the viewpoint of acoustic, pronunciation and language modeling.

We also revised our recognition engine Julius [1] so that very long speech can be handled without prior segmentation[5].

2.1. Acoustic Model

A large portion of the CSJ consists of two styles of monologues. One is academic presentation speech at technical conferences and meetings, and the other is extemporaneous public speech on given topics such as hobbies and travels. We have set up a variety of baseline acoustic models[6]. Since the speaking style is apparently different for academic presentation speech and extemporaneous public speech, respective models are trained.

In this paper, we focus on academic presentation speech given by male speakers. The training data consist of 781 presentations that amount to 106 hours of speech.

Acoustic models are based on diagonal-covariance 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. We trained a PTM (phonetic tied-mixture) triphone model[7]. Decision-tree clustering was performed to set up 3000 shared states. In PTM modeling, triphone states of the same phone share Gaussians but have different weights. Here, 129 (=43*3) codebooks of 192 mixture components were used. As a whole, there are 25K Gaussian components and 576K mixture weights.

Increase of training data thanks to the increased size of the CSJ consistently, though modestly, improved the word accuracy. For example, increase from 38 hours to 60 hours results in the reduction of WER (Word Error Rate) from 35.8% to 34.7% with the former language model. For reference, the standard read speech model[8] obtained a higher WER by about 10% absolute.

2.2. Language and Pronunciation Model

A baseline language model is constructed using the transcriptions of 2592 talks excluding the test-set. The total text size is about 6.67 million words including fillers and word fragments. Word segmentation was automatically done using a morphological analyzer that was trained with the maximum entropy criterion[9].

In spontaneously spoken Japanese, pronunciation variation is so large that a number of surface form entries are needed for a lexical item. We found that statistical modeling of pronunciation variations integrated with the language modeling was effective in suppressing false matching of less frequent entries[10]. Here, we adopt a simple trigram model of word-pronunciation entries.

Transcription of the CSJ was made manually both in an orthographic notation and a phonetic (*kana*) one for each utterance unit. Thus, automatic alignment of the two by the word unit is needed to obtain the word-pronunciation entries. This was incorporated as a post-processor of the morphological analyzer[9]. Some heuristic thresholding is applied to eliminate erroneous patterns. As a result, we get 30820 word-pronunciation entries (24437 distinct words), for which a trigram model is trained.

The effect of training data size is clearly confirmed in Table 1. WER (Word Error Rate) is significantly reduced according to the increase of the data. For reference, the addition of lecture note archives that were post-edited for document style has little effect[5] when the matched training data are increased. The result strongly demonstrates that the corpus of this scale is meaningful in modeling spoken language.

Next, the effect of statistical pronunciation modeling is shown in Table 2, where the cases of single pronunciation and multiple pronunciation entries without statistics are compared with statistical models. Here, pron-unigram is a model that adopts pronunciation unigram within individual word entries, for which a trigram model is trained. On the other hand, pron-trigram is trained for word-pronunciation pairs. The result shows that the statistical modeling, especially the word-pronunciation trigram model, is effective. The model training was also made possible thanks to the large scale corpus.

2.3. Model Adaptation and Speaking Rate Dependent Decoding

Next, we incorporate speaker adaptation of acoustic and language models. Since lecture speech has long duration (large data) per speaker, the unsupervised adaptation scheme works very well.

First, we generate transcriptions for the test utterances using the baseline speaker-independent model. For acoustic model, MLLR adaptation of Gaussian means is performed

[1] downloadable at `http://julius.sourceforge.jp`

Table 1. Effect of language model training data

	LM1	LM2	LM3	LM4	current*
# talks	186	316	612	1125	2592
text size	0.5M	0.8M	1.5M	2.7M	6.3M
voca. size	10K	13K	19K	21K	24K
OOV rate	4.7	4.0	3.2	3.0	1.5
perplexity	152.8	143.2	134.1	115.4	105.6
WER	38.5	36.2	34.9	34.5	33.?

Since the acoustic model used in ASR is a former version, the overall WER is lower than the latest result.

* Current system adopts a different morphological system, thus cannot be directly compared with former versions. The figures are estimated.

Table 2. Effect of pronunciation modeling

method	WER
single pron. per word	31.6
multiple pron. per word	31.4
pron-unigram	30.7
pron-trigram	30.5

using the phone labels of the initial recognition result, and a speaker-adapted model is generated.

We have also studied unsupervised methods of language model adaptation to a specific speaker and a topic[10], which are based on a model trained with the initial transcription. The first method is to select similar texts using the word perplexity and TF-IDF measure and weight them in re-training. The second method makes direct use of the model generated from the initial recognition result by linear interpolation with the baseline model. It was shown that all proposed adaptation methods and their combinations reduce the perplexity and WER[10].

We also proposed a decoding strategy adapted to the speaking rate[11]. In spontaneous speech, speaking rate is generally fast and may vary a lot within a presentation. We also observe different error tendencies for portions of presentations where speech is fast or slow. The proposed speaking rate dependent decoding strategy applies the most appropriate acoustic analysis, phone models, and decoding parameters according to the speaking rate. Several methods were investigated and their selective application led to improved accuracy[11].

The effect of these methods for the task of transcription of 15 academic presentations is summarized in Table 3. The unsupervised acoustic model adaptation reduced WER by 4.9% absolute from 30.9% to 26.0%, and the combination with the language model adaptation methods reduced WER further by 2.1% absolute. The speaking rate dependent decoding strategy gained additional improvement of 1.9% absolute. Finally, WER of 22.0% is achieved.

Table 3. Effect of model and decoding adaptation

method	WER
baseline	30.9
+ acoustic model adaptation	26.0
+ language model adaptation	23.9
+ speaking rate adaptation	22.0

3. AUTOMATIC TRANSFORMATION OF TRANSCRIPTION INTO DOCUMENT STYLE

Transcriptions of lecture speech include many colloquial expressions peculiar to spoken language. The Japanese spoken language in particular is quite different from the written language, and is not suitable for documents in terms of readability. Thus, it is necessary to transform transcriptions and recognition results into document style for practical archives. This process is also important as a pre-process of automatic summarization.

We approach the problem by using a statistical framework that has become popular in machine translation. We regard the spoken and written Japanese languages as different languages and apply the translation methodology to transform the former into the latter. Within this framework, correction of colloquial expressions, deletion of fillers, insertion of periods (end-of-sentence symbols), and insertion of particles are performed in an integrated manner[12].

The statistical machine translation framework is formulated by finding the best output sequence Y for an input sequence X, such that a posteriori probability $P(Y|X)$ is maximum. According to Bayes rule, maximization of $P(Y|X)$ is equivalent to the maximization of the product (sum in log scale) of $P(Y)$ and $P(X|Y)$, where $P(Y)$ is the probability of the source language model and $P(X|Y)$ is the probability of the transformation model. The transformation model represents correspondence of input and output word sequences.

In the task of style conversion, the input X is a word sequence of spoken language transcriptions that do not have periods but include pause duration. The output Y is a word sequence of the written language. For $P(Y)$ calculation, we use a word 3-gram model trained with a written language corpus. Since the conversion of one word affects neighbor words in an N-gram model, the decoding is performed for a whole input word sequence with beam pruning.

3.1. Correction of Colloquial Expressions

$P(X|Y)$ represents the probability that a colloquial expression X arises for a written expression Y. It is estimated from the parallel corpus of exact transcriptions of spoken language and texts after correction by a human editor. We define 64 conversion pairs and estimate their probabilities with a parallel corpus of 18 lectures of the CSJ.

3.2. Insertion of Particles

Since particles are often omitted in spontaneous Japanese, they needs to be recovered for document-style text. As the phenomena is dependent on adjacent words, we define the deletion probabilities of particles $P(X|Y)$ for the triplet of the preceding part of speech, the particle itself, and the following part of speech, such as *"Noun Particle Noun"*, *"Noun Particle Verb"* and *"Noun Particle Adjective"*.

3.3. Insertion of Periods

In recognizing read speech, periods are conventionally assigned to pauses at the end of utterances because an utterance is assumed to be a sentence. In spontaneous speech, however, pauses are put not only at the end of sentences but at arbitrary places. Thus, the CSJ has pause marks with their duration instead of periods, and the speech recognizer does not output periods. However, periods are needed in document-style text for better readability.

Our proposed method converts pauses into periods selectively using a threshold function $P(Y|X)$ that considers duration information and the adjacent parts of speech, as well as the language model score $P(Y)$. Specifically, the pause duration threshold of X with which pauses are converted to periods is set up depending on the contextual words of Y.

4. AUTOMATIC INDEXING OF KEY SENTENCES

Next, we address automatic extraction of key sentences, which will be useful indices in lectures. Collection of these sentences may suffice summarization of the talk. The framework extracts a set of natural sentences, which can be aligned with audio segments for alternative summary output.

4.1. Discourse Modeling of Lecture Presentations

There is a relatively clear prototype in the flow of presentation, which is similarly observed in technical papers[13]. When using slides for presentation, one or a couple of slides constitute a topic discourse unit we call 'section' in this paper. The unit in turn usually corresponds to the (sub-)sections in the proceedings paper.

It is also observed that there is a typical pattern in the first utterances of the units. Speakers try to briefly tell what comes next and attract audiences' attention. For example, "Next, I will explain how it works." and "Now, move on to experimental evaluation". We define such characteristic expressions that appear at the beginning of section units as discourse markers. We proposed a method to automatically train a set of discourse markers without any manual tags, and shown the effectiveness in segmentation of lecture speech[4].

The boundary of sections is known as useful for extracting key sentences in the text-based natural language processing. However, the methodology cannot be simply applied to spoken language because the boundary of sections is not explicit in speech. Thus, we apply the discourse segmentation to extraction of key sentences from lectures[14].

4.2. Statistical Derivation of Discourse Markers

It is expected that speakers put relatively long pauses in shifting topics or changing slides, although a long pause does not always mean a section boundary. Here, we set a threshold on pause duration to pick up boundary candidates. We use the average of pause length during a talk as the threshold.

From the candidates of the first sentences picked up by the pause information, we extract characteristic expressions, namely select discourse markers useful for indexing. Discourse markers should frequently appear in the first utterances, but should not appear in other utterances so often. Word frequency is used to represent the former property and sentence frequency is used for the latter. For a word w_j, the word frequency wf_j is defined as its occurrence count in the set of first sentences. The sentence frequency sf_j is the number of sentences in all lectures that contain the word. We adopt the following evaluation function.

$$S_{DM}(w_j) = wf_j * \log(\frac{N_s}{sf_j}) \qquad (1)$$

Here, N_s is the total number of sentences in all lectures. A set of discourse markers are selected by the order of $S_{DM}(w_j)$.

4.3. Measure of Importance based on Discourse Markers

In the text-based natural language processing, a well-known heuristics for key sentence extraction is to pick up initial sentences of the articles or paragraphs. Using the automatically-derived discourse markers that characterize the beginning of sections, the heuristics is now applicable to speech materials.

The importance of sentences is evaluated using the same function (equation (1)) that was used as appropriateness of discourse markers. For each sentence s_i, we compute a sum score $S_{DM}(s_i) = \sum_{w_j \in s_i} S_{DM}(w_j)$.

Then, key sentences are selected based on the score up to the specified number (or ratio) of sentences from the whole lecture.

4.4. Combination with Keyword-based Method

The other approach to extraction of key sentences is to focus on keywords that are characteristic to the lecture. The most orthodox statistical measure to define and extract such keywords is the following TF-IDF criterion.

$$S_{KW}(w_j) = tf_j * \log(\frac{N_d}{df_j}) \qquad (2)$$

Here, term frequency tf_j is the occurrence count of a word w_j in the lecture, and document frequency df_j is the number of lectures (=documents) in which the word w_j appears. N_d is the number of lectures used for normalization. For each sentence s_i, we compute $S_{KW}(s_i) = \sum_{w_j \in s_i} S_{KW}(w_j)$.

Then, we introduce a new measure of importance that combines it with the discourse marker-based method by taking a geometric mean with a weight α.

$$S_{final}(s_i) = S_{DM}(s_i)^{\alpha} \cdot S_{KW}(s_i)^{(1-\alpha)}$$

4.5. Experimental Evaluation

For part of the CSJ, key sentences labeled by human subjects are included. In this work, we made use of those available as of August 2003. A set of key sentences were labeled by three human subjects for 19 academic presentations. The subjects were instructed to select sentences which seemed important by 50% of all, and then 10% from those 50%.

We set up experiments based on the agreed portion of the 50% extraction data. Specifically, we picked up sets of sentences agreed upon by two subjects. Since three combinations exist for picking up two subjects out of three, we derived three answer sets. The performance is evaluated by averaging for these three sets. Using this scheme, we can also estimate the human performance by matching one subject's selection with the answer set derived from the other two. The recall, precision and F-measure are 83.2%, 62.7% and 0.715, respectively. These figures are regarded as a target for the proposed system.

The proposed method based on the discourse markers (DM) and its combination with the keyword-based method (KW) were evaluated. Indexing performance of the key sentences for correct transcriptions is listed in Table 4. The method using the discourse marker (DM) was comparable to the keyword-based method (KW), and the synergetic effect of their combination (DM+KW) was clearly confirmed. When we compare the system performance against the human judgment, the accuracy by the system is lower by about 10%. The proposed method performs reasonably, but it still has room for improvement.

Then, we made evaluation using the transcriptions generated by the ASR system. Since ASR results do not include periods, we incorporate the automatic period insertion procedure presented in Section 3.3 in order to segment the lecture into sentences. The indexing method is based on the discourse marker and keyword combination (DM+KW). Table 5 lists the recall, precision and F-measure in comparison with the case of manual transcription. Here, we also tested the case where the sentence segmentation or period insertion is done automatically on the manual transcriptions to see individual effects. It is observed that the automatic segmentation has a bad effect on the accuracy, especially on the precision. On the other hand, no degradation is observed by adopting automatic speech recognition regardless of the word error rate of 23%. These results demonstrate that the statistical evaluation of importance of the sentences is robust.

Table 4. Performance of key sentence indexing (text)

method	recall	precision	F-measure
DM	71.0%	53.3%	0.609
KW	71.7%	53.8%	0.614
DM+KW	74.0%	55.5%	0.635
human	83.2%	62.7%	0.715

DM: discourse marker, KW: keyword

Table 5. Performance of key sentence indexing (ASR results)

transcript	segment	recall	precision	F-measure
manual	manual	74.0%	55.5%	0.635
manual	automatic	73.1%	45.8%	0.563
automatic	automatic	72.7%	45.6%	0.561

5. CONCLUSIONS

The paper gave an overview of our archiving system of lectures, which consists of not only automatic transcription but also automatic editing and indexing of key sentences. It is shown that the large-scale corpus had strong impact in developing acoustic and language models for spontaneous speech. It is also confirmed that speaker adaptation of these models is very effective. The proposed method combining statistical measures of discourse markers and topic words realizes indexing of key sentences with the accuracy close to the human performance.

Ongoing work includes application of the method to other domains such as panel discussions and lectures at universities, and automatic annotation of more specific tags for a comprehensive digital archiving system.

Acknowledgments: The work was conducted as part of the Science and Technology Agency Priority Program on "Spontaneous Speech: Corpus and Processing Technology". The authors are grateful to Prof. Sadaoki Furui and other members of the project for their fruitful collaboration.

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] S.Furui. Recent advances in spontaneous speech recognition and understanding. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition,* pages 1–6, 2003.

[3] K.Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition,* pages 7–12, 2003.

[4] T.Kawahara and M.Hasegawa. Automatic indexing of lecture speech by extracting topic-independent discourse markers. In *Proc. IEEE-ICASSP*, pages 1–4, 2002.

[5] T.Kawahara, H.Nanjo, and S.Furui. Automatic transcription of spontaneous lecture speech. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001.

[6] T.Kawahara, H.Nanjo, T.Shinozaki, and S.Furui. Benchmark test for speech recognition using the Corpus of Spontaneous Japanese. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 135–138, 2003.

[7] 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.

[8] T.Kawahara, A.Lee, T.Kobayashi, K.Takeda, N.Minematsu, S.Sagayama, K.Itou, A.Ito, M.Yamamoto, A.Yamada, T.Utsuro, and K.Shikano. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proc. ICSLP*, volume 4, pages 476–479, 2000.

[9] K.Uchimoto, C.Nobata, A.Yamada, S.Sekine, and H.Isahara. Morphological analysis of Corpus of Spontaneous Japanese. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 159–162, 2003.

[10] H.Nanjo and T.Kawahara. Unsupervised language model adaptation for lecture speech recognition. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 75–78, 2003.

[11] H.Nanjo and T.Kawahara. Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. In *Proc. IEEE-ICASSP*, pages 725–728, 2002.

[12] H.Nanjo, K.Shitaoka, and T.Kawahara. Automatic transformation of lecture transcription into document style using statistical framework. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 215–218, 2003.

[13] S.Teufel and M.Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[14] T.Kawahara, K.Shitaoka, T.Kitade, and H.Nanjo. Automatic indexing of key sentences for lecture archives. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.