

Topic Segmentation and Retrieval System for Lecture Videos Based on Spontaneous Speech Recognition

Natsuo Yamamoto, Jun Ogata and Yasuo Arika

Faculty of Science and Technology, Ryukoku University

{ymmt, ogata}@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

Abstract

In this paper, we propose a segmentation method of continuous lecture speech into topics. A lecture includes several topics but it is difficult to judge their boundaries. To solve this problem, transcriptions obtained by spontaneous speech recognition of a lecture speech is associated with the textbook used in the lecture. This method showed high performance of the topic segmentation with an average of 93.7%. Incorporating this method, we constructed a system where we can view an interesting part of lecture videos, by specifying the chapters or sections as well as keywords.

1. Introduction

Due to computer progress and their wide spread in the world, various information has been accumulated in large quantities. However, it is still difficult to retrieve required information efficiently. Especially, in the educational field, systems are strongly required which can present the information on demand from accumulated data. In this paper, we propose a system to retrieve an interesting part of lecture videos efficiently on demand, based on automatic topic segmentation of a lecture speech, using spontaneous speech recognition techniques.

In lecture videos, a lecture proceeds according to a textbook. Usually one lecture contains several topics, each of which correspond to sections in the textbook. Therefore, one lecture can be segmented into topics using the textbook. A user can study the subject using the video and textbook so that, if the video is segmented into topics, the user can easily access and repeatedly view an interesting or inversely incomprehensive topics in the video.

However there are some problems in segmenting the lecture speech into topics. First, speech recognition is difficult in lecture due to disfluency such as repetition, mistakes and rephrasing caused by spontaneous characteristics, compared to news speech or address speech. Second, it is difficult to perform the topic segmentation at good accuracy because the cue is weak to detect the topic boundaries compared to news speech.

It may be possible to detect the topic boundaries by extracting the keywords signing the purpose or result because even the lecture proceeds according to some flow such as plan-do-see. Therefore it can detect boundaries between large topics, but it is difficult to detect the boundaries between small topics.

To solve these problems, in this paper, we propose following two methods. First, to improve the spontaneous speech recognition in lecture video, unsupervised adaptation of an acoustic model is employed and keyword indices in the textbook is added to the unknown word category in the language model. The keyword indices are the words judged to be important by the author of the textbook. Therefore, the accurate

recognition of the keyword indices is important for topic segmentation of the lecture speech.

Second, to improve the topic segmentation, the textbook used in the lecture is utilized. Since small topics are included in the textbook and the lecture follows it, the boundaries between small topics can be detected by associating the transcriptions obtained through spontaneous speech recognition of a lecture speech with the textbook used in the lecture.

In this paper, we describe the spontaneous speech recognition of the lecture speech in section 2. Next, in section 3, the topic segmentation by associating with the textbook is described along with TF-IDF and the vector space model which are basic technologies in topic segmentation. Finally, the experimental result and a topic retrieval system based on these techniques are shown in section 4.

2. Lecture speech recognition

2.1. Speech recognition system

As a baseline LVCSR system, we adopted 2-pass search strategy[1][2]. At the 1st-pass, a word graph is generated using the lexical tree search with bigram language model. Then, at the 2nd-pass, the best sentence is searched in the word graph using the acoustic score computed at the 1st-pass and trigram language model.

2.2. Acoustic model

A basic acoustic model is a context-dependent triphone and was trained using 21,783 sentences spoken by 137 males collected from JNAS (Japanese Newspaper Article Sentences) corpus [3]. Table 1 shows the experimental conditions for acoustic analysis.

In order to absorb the acoustic mismatch between training speech and real lecture speech, MLLR (Multiple linear regression) adaptation technique is employed[4]. At first, the input lecture speech is recognized using a basic acoustic model and the language model. Then the transcription obtained from the speech recognition is used for MLLR adaptation.

Table 1: Acoustic analysis and HMM

Sampling frequency	16kHz
Feature parameter	MFCC (39 dimensions)
Analysis frame length	20ms
Analysis frame shift	10ms
Analysis window	Hamming window

2.3. Language model

Language model was trained by 186 lecture speech transcriptions (549,612 words) available from CSJ (The Corpus of Spontaneous Japanese). CSJ is a spontaneous Japanese speech corpus collected under the Japanese national project [5]. Here, all the experiments were conducted using the bigram model. Vocabulary size of language model is 10K.

2.4. Keyword set

For the topic segmentation, the keywords are important to discriminate different topics. From this viewpoint, we put keyword indices listed at the end of the textbook into the unknown word category of the language model. The number of keyword indices depends on the lecture contents, varying from 259 to 594 for one lecture.

2.5. Experimental result

Two sets of lecture speech with about 40 minutes duration respectively were used for evaluation in the speech recognition. Here, each lecture speech was divided automatically by detecting sections with the power lower than some threshold. The divided speech data is called utterance hereafter. Therefore, one utterance does not necessarily correspond to meaningful and grammatically correct Japanese sentence.

The recognition experiment was carried out. The results are shown in Table 2 and Table 3. Table 2 shows the word recognition rate, namely word correct and word accuracy. On the other hand, Table 3 shows the correct rate of the keyword indices, namely the ratio of the correctly recognized keywords to the truly existing keywords.

The word recognition for the speech data Lec2 is lower than that of Lec1. This can be attributed to the unclearness and the lower power at the end of utterances in Lec2. The keywords important for topic segmentation achieved high correct rate about 91.1% at average, even there is about 5% difference between the speech data Lec1 and Lec2.

Table 2: Recognition result (%)

	Correct	Accuracy
Lec1	61.7	58.23
Lec2	50.5	42.03
Average	56.1	50.1

Table 3: Correct rate of keyword indices (%)

	Correct
Lec1	93.64
Lec2	88.73
Average	91.2

3. Method of topic segmentation

3.1. Overview

The recognition result of the lecture speech is a sequence of transcribed sentences (utterances). The purpose of the topic segmentation is to divide this sequence of recognized sentences into sections or chapters where the meaning is coherent by associating them with the textbook. For this purpose, a vector is created from respective section in the textbook. We call this vector a topic vector. Consequently, the topic vectors are created by the number of chapters and sections in the textbook.

The easiest way for the topic segmentation is to regard a sentence in the recognized speech as a vector. We call this vector a lecture vector here. Then the similarity is computed between a sequence of lecture vectors and the topic vectors. If a sequence of lecture vectors from the recognized speech is well

matched with some topic vector from the textbook, then the section or chapter is identified.

However, it is easily understood that a sentence is too short as a vector so that more longer section including several sentences is desired. We call this long section as an analysis section as shown in Figure1. The lecture vector from this analysis section is created by extracting nouns and computing their weight in the analysis section because nouns correspond to keyword indices and plays an important role to discriminate the topic differences.

In our experiment, the analysis section includes 10 sentences based on a preliminary experiment and shifted as shown in Figure1. For the morphological analysis to extract nouns, Chasen[6] was used. As for noun weight, TF-IDF was employed to reduce the falsely recognized nouns and unimportant nouns which appears all over the analysis sections. The similarity is computed between a lecture vector and topic vectors using a cosine function in a vector space. Then the topic vector is found with the highest similarity to a lecture vector and its section or chapter is associated. The details of TF-IDF and similarity computation are described in the following section.

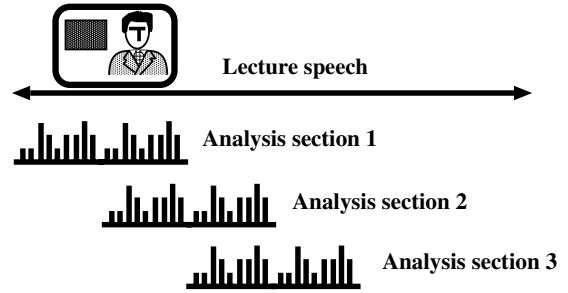


Figure 1: Analysis section for lecture speech

3.2. TF-IDF

TF-IDF is shown in Eq.(1). If the word w_i appears in document d_j frequently, then the TF(Term Frequency) increases. If the word w_i appears in the limited number of documents, the IDF(Inverse Document Frequency) increases. The high TF-IDF indicates the word w_i appears frequently in the specific documents. In this sense, the words with high TF-IDF are important as the keywords.

$$\begin{aligned}
 TF \cdot IDF &= TF(w_i, d_j) \cdot IDF(w_i) & (1) \\
 TF(w_i, d_j) &= \text{Frequency of word } w_i \text{ in document } d_j \\
 IDF(w_i) &= \log \frac{\text{The number of documents}}{\text{Number of documents including } w_i}
 \end{aligned}$$

3.3. Similarity computation in vector space

A vector space model is frequently used in the field of information retrieval. Table4 shows an example of matrix $A_{t \times d}$ whose columns and rows correspond to documents and keyword indices respectively. The value of the matrix element is TF-IDF of the keyword indices t_i in the document d_j .

When each column of A is regarded as a vector, these column vectors correspond to the document vectors \mathbf{d}_j represented in terms of keyword indices. The dimension of the document vector equals to the number of different words appearing in the

textbook or lecture speech. These document vectors \mathbf{d}_j can be presented in a vector space.

Similarly, a query vector \mathbf{q} can be presented in terms of TF-IDF of keyword indices. The similar $sim(\mathbf{d}_j, \mathbf{q})$ between a document vector \mathbf{d}_j and a query vector \mathbf{q} can be computed according to Eq.(2). It is well known that Eq.(2) shows an excellent retrieval performance [7]-[8].

$$\begin{aligned} sim(\mathbf{d}_j, \mathbf{q}) &= \frac{\mathbf{d}_j^t \mathbf{q}}{\|\mathbf{d}_j\| \|\mathbf{q}\|} \\ &= \frac{\sum_{i=1}^m (a_{ij} q_i)}{\sqrt{\sum_{i=1}^m (a_{ij})^2} \sqrt{\sum_{i=1}^m (q_i)^2}} \quad (2) \end{aligned}$$

Table 4: word×document matrix A

	d_1	d_2	...	d_d
t_1	a_{11}	a_{12}	...	a_{1n}
t_2	a_{21}	a_{22}	...	a_{2n}
⋮	⋮	⋮	⋮	⋮
t_t	a_{m1}	a_{m2}	...	a_{mn}

3.4. Post-processing

Figure 2 shows the topic segmentation process by associating a lecture speech with the textbook. However, in this method, the order of the lecture vectors are disregarded so that following wrong association is caused.

1. The first lecture vector is associated with section2 in the textbook.
2. There is a lecture vector associated with section4 just after the lecture vector associated with section2.
3. There is a lecture vector associated with section2 just after the lecture vector associated with section3.

These wrong association suggests that if the i th lecture vector is associated with sections s_i in the textbook, the succeeding $i + 1$ th lecture vector has to be associated with sections s_i or sections $s_i + 1$ in the textbook. According to this suggestion, following post-processing is carried out after the basic association between a sequence of lecture vectors and topic vectors. Figure 3 shows an example of this post-processing.

- (1) If vector $i = 1$, then its section $s_i = 1$.
- (2) If section $s_{i+1} \neq s_i$ and section $s_{i+1} \neq s_i + 1$ then section $s_{i+1} = s_i$.(regarding noises)
- (3) In a case that section $s_{i+1} = s_i + 1$ and $s_i + 1$ continues for m successive vectors and section s_i continues for further n successive vectors, if $m < n$, then section $s_i + 1$ for m vectors is changed to section s_i .

4. Experimental result

4.1. Experimental conditions

We carried out topic segmentation experiments to the speech recognition result described in section2. The result of morphological analysis is shown in Table 5. In the table, "Number of different words" indicates the number of words after removing the repetition from the total number of words used in the lecture speech. Table 6 shows the characteristics of the lecture

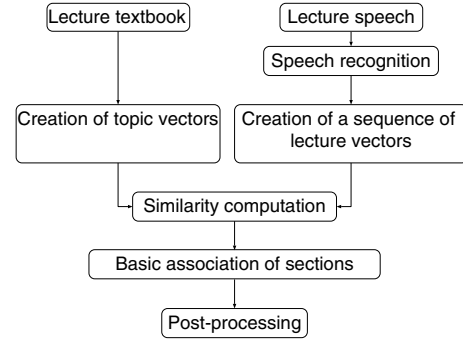


Figure 2: Association process between speech and textbook

Table 5: Number of noun words in lecture speech

	Number of words	Number of different words
Lec1	2423	697
Lec2	2104	839

Table 6: Lecture textbook

	Words	Different words	Chapters	Sections
Lec1	1179	459	4	13
Lec2	1180	626	4	10

textbook; number of words, number of different words, number of chapters and sections.

The experimental results was evaluated by "correct rate" shown in Eq.(3) before and after the post-processing described in section3.4. True label data was manually investigated by listening to the lecture speech and reading the textbook.

Correct rate=

$$\frac{\text{Number of lecture vectors correctly associated}}{\text{Number of all lecture vectors}} \quad (3)$$

4.2. Experimental result

Table7 and Table8 show the result of association with the chapters and sections in the textbook respectively. In the table, (a) and (b) show the result before and after post-processing respectively.

In total, high correct rate was achieved, especially it was 98.3% in chapter association. On the other hand, the correct rate in section association was 89.0%, a little lower than the chapter. It can be explained that the feature of the section was lower than the chapter as a topic because the number of words used in a section is less than those in a chapter.

4.3. Retrieval and browsing system

Based on the described speech recognition and topic segmentation, We constructed a retrieval and browsing system of lecture videos. This system is composed of the following three components. Figure4 shows an interface windows including three components.

- Video display
- Contents browser

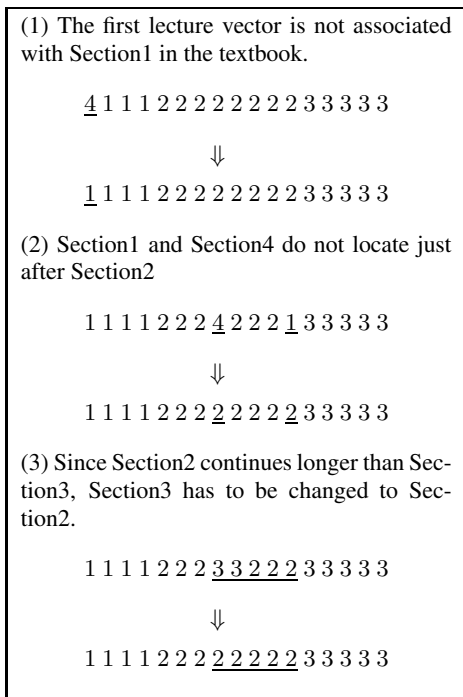


Figure 3: Example of post-processing

- Retrieval window

The video display is a window where a specified lecture video is displayed. In the contents browser, table of contents in the textbook is hierarchically displayed. When a user select the first lecture, the chapters included in the first lecture are displayed. Furthermore, when he selects Chapter2 among them, the sections included in Chapter2 are displayed. He can view a part of the lecture video corresponding to the selected section.

In the retrieval window, the keyword indices included in the lecture textbook are listed. These indices are internally linked to the corresponding section. Therefore, when the user selects one of the indices, he can view the video where the specified index is explained.

Table 7: Result of chapter association

(a) Before post-processing		(b) After post-processing	
	Correct(%)		Correct(%)
Lec1	80.2	Lec1	98.6
Lec2	88.0	Lec2	97.9
Total	84.1	Total	98.3

Table 8: Result of section association

(a) Before post-processing		(b) After post-processing	
	Correct(%)		Correct(%)
Lec1	60.4	Lec1	91.1
Lec2	74.9	Lec2	86.9
Total	67.7	Total	89.0



Figure 4: Interface window in a retrieval system

5. Conclusion

In this paper, we proposed a method for topic segmentation of lecture videos by associating lecture speech with the lecture textbook. The association was performed by computing the similarity between topic vectors and a sequence of lecture vectors obtained through spontaneous speech recognition. The result of the topic segmentation showed high accuracy, 98.3% in the association with chapters in the textbook. We also constructed a system to retrieve and browse the lecture video by specifying the section, chapter or keyword indices.

In future, we are planning to study topic segmentation without the textbook and to increase the number of lecture videos to retrieve and browse in the constructed system.

6. References

- [1] S.Ortmanns, H.Ney, X.Aubert: "A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition", Computer Speech and Language, Vol.11, No.1, pp.43-72(1997).
- [2] J.Ogata and Y.Ariki: "An Efficient Lexical Tree Search for Large Vocabulary Continuous Speech Recognition" ICSLP'2000, Vol.II, pp.967-970 (2000-10).
- [3] K.Itou, M.Yamamoto, K.Takeda, T.Takezawa, T.Matsuoka, T.Kobayashi, K.Shikano and S.Itahashi: "The Design of the Newspaper-Based Japanese Large Vocabulary Continuous Speech Recognition Corpus", Proc. ICSLP'98, 1998.
- [4] C.L.Leggetter and P.C.Woodland: "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol.9, pp.171-185, 1995.
- [5] S.Furui, K.Maekawa, H.Isahara, T.Shinozaki and T.Ohdaira: "Toward the Realization of Spontaneous Speech Recognition - Introduction of a Japanese Priority Program and Preliminary Results-", Proc. ICSLP'2000, pp.518-521, (2000).
- [6] Chasen: <http://chasen.aist-nara.ac.jp/>
- [7] G.Salton, A.Wong and C.Yang: "A Vector Space Model for Information Retrieval", Journal of the ASIS, pp.613-620, November 1975.
- [8] G.Salton: "Automatic Text Processing", Addison-Wesley, Reading, Massachusetts, 1989.