A System for On-demand Video Lectures

Atsushi Fujii[†,†††] **Katunobu Itou**[††,†††] **Tetsuya Ishikawa**[†]
[†] Institute of Library and Information Science
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
[††] National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini Umezono, Tsukuba, 305-8568, Japan
[†††] CREST, Japan Science and Technology Corporation
fujii@slis.tsukuba.ac.jp

Abstract

We propose a lecture-on-demand system, which searches lecture videos for segments relevant to user information needs. We utilize the benefits of textbooks and audio/video data corresponding to a single lecture. Our system extracts the audio track from a target lecture video, generates a transcription by large vocabulary continuous speech recognition, and produces a textual index. Users can selectively view specific video segments by submitting textual queries associated with the textbook for the target lecture. Experimental results showed that by adapting speech recognition to the lecture topic, the recognition accuracy increased and the retrieval accuracy was comparable with that obtained by human transcriptions. Our system is implemented as a client-server system over the Web to facilitate e-education.

Introduction

Given the growing number of multimedia contents available via the World Wide Web, CD-ROMs, and DVDs, information technologies across speech, image, and text processing have of late become crucial. Among various types of contents, lectures (audio/video) are very typical and valuable multimedia contents, in which speeches (i.e., oral presentations) are usually organized based on textual materials, such as resumes, slides, and textbooks. In lecture videos, image information, such as flip charts, is often additionally used. In other words, a single lecture consists of different types of compatible multimedia contents.

However, since a single lecture often includes multiple stories and takes long time, it is useful to selectively obtain specific segments (passages) so that audience can satisfy their information needs with a minimal cost. To resolve this problem, in this paper we propose a lecture-on-demand system, which retrieves relevant video/audio passages in response to user queries. For this purpose, we utilize the benefits of different media types to improve retrieval performance.

On the one hand, textual contents are advantageous in the sense that users can view/scan the entire contents quickly and easily identify relevant passages using layout information (e.g., text structures based on sections and paragraphs). In other words, textual contents can be used for random-access purposes.

On the other hand, speech contents are fundamentally used for sequential-access purposes. Thus, it is difficult to identify relevant passages unless target video/audio data include additional annotations, such as indexes. Even if target data are indexed, users are not necessarily able to come up with effective queries. To resolve this problem, textbooks are desirable materials, from which users can extract effective keywords and phrases.

However, while textbooks are usually concise, speeches are relatively redundant and thus are easy to understand more than textbooks, specifically in the case where additional image information is provided.

In view of the above discussion, we model our lecture-on-demand (LOD) system as follows. A user selects text segments (i.e., keywords, phrases, sentences, and paragraphs) relevant to their information needs, from a textbook for a target lecture. By using selected segments, a textual query is automatically generated. In other words, queries can be formulated even if users cannot come up with effective keywords and phrases. Users can also submit additional keywords as queries, if necessary. Video passages relevant to a given query are retrieved and presented to the user.

To retrieve video passages in response to textual queries, we extract the audio track from a lecture video, generate a transcription by means of large vocabulary continuous speech recognition, and produce a textual index, prior to the system usage.

Our on-demand system should not be confused with video-on-demand (VOD) systems, which search video archives for specific videos in response to user requests. While in VOD systems, minimal unit for retrieval is the entire program, in our system, retrieval units are passages smaller than the entire program.

System Description

Overview

Figure 1 depicts the overall design of our lecture-on-demand system, in which left/right-hand regions correspond to the on-line and off-line processes, respectively. Although our system is currently implemented for Japanese, our methodology is fundamentally language-independent. For the pur-

pose of research and development, we tentatively target lecture programs on TV for which textbooks are published. We explain the basis of our system using this figure.

In the off-line process, given the video data of a target lecture, audio data are extracted and segmented into more than one passage. Then, speech recognition transcribes each passage. Finally, transcribed passages are indexed as performed in conventional text retrieval systems, so that each passage can be retrieved efficiently in response to textual queries.

To adapt speech recognition to a specific lecturer, we perform unsupervised speaker adaptation using an initial speech recognition result (i.e., a transcription).

To adapt speech recognition to a specific topic, we perform language model adaptation, for which we search a general corpus for documents relevant to the textbook related to a target lecture. Then, retrieved documents (i.e., a topic-specific corpus) are used to produce a word-based N-gram language model.

We also perform image analysis to extract textual contents (e.g., keywords and phrases) in flip charts. These contents are also used later to improve our language model.

In the on-line process, a user can view specific video passages by submitting any textual queries, i.e., keywords, phrases, sentences, and paragraphs, extracted from the textbook. Any queries not in the textbook can also additionally be used. The current implementation is based on a client-server system over the Web. While both the off-line and on-line processes are performed on servers, users can utilize our system with Web browsers on their own PCs.

Figure 2 depicts a prototype interface of our LOD system, in which a lecture associated with "nonlinear multivariate analysis" is given. In this interface, an electronic version of a textbook is displayed in the left-hand side, and a lecture video is played in the right-hand side. In addition, users can submit any textual queries to the box in the bottom of the interface. The operation is similar to that for existing Web search engines.

In Figure 3, a text paragraph related to "discriminant analysis" is copied and pasted into the query input box. It should be noted that unlike conventional keyword-based retrieval systems, in which users usually submit a small number of keywords, in our system users can easily submit longer queries using textbooks. In the case where submitted keywords are misrecognized in transcriptions, the retrieval accuracy decreases. However, longer queries are relatively robust for speech recognition errors, because the effect of misrecognized words are overshadowed by a large number of words correctly recognized.

Figure 4 depicts retrieval results, in which top-ranked transcribed passages for the query in Figure 3 are listed according to the degree of relevance. Users can select (click) transcriptions to play the corresponding video passage. We explain each module in the following three sections.

Passage Segmentation

The basis of passage segmentation is to divide the entire video data for a single lecture into more than one minimal unit to be retrieved. We shall call those units passages.

For this purpose, both speech and image data can be promising clues. For example, Hamada et al. (2000) performed a structural analysis on cooking TV programs by means of speech/image/text processing. However, in lecture TV programs, it is often the case that a lecturer sitting still is mainly focused and a small number of flip charts are occasionally used. In such cases, image data is less informative. Thus, we tentatively use only speech data for the passage segmentation process.

However, unlike the case where a target speech (e.g., a news program) consists of multiple different topics (Allan 2002; Takao, Ogata, & Ariki 2000), topic segmentation for lectures is relatively difficult, because a single lecture consists of sub-topics closely related to one another, and thus topic boundaries are not necessarily clear.

Existing methods to segment written texts (e.g., one proposed by Hearst (1997)) rely only on lexical information in texts, and thus are not robust against errors in automatic transcriptions. Additionally, in our LOD system, segmentation can potentially vary depending on the user query. Thus, it is difficult to predetermine a desirable segmentation in the off-line process.

In view of the above problems, we first extract the audio track from a target video, and perform a simple pause-based segmentation method to obtain minimal speech units, such as sentences and long phrases. In other words, speech units are continuous audio segments that do not include pauses longer than a certain threshold. Finally, we generate variable-length passages from one or more speech units. To put it more precisely, we combine N speech units into a single passage, with N ranging from 1 to 5 in the current implementation.

Figure 5 shows an example of variable-length passages, in which any sequences of speech units that are 1-5 in length are identified as passages.

Speech Recognition

The speech recognition module generates word sequence W, given phone sequence X. In a stochastic speech recognition framework (Bahl, Jelinek, & Mercer 1983), the task is to select the W maximizing $P(W|X)$, which is transformed as in Equation (1) through the Bayesian theorem.

$$\arg\max_{W} P(W|X) = \arg\max_{W} P(X|W) \cdot P(W) \qquad (1)$$

$P(X|W)$ models a probability that word sequence W is transformed into phone sequence X, and $P(W)$ models a probability that W is linguistically acceptable. These factors are called acoustic and language models, respectively.

We use the Japanese dictation toolkit (Kawahara *et al.* 2000)[1], which includes the Julius decoder and acoustic/language models. Julius performs a two-pass (forward-backward) search using word-based forward bigrams and backward trigrams.

The acoustic model was produced from the ASJ speech database (Itou *et al.* 1998), which contains approximately 20,000 sentences uttered by 132 speakers including the both

[1] http://winnie.kuis.kyoto-u.ac.jp/dictation/

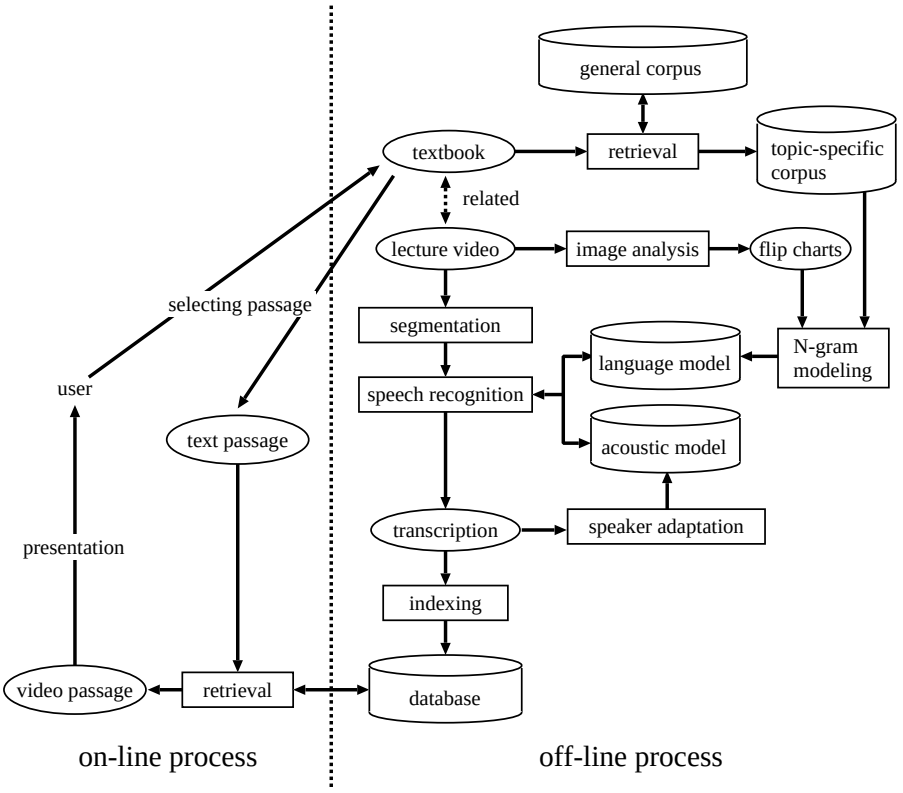

Figure 1: The overview of our lecture-on-demand system.

gender groups. A 16-mixture Gaussian distribution triphone Hidden Markov Model, in which states are clustered into 2,000 groups by a state-tying method, is used. We adapt the provided acoustic model by means of an MLLR-based unsupervised speaker adaptation method (Leggetter & Woodland 1995), for which in practice we use the HTK toolkit[2].

Existing methods to adapt language models can be classified into two fundamental categories. In the first category, the *integration* approach, general and topic-specific (domain-specific) corpora are integrated to produce a topic-specific language model (Auzanne *et al.* 2000; Seymore & Rosenfeld 1997). Since the sizes of those corpora are different, N-gram statistics are calculated by the weighted average of statistics extracted independently from those corpora. However, it is difficult to determine the optimal weight depending on the topic.

In the second category, the *selection* approach, a topic-specific subset is selected from a general corpus and is used to produce a language model. This approach is effective if general corpora contain documents associated with target topics, but N-gram statistics in those documents are overshadowed by other documents in resultant language models.

We followed the selection approach, because the 10M Web page corpus (Eguchi *et al.* 2002)[3] containing mainly Japanese pages associated with various topics was available to the public. The quality of the selection approach is dependent of the method to select topic-specific subsets. An existing method (Chen *et al.* 2001) uses hypotheses in the initial speech recognition phase as queries to retrieve topic-specific documents from a general corpus. However, errors in the initial hypotheses potentially decrease the retrieval accuracy. Instead, we use textbooks related to target lectures as queries to improve the retrieval accuracy and consequently the quality of language model adaptation.

Retrieval

Given transcribed passages and textual queries, the basis of the retrieval module is the same as conventional text retrieval. We use an existing probabilistic text retrieval method (Robertson & Walker 1994) to compute the relevance score between the query and each passage in the database. The relevance score for passage p is computed by Equation (2).

$$\sum_t f_{t,q} \cdot \frac{(K+1) \cdot f_{t,p}}{K \cdot \{(1-b) + \frac{dl_p}{b \cdot avgdl}\} + f_{t,p}} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5} \tag{2}$$

Here, $f_{t,q}$ and $f_{t,p}$ denote the frequency that term t appears in query q and passage p, respectively. N and n_t denote the total number of passages in the database and the number of passages containing term t, respectively. dl_p denotes the length of passage p, and $avgdl$ denotes the average length of

[2]http://htk.eng.cam.ac.uk/

[3]http://research.nii.ac.jp/ntcir/index-en.html

Figure 2: The interface of our LOD system over the Web.

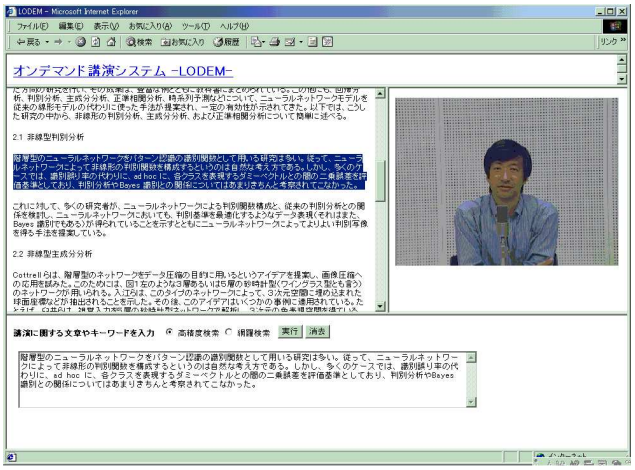

Figure 3: An example scenario, in which a paragraph in the textbook is copied and pasted into the query input box.

Figure 4: Example retrieved transcriptions.

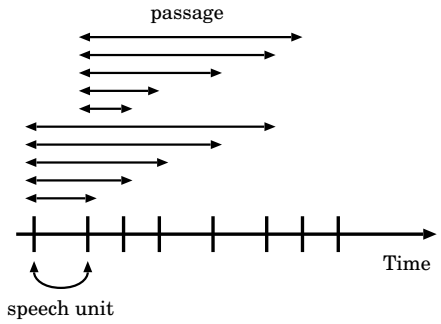

Figure 5: An example of the passage segmentation process.

passages in the database. We empirically set $K = 2.0$ and $b = 0.8$, respectively.

It should be noted that in Equation (2), the score is normalized with the length of each passage. Thus, longer passages, which potentially include more index terms, are not necessarily assigned with a higher score. This property is important, because variable-length passages are considerably different in terms of length.

We use content words, such as nouns, extracted from transcribed passages as index terms, and perform word-based indexing. We use the ChaSen morphological analyzer[4] to extract content words. We also extract terms from queries using the same method.

However, retrieved passages are not disjoint, because top-ranked passages often overlap with one another in terms of the temporal axis. It is redundant to simply list the top-ranked retrieved passages as they are. Thus, we reorganize those overlapped passages into a single passage. In Figure 6, which uses the same basic notation as Figure 5, illustrates an example scenario. In this figure, top-ranked passages are organized into three groups.

The relevance score for a group (a new passage) is computed by averaging scores for all passages belonging to the group. New passages are sorted according to the degree of relevance and are presented to users as the final result.

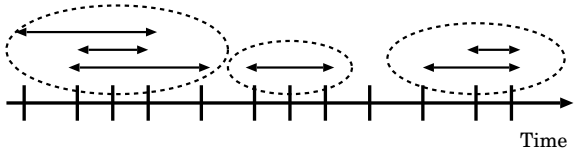

Figure 6: An example of grouping retrieved passages.

Experimentation

Methodology

To evaluate the performance of our LOD system, we produced a test collection (a kind of benchmark data set) and performed experiments partially resembling one performed

[4] http://chasen.aist-nara.ac.jp/

in the TREC spoken document retrieval (SDR) track (Garofolo *et al.* 1997).

Two lecture programs on TV, for which printed textbooks were also published, were videotaped in DV and were used as target lectures. Both lectures were manually transcribed and sentence boundaries with temporal information (i.e., correct speech units) were also manually identified. The textbooks for the two target lectures were read by an OCR software and were manually revised. The accuracy of the OCR software was roughly 97% on a word-by-word basis.

For both lectures, each paragraph in the corresponding textbook was used as a query independently. For each query, a human assessor (a graduate student other than authors of this paper) identified one or more relevant sentences in the human transcription.

Table 1 shows details of our test collection, in which lectures #1 and #2 were associated with the criminal law and histories of ancient Greece, respectively. Each lecture was 45 minutes long. In this table, we shall use the term "word token" to refer to occurrences of words, and the term "word type" to refer to vocabulary items. The column "# of Fillers" denoting the number of interjections in speech partially shows the fluency of each lecturer.

Table 1: Details of our test collection used for experiments.

ID	#1	#2
Topic	Law	History
# of Word tokens in lecture	6917	8092
# of Word types in lecture	1029	1219
# of Fillers in lecture	3	953
# of Sentences in lecture	181	191
# of Queries	25	13
Avg. # of relevant sentences per query	7.6	6.8
Avg. length of queries (Avg. # of words)	154	247

By using our test collection, we evaluated the accuracy of speech recognition and passage retrieval. It may be argued that passage segmentation should also be evaluated. However, to evaluate the extent to which the accuracy of passage segmentation affects the entire system performance, relevance assessment for passage retrieval has to be performed for multiple segmentations, which is expensive.

For both target lectures, our system used the sentence boundaries in human transcriptions to identify speech units, and performed speech recognition. We also used human transcriptions as perfect speech recognition results and investigated the extent to which speech recognition errors affect the retrieval accuracy. Our system retrieved top-ranked passages in response to each query. It should be noted that passages here are those grouped based on the temporal axis, which should not be confused with those obtained in the passage segmentation method.

For lecture #1, we adapted the acoustic model to the lecturer by means of the MLLR-based method. However, for lecture #2 we did not perform acoustic model adaptation, because the speech data contained constant background noise and the sound quality was not good enough to adapt the acoustic model. For both lectures #1 and #2, we did not use flip chart information obtained by means of image analysis.

Results

To evaluate the accuracy of speech recognition, we used word error rate (WER), which is the ratio between the number of word errors (i.e., deletion, insertion, and substitution) and the total number of words. We also used test-set out-of-vocabulary rate (OOV) and trigram test-set perplexity (PP) to evaluate the extent to which our language model was adapted to target topics.

We used human transcriptions as test set data. For example, OOV is the ratio between the number of word tokens not contained in the language model for speech recognition and the total number of word tokens in the transcription. It should be noted that smaller values of OOV, PP, and WER are obtained with better methods.

The final outputs (i.e., retrieved passages) were evaluated based on recall and precision, averaged over all queries. Recall (R) is the ratio between the number of correct speech units retrieved by our system and the total number of correct speech units for the query in question. Precision (P) is the ratio between the number of correct speech units retrieved by our system and the total number of speech units retrieved by our system. To summarize recall and precision into a single measure, we used F-measure (F), which is calculated by Equation (3).

$$\frac{(\beta^2+1)\cdot \mathrm{R}\cdot \mathrm{P}}{\beta^2\cdot \mathrm{R}+\mathrm{P}} \tag{3}$$

Here, β is a parametric constant used to control the preference between recall and precision. In our case, recall and precision were equally important, and thus we set $\beta = 1$.

Table 2 shows the accuracy of speech recognition (WER) and passage retrieval (R, P, and F), for each lecture. In this table, the columns "HUM" and "ASR" correspond to the results obtained with human transcriptions and automatic speech recognition, respectively. The results for ASR are also divided into those obtained with/without language model adaptation (LA).

To adapt language models, we used the textbook corresponding to a target lecture and searched the 10M Web page corpus for 2,000 relevant pages, which were used as a source corpus. In the case where the language model adaptation was not performed, all 10M Web pages were used as a source corpus. In either case, 20,000 high frequency words were selected from a source corpus to produce word-based trigram language model. We used the ChaSen morphological analyzer to extract words (morphemes) from source corpora, because Japanese sentences lack lexical segmentation.

In passage retrieval, we regarded the top N passages as the final outputs. In Table 2, the value of N ranges from 1 to 3. As the value of N increases, the recall improves, but potentially sacrificing the precision.

Discussion for Speech Recognition

By comparing the results of ASR with/without LA in Table 2, OOV, PP, and WER decreased by adapting language

models to target topics, irrespective of the lecture. This suggests that our language model adaptation method was effective to improve the quality of speech recognition.

The values of OOV, PP, WER for lecture #2 were generally greater than those for lecture #1. One possible rationale is that the lecturer of #1 spoke more fluently and the number of erroneous utterances were smaller, when compared with the lecturer of #2. This tendency was partially observable in the column "# of Fillers in lecture" of Table 1. Additionally, the acoustic model was not adapted to the lecturer of #2, because the sound quality of the speech data for lecture #2 was not good enough to perform acoustic model adaptation.

Discussion for Passage Retrieval

By comparing the results of ASR with/without LA in Table 2, recall, precision, and F-measure increased by adapting language models to the topic of lecture #2, irrespective of the number of passages retrieved. This suggests that our language model adaptation method was effective to improve the retrieval accuracy.

For lecture #1, the retrieval accuracy did not significantly differ whether or not we adapted the language model to the topic. One possible rationale is that WER of lecture #1 without language model adaptation (20.9%) was small enough to obtain the retrieval accuracy comparable with text retrieval (Jourlin *et al.* 2000). In fact, the difference between HUM and ASR was marginal in terms of the retrieval accuracy. Thus, the effect of the language model adaptation method was overshadowed in passage retrieval.

Surprisingly, for lecture #2, recall, precision, and F-measure of ASR with LA were better than those of HUM except for the case of $N = 3$. In other words, the automatic transcription was more effective than the human transcription for passage retrieval purposes.

One possible rationale is associated with Japanese variants (i.e., more than one spelling form corresponding to the same word), such as "*girisha*/*girishia* (Greece)." Since the language model was adapted by means of the textbook corresponding to a target lecture, the spelling in automatic transcriptions systematically resembled one in queries extracted from textbooks. In contrast, it is difficult to standardize the spelling in human transcriptions. Thus, relevant passages in automatic transcriptions were retrieved more likely than passages in human transcriptions.

For all cases, recall was better than precision. This is attributed to our retrieval method. Since passages (one or more sentences) obtained by the initial phase were grouped into a single passage based on the temporal axis, irrelevant sentences were often contained in the retrieval results.

The retrieval accuracy for lecture #1 was generally higher than those for lecture #2. While the story of lecture #1 was organized based primarily on the textbook, the story of lecture #2 was relatively independent of the contents in the textbook. This suggests that the performance of our LOD system is dependent of the organization of target lectures.

At the same time, since our test collection includes only two lectures, experiments using larger test collections associated with various topics should be further explored.

Table 2: Experimental results for speech recognition (OOV: test-set out-of-vocabulary rate, PP: trigram test-set perplexity, WER: word error rate) and passage retrieval (N: # of passages retrieved, R: recall, P: precision, F: F-measure).

ID		#1			#2		
			ASR			ASR	
		HUM	w/o LA	w/ LA	HUM	w/o LA	w/ LA
OOV		—	.0444	.0203	—	.0729	.0821
PP		—	48.91	43.27	—	122.1	96.69
WER		—	.2088	.1335	—	.5161	.4232
N=1	R	.6947	.7263	.7316	.4494	.2584	.5506
	P	.5344	.5476	.5187	.3774	.3194	.3858
	F	.6041	.6244	.6070	.4103	.2857	.4537
N=2	R	.8474	.8579	.8316	.6629	.3596	.6742
	P	.4411	.4478	.4580	.3010	.2105	.3141
	F	.5802	.5884	.5907	.4140	.2656	.4286
N=3	R	.8789	.8684	.8737	.7640	.4382	.7079
	P	.4103	.4054	.4010	.2688	.1625	.2520
	F	.5595	.5528	.5497	.3977	.2371	.3717

Related Work

Informedia (Hauptmann & Witbrock 1997) retrieves video passages from TV news programs in response to textual queries, for which users have to type the entire queries. This feature is problematic in the case where users have difficulty formulating effective queries. However, in our case, users can utilize segments of the textbook associated with a lecture as queries even if they cannot come up with effective keywords and phrases.

Hamada et al. (2000) performed structural analysis on cooking TV programs by means of speech/image/text processing, in which the textbook for a program was additionally used. However, while they focused mainly on analyzing video data, we intended to retrieve video passages.

Unlike our study in this paper, in the above two cases no quantitative experimental results were shown with respect to the accuracy of retrieving video data. Thus, it is difficult to compare the performance of our system with those for those existing systems.

Spoken document retrieval (SDR), in which textual queries are used to search speech archives for relevant information, is primarily related to our research. Initiated partially by the TREC-6 SDR track (Garofolo *et al.* 1997), various SDR methods targeting broadcast news have been proposed (Johnson *et al.* 1999; Jones *et al.* 1996; Sheridan, Wechsler, & Schäuble 1997). State-of-the-art SDR methods, with WER being approximately 20%, are comparable with text retrieval methods in performance (Jourlin *et al.* 2000), and thus are already practical.

However, as shown in Table 2 (lecture #2), the speech recognition accuracy for lectures was not necessarily high when compared with broadcast news. While the TREC conference concluded that SDR in English was a solved problem, SDR for lectures remains unsolved and should be further explored, specifically for languages other than English.

Segmenting lectures into passages is associated with

the Topic Detection and Tracking (TDT) evaluation workshop (Allan 2002), in which one task is to segment a single broadcast news stream into topically coherent stories. However, in the case of lectures, stories in a single lecture are closely related to one another, and therefore topic segmentation is more difficult than that for broadcast news programs.

Our research is also associated with speech summarization (Hori & Furui 2000), because a specific number of passages extracted from the entire speech data are organized so that users can understand important contents with a minimal cost. However, unlike existing methods for generic summaries, our method is classified as a query-biased (user-focused) summarization (Mani & Bloedorn 1998; Tombros & Sanderson 1998), in which different summaries are generated depending on the user information needs.

Finally, our research is crucial for e-education purposes, in which educational contents, such as lecture video/audio data are provided in real-time over computer networks. For example, in the WIDE University, School of Internet[5], lecture video data manually synchronized with presentation slides are available to the public over the Web.

Jones and Edens (2002) proposed a system to automatically synchronize an audio track with presentation slides, which is expected to reduce a cost required for manual indexing. Their system is similar to our system, because textual materials (slides and textbooks) are used to identify corresponding passages in a presentation. However, while their system was mainly intended to match transcriptions with slides, we also addressed problems in adapting language models for speech recognition, and showed its effectiveness by means of experiments.

Conclusion

Reflecting the rapid growth in the utilization of multimedia contents, information technologies across speech, image, and text processing are crucial. Among various content types, in this paper we focused video data of lectures organized based on textbooks, and proposed a system for on-demand lectures, in which users can formulate textual queries using the textbook for a target lecture to retrieve specific video passages.

To retrieve video passages in response to textual queries, we extract the audio track from a lecture video, generate a transcription by large vocabulary continuous speech recognition, and produce a textual index, prior to the system usage. The current system is implemented as a server-client system over the Web to facilitate e-education.

We also evaluated the performance of our system by means of experiments, for which two TV lecture programs were used. The experimental results showed that the accuracy of speech recognition varied depending on the domain and presentation style of lectures. However, the accuracy of speech recognition and passage retrieval was improved by adapting language models to the topic of a target lecture. In addition, even if the word error rate was approximately 40%, the accuracy of retrieval was comparable with that obtained by human transcriptions.

[5]http://www.soi.wide.ad.jp/

Future work will include improvement of each component in our system and extensive experiments using larger test collections related to various domains.

References

Allan, J., ed. 2002. *Topic Detection and Tracking: Event-based News Organization*. Kluwer Academic Publishers.

Auzanne, C.; Garofolo, J. S.; Fiscus, J. G.; and Fisher, W. M. 2000. Automatic language model adaptation for spoken document retrieval. In *Proceedings of RIAO 2000 Conference on Content-Based Multimedia Information Access*.

Bahl, L. R.; Jelinek, F.; and Mercer, R. L. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5(2):179–190.

Chen, L.; Gauvain, J.-L.; Lamel, L.; Adda, G.; and Adda, M. 2001. Language model adaptation for broadcast news transcription. In *Proceedings of ISCA Workshop on Adaptation Methods For Speech Recognition*.

Eguchi, K.; Oyama, K.; Kuriyama, K.; and Kando, N. 2002. The Web retrieval task and its evaluation in the third NTCIR workshop. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 375–376.

Garofolo, J. S.; Voorhees, E. M.; Stanford, V. M.; and Jones, K. S. 1997. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, 83–91.

Hamada, R.; Ide, I.; Sakai, S.; and Tanaka, H. 2000. Associating cooking video with related textbook. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, 237–241.

Hauptmann, A. G., and Witbrock, M. J. 1997. Informedia: News-on-demand multimedia information acquisition and retrieval. In Maybury, M. T., ed., *Intelligent Multimedia Information Retrieval*. AAAI Press/MIT Press. 215–239.

Hearst, M. A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64.

Hori, C., and Furui, S. 2000. Automatic speech summarization based on word significance and linguistic likelihood. In *Proceedings of ICASSP2000*, 1579–1582.

Itou, K.; Yamamoto, M.; Takeda, K.; Takezawa, T.; Matsuoka, T.; Kobayashi, T.; Shikano, K.; and Itahashi, S. 1998. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proceedings of the 5th International Conference on Spoken Language Processing*, 3261–3264.

Johnson, S.; Jourlin, P.; Moore, G.; Jones, K. S.; and Woodland, P. 1999. The Cambridge University spoken document retrieval system. In *Proceedings of ICASSP'99*, 49–52.

Jones, G. J. F., and Edens, R. J. 2002. Automated alignment and annotation of audio-visual presentations. In *Pro-*

ceedings of the 6th European Conference on Development for Digital Libraries, 276–291.

Jones, G.; Foote, J.; Jones, K. S.; and Young, S. 1996. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 30–38.

Jourlin, P.; Johnson, S. E.; Jones, K. S.; and Woodland, P. C. 2000. Spoken document representations for probabilistic retrieval. *Speech Communication* 32:21–36.

Kawahara, T.; Lee, A.; Kobayashi, T.; Takeda, K.; Minematsu, N.; Sagayama, S.; Itou, K.; Ito, A.; Yamamoto, M.; Yamada, A.; Utsuro, T.; and Shikano, K. 2000. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proceedings of the 6th International Conference on Spoken Language Processing*, 476–479.

Leggetter, C., and Woodland, P. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language* 9:171–185.

Mani, I., and Bloedorn, E. 1998. Machine learning of generic and user-focused summarization. In *Proceedings of AAAI/IAAI-98*, 821–826.

Robertson, S., and Walker, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 232–241.

Seymore, K., and Rosenfeld, R. 1997. Using story topics for language model adaptation. In *Proceedings of Eurospeech97*, 1987–1990.

Sheridan, P.; Wechsler, M.; and Schäuble, P. 1997. Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 99–108.

Takao, S.; Ogata, J.; and Ariki, Y. 2000. Topic segmentation of news speech using word similarity. In *Proceedings of the Eighth ACM International Conference on Multimedia*, 442–444.

Tombros, A., and Sanderson, M. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2–10.