Developing an SDR Test Collection from Japanese Lecture Audio Data

Tomoyosi Akiba*, Kiyoaki Aikawa†, Yoshiaki Itoh‡, Tatsuya Kawahara§, Hiroaki Nanjo¶, Hiromitsu Nishizaki‖, Norihito Yasuda**, Yoichi Yamashita††, and Katunobu Itou‡‡
* Toyohashi University of Technology
† Tokyo University of Technology
‡ Iwate Prefectural University
§ Kyoto University
¶ Ryukoku University
‖ University of Yamanashi
** Nippon Telegraph and Telephone Corporation
†† Ritsumeikan University
‡‡ Hosei University

***Abstract*—The lecture is one of the most valuable genres of audiovisual data. However, spoken lectures are difficult to reuse because browsing and efficient searching within spoken lectures is difficult. To promote the research activities in the spoken lecture retrieval, this paper reports a test collection for its evaluation. The test collection consists of the target spoken documents of about 2,700 lectures (604 hours) taken from the Corpus of Spontaneous Japanese (CSJ), 39 retrieval queries, the relevant passages in the target documents for each query, and the automatic transcription of the target speech data. We report the retrieval performance targeting the constructed test collection by applying a standard spoken document retrieval (SDR) method, which serves as a baseline for the forthcoming SDR studies using the test collection. We also introduce the several studies conducted by the users of the test collection.**

I. Introduction

Traditionally, human beings have used spoken language mainly for communication. However, advances in speech recognition technologies will make it possible to use spoken language in addition to written language as a medium for storing and transmitting knowledge. In practice, audio data such as broadcast news, lectures, and Weblog-style recording in podcasts is increasingly available via the Internet. Among others, the lecture is one of the most valuable genres of audiovisual data.

Spoken document processing is a promising technology for utilizing the lectures in various way. Spoken document processing deals with speech data, using techniques similar to text processing. These include transcription, translation, search, alignment to parallel materials such as slides, textbooks, and related papers, structuring, summarizing, and editing. As this technology improves, there will be advanced applications such as computer-aided remote lecture systems and self-learning systems with efficient searching and browsing. Indeed, several multimedia retrieval systems and prototype self-learning systems targeting spoken lectures have been reported so far [1], [2], [3], [4]. However, spoken document processing methods are difficult to evaluate because they require a subjective judgment and/or the checking of large quantities of evaluation data. In certain situations, a test collection can be used for a shareable standard of evaluation.

To date, test collections for information retrieval research have been constructed from sources such as newspaper articles [5], Web documents [6], and patent documents [7]. Test collections for cross-language retrieval [8], [9], open-domain question answering [10], [11], and text summarization [12] have also been constructed.

A test collection for spoken document retrieval (SDR) is usually based on a broadcast news corpus. Compared to broadcast news, lectures are more challenging for speech recognition because the vocabulary can be technical and specialized, the speaking style can be more spontaneous, and there is a wider variety of speaking styles and structure types for lectures. Moreover, a definition of the semantic units in lectures is ambiguous because it is highly dependent on the queries. We aim to construct a test collection for ad hoc retrieval and term detection.

The rest of this paper is organized as follows. Section II describes how we constructed the test collection for spoken document retrieval, targeting lecture audio data. In Section III, we evaluate the test collection by investigating its baseline retrieval performance, which was obtained by applying a conventional document retrieval method. We also show several SDR research activities conducted by the users of the test collection in Section IV.

II. Constructing a Test Collection for SDR

A test collection for text document retrieval comprises three elements: (1) a huge document collection in a target domain, (2) a set of queries, and (3) results of relevance judgments, i.e., sets of relevant documents that are selected from the collection for each query in the query set.

In the spoken document case, the text collection should not merely be replaced with a spoken document collection. Two additional elements are necessary for an SDR test collection:

TABLE I
SUMMARY OF THE TARGET DOCUMENT COLLECTION FROM CSJ.

	Speakers	Lectures	Data size (hours)
Academic lectures	819	987	274.4
Simulated lectures	594	1715	329.9

(4) manual transcriptions and (5) automatic transcriptions of the spoken document collection. The manual transcriptions are necessary for relevance judgment by the test collection constructors and can be used as a "gold standard" for automatic transcriptions by test collection users. The automatic transcriptions obtained by using a large vocabulary continuous speech recognition (LVCSR) system are also desirable for supporting those researchers who do not have their own facilities for speech recognition and yet are interested in aspects of text processing in SDR.

These elements of our SDR test collection are described in the following subsections.

A. Target Document Collection

We chose the Corpus of Spontaneous Japanese (CSJ) [13] as the target collection. It includes several kinds of spontaneous speech data, such as lecture speech and spoken monologues, together with their manual transcriptions. From them, we selected two kinds of lecture speech from the CSJ: lectures at academic societies, and simulated lectures on a given subject. The collection contains 2702 lectures and more than 600 hours of speech, which are segmented into utterances by the boundaries defined in the CSJ. Table I summarizes the collection [14]. Because its size is comparable to the Text Retrieval Conference (TREC) SDR test collection [15], the size is sufficient for the purposes of retrieval research.

B. Queries

Queries, or information needs, for spoken lectures can be categorized into two types: those searching for a whole lecture and those looking for some information described in a part of a lecture. We focus on the latter type of query in our test collection, because this is much more likely than the former in terms of the practical use of lecture search applications. For such a query, the length of the relevant segment will vary, so a document, in information retrieval (IR) terms, must be a segment with a variable length. In this paper, we refer to such a segment as a "passage."

Another reason why we focused on partial lectures arises from technical issues involved in constructing a test collection for retrieval research. If we regard each lecture in the collection as a document, the corresponding ad hoc task is defined as searching for relevant documents from among the 2702 documents. This number is far less than that used for the TREC SDR task, which has 21,754 documents (stories) in the target collection.

Therefore, we constructed queries that ask for passages of varying lengths from lectures. In order to uniform the granularities of the answers, we tried to control the length to about one minute on average, which is approximately equivalent to the length of an explanation for a presentation slide , by specifying this in the guidelines. It is observed that the constructed query tends to be less like a query in document retrieval, but more like a question submitted to a question answering system. In addition to the guidelines, nine subjects are relied upon to invent such queries by investigating the target documents and we obtained about 100 initial queries in total, from which we planned to select the appropriate subset by conducting a relevance judgment in the next step.

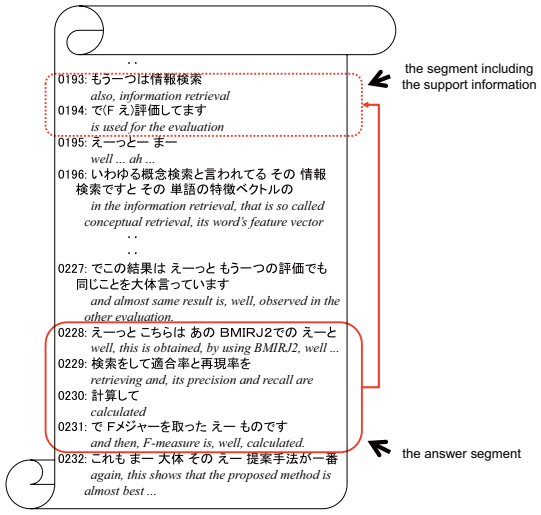

Fig. 1. An example of the answer and the supporting segment.

C. Relevance Judgment

The relevance judgment for the queries was conducted manually and performed against every variable length segment (or passage) in the target collection. One of the difficulties related to the relevance judgment comes from the treatment of the supporting information. We regarded a passage as irrelevant to a given query even if it was a correct answer in itself to the query, when it had no supporting information that would convince the user who submitted the query of the correctness of the answer. For example, for the query "How can we evaluate the performance of information retrieval?," the answer "F-measure" is not sufficient, because it does not say by itself that it is really an evaluation measure for information retrieval. The relevant passage must also include supporting information indicating that "F-measure" is one of the evaluation metrics used for information retrieval. Figure 1 shows an example of an answer and its supporting information for the query "How can we evaluate the performance of information retrieval?"

As shown in Figure 1, the supporting information does not always appear together with the relevant passage, but may appear somewhere else in the same lecture. Therefore, we regarded a passage as relevant to a given query if it had some supporting information in some segment of the same lecture. If a passage in a lecture was judged relevant, the range of

TABLE II
STATISTICS FOR THE RESULTS OF THE RELEVANCE JUDGMENT.

Label	Passages per query	Unique lectures per query	Utterances per passage
Relevant	11.18	7.90	10.39
Relevant & Partially Relevant	12.69	9.26	10.88

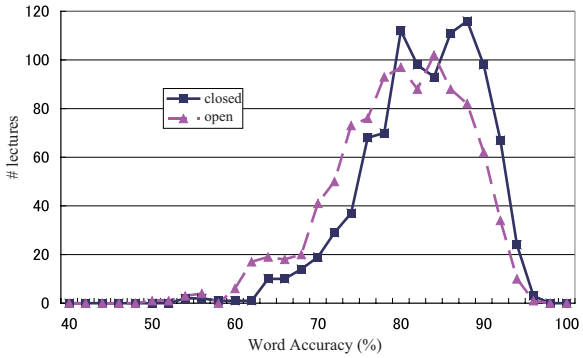

Fig. 2. Distribution of word error rates in CSJ lectures.

the passage and the ranges of the supporting segments, if any, along with the lecture ID, were recorded in our "golden" file.

The relevance judgment against the 100 initial queries was performed by the nine query constructors themselves. For each query, one assessor, i.e. its constructor, searched its relevant passages and judged their degrees of relevancy. The assessor manually selected the candidate passages from the target document collection and labeled them into three classes according to the degree of their relevancy: "Relevant," "Partially relevant," and "Irrelevant." For this task, the assessor used the document search engine for the initial retrieval, and then investigated the search results to find the passage.

Finally, after we filtered out the queries that had no more than four relevant passages in the target collection, 39 queries were selected for our test collection. Table II shows some statistics of the result.

D. Automatic Transcription

A Japanese LVCSR decoder [16] was used to obtain automatic transcriptions of the target spoken documents. Because the target spoken documents of the lecture speech are more spontaneous than those of broadcast news, the speech recognition accuracy was expected to be worse than for TREC SDR. To achieve better recognition results, both the acoustic model and the language model were trained by using the CSJ itself [17]. Specifically, the language model is trained by using all target lectures except the *core lectures*, which are defined in CSJ and consist of 70 academic lectures and 107 simulated lectures, while the acoustic model is trained by using all target lectures [1].

[1]More specifically, all lectures excluding ten *test-set* lectures. See [17] for more details.

TABLE III
A COMPARISON BETWEEN TREC-9 SDR AND OUR CSJ SDR TEST COLLECTIONS.

	TREC9 SDR	CSJ SDR
Language	English	Japanese
Target documents	Broadcast news	Lecture speech
Quantity	557 hours	604.3 hours
Documents	21,754	2702 (30,762 seg. *)
Words per document	169	2324.9 (204.2 per seg. *)
Queries	50	39
Reference Transcription	closed caption (WER 10.3%)	manual transcription
ASR WER	26.7%	21.4%

* A succession of 30 utterances is considered to be a segment.

For the sake of comparison, another acoustic model trained by using only the simulated lectures was prepared to obtain recognition results using an open setting. The recognition results targeting the academic lectures obtained by these two acoustic models were compared. Figure 2 shows the two distributions of the word accuracy of the CSJ lectures, obtained by using the closed and open settings. They differ in their average, but have almost the same shape, which ranges between about 0.65 and 0.95. For the first attempt, we decided to use the recognition results in a closed setting. The word error rate (WER) was about 20%, which is comparable to that of the TREC SDR task.

E. Summary of the Test Collection

Table III summarizes the constructed test collection compared with the TREC-9 SDR test collection. Although there are some differences between them especially in the language (English vs. Japanese) and the target domain (broadcast news vs. lecture speech), the task size is almost comparable if 30 utterances are used for a document in our task.

III. EVALUATION

To evaluate the test collection and to assess the baseline retrieval performance obtained by applying a standard method for SDR, an ad hoc retrieval experiment targeting the test collection was conducted.

A. Alignment between Automatic and Manual Transcriptions

The relevance judgment described in Section II-C is performed against the CSJ transcriptions. On the contrary, the automatic transcription described in Section II-D does not include the sentence boundaries defined in the CSJ transcriptions. Therefore, the results of the relevance judgment cannot be mapped into the automatic transcriptions straightforwardly.

Relying on the fact that the recognition accuracy of the automatic transcription is relatively high, we aligned the utterances defined in the CSJ transcriptions with the segments in the automatic transcriptions by using the text-based DP-matching guided by the edit distance described as follows.

1) From the automatic transcriptions, the text and the boundary information between the recognition units are

TABLE IV
STATISTICS OF THE REDEFINED TASK.

Utterances per passage	15	30	60	Lecture
Target documents	60,202	30,762	16,060	2,702
Average relevant documents (**R**)	16.36	12.77	10.90	8.13
Average relevant documents (**R+P**)	19.03	14.79	12.54	9.44

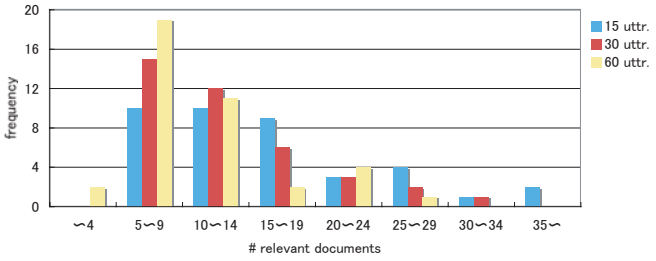

Fig. 3. The distribution of the relevant documents (**R** degree).

extracted. From the CSJ transcriptions, the text and utterance boundary information are extracted. Both types of boundary informations are annotated with a unique identical marker, with the expectation that the two symbols from the transcriptions will be aligned together in the following matching process.

2) The texts of both sides are morphologically segmented by using a Japanese morphological analyzer, with the boundary markers retained at their original positions. For each side, the sequences of the morphemes and boundary markers are obtained.
3) The two sequences are aligned by using DP-matching, which minimizes the edit distance between them.
4) For each utterance in the CSJ transcriptions, the corresponding morpheme sequence in the automatic transcription can be obtained by investigating the resulting alignment.

Here we rely on the high recognition accuracy. However, if the accuracy is low, the text-based method is not appropriate, and the method using the time information should be adopted.

B. Task Definition

The purpose of the evaluation is to observe the performance obtained by applying the standard method for SDR, i.e., term indexing and a vector space model for retrieval, and to compare the results with other studies in SDR and IR research. However, the primary task of our test collection, i.e., to find passages with variable utterance length, is not conventional. Therefore, we redefined the conventional retrieval task, in which a fixed set of documents is predefined and indexed statically to prepare for the retrieval.

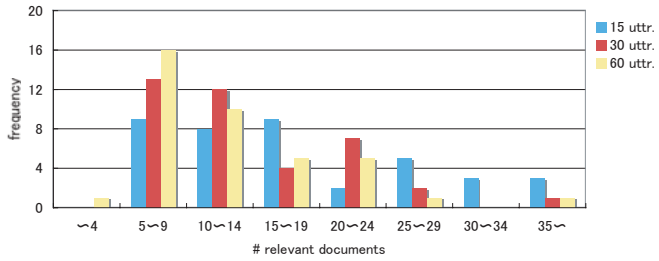

Fig. 4. The distribution of the relevant documents (**R+P** degree).

First, we defined pseudopassages by automatically segmenting each lecture into sequences of segments with fixed numbers of sequential utterances: 15, 30, and 60. When 30 utterances are used in a segment, the number of pseudopassages is 30,762, and the number of words in a document is 204.2 on average, which are comparable numbers to those for TREC SDR.

Next, we assigned retrieved pseudopassages a relevance label as follows: if the pseudopassage shared at least one utterance that came from the relevant passage specified in the "golden file," then the pseudopassage was labeled as "relevant." Two degrees of relevance were used for the evaluation as follows.

R The passages labeled "Relevant" are used for deciding the relevant pseudopassages.

R+P The passages labeled either "Relevant" or "Partially relevant" are used for deciding the relevant pseudopassages.

Table IV lists the size of the target documents (the number of pseudopassages) and the number of relevant documents for each task. Figure 3 and 4 show the distribution of the relevant documents found in our redefined ad hoc retrieval task in **R** and **R+P** degrees, respectively.

C. Ad hoc Retrieval Methods

All pseudopassages were then indexed by using either their words, their character bi-grams, or a combination of the two. At the retrieval time, the query is also pre-processed into the same representation as the indexing unit. The vector space model was used as the retrieval model, and TF–IDF (Term Frequency–Inverse Document Frequency) with pivoted normalization [18] was used for term weighting. We compared three representations of the pseudopassages: the 1-best automatically transcribed text, the union of the 10-best automatically transcribed texts, and the manually transcribed reference text.

D. Evaluation Metric

We used 11-point average precision [19] as our evaluation metric, which is obtained by averaging the following AP over

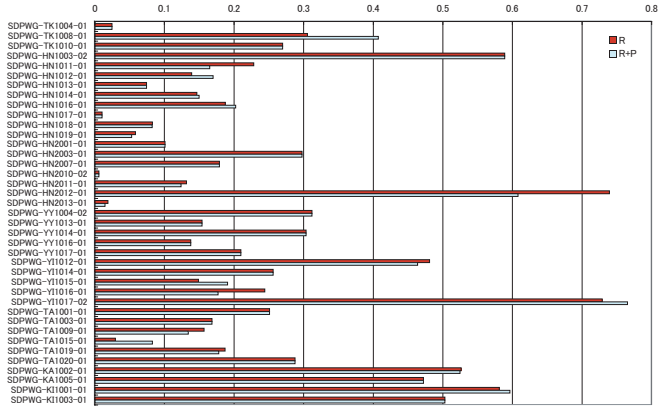

Fig. 5. 11-point average precision for each query (using 30 utterances as a document, and manual transcription for the indexing.)

TABLE V
11-POINTS AVERAGE PRECISIONS USING 15 UTTERANCES AS A PSEUDOPASSAGE.

		Indexing unit		
Relevance degree	Transcription	Word	Char. 2-gram	Word + char. 2-gram
R	Reference	0.180	0.165	0.185
	10-best	0.177	0.145	0.167
	1-best	0.155	0.135	0.146
R+P	Reference	0.181	0.166	0.188
	10-best	0.179	0.150	0.171
	1-best	0.159	0.143	0.152

TABLE VI
11-POINTS AVERAGE PRECISIONS USING 30 UTTERANCES AS A PSEUDOPASSAGE.

		Indexing unit		
Relevance degree	Transcription	Word	Char. 2-gram	Word + Char. 2-gram
R	Reference	0.249	0.216	0.240
	10-best	0.225	0.205	0.232
	1-best	0.213	0.188	0.207
R+P	Reference	0.249	0.220	0.242
	10-best	0.227	0.210	0.234
	1-best	0.211	0.194	0.211

the queries.

$$IP(x) = \max_{x \le R_i} P_i$$

$$AP = \frac{1}{11} \sum_{i=0}^{10} IP(\frac{i}{10}),$$

where R_i and P_i are the recall and precision up to the i-th retrieved documents. In practice, we retrieved 1000 documents for each query to calculate the AP.

E. Results

Figure 5 shows the 11-point average precision for each query, where 30 utterances were used as a pseudo-passage, and the reference transcriptions were used for indexing. It indicates that the variance in difficulty is high. For example, the hardest query (SDPWG-HN2010-02: *"How does smoking influence our health and what hazards does smoking have?"*) can find only one (**R** degree) relevant passage in the 100-best candidates. On the other hand, the easiest query (SDPWG-HN2012-01: *"Where are some wine production areas? I especially want to know about very famous or personally preferred areas."*) can find eight (**R** degree) relevant passages in the 10-best candidates.

TABLE VII
11-POINTS AVERAGE PRECISIONS USING 60 UTTERANCES AS A PSEUDOPASSAGE.

		Indexing unit		
Relevance degree	Transcription	Word	Char. 2-gram	Word + Char. 2-gram
R	Reference	0.294	0.269	0.297
	10-best	0.256	0.236	0.265
	1-best	0.251	0.227	0.253
R+P	Reference	0.305	0.278	0.308
	10-best	0.261	0.243	0.271
	1-best	0.256	0.235	0.263

TABLE VIII
11-POINTS AVERAGE PRECISIONS USING TH WHOLE LECTURE AS A PSEUDOPASSAGE.

		Indexing unit		
Relevance degree	Transcription	Word	Char. 2-gram	Word + Char. 2-gram
R	Reference	0.453	0.443	0.468
	10-best	0.399	0.384	0.414
	1-best	0.411	0.397	0.426
R+P	Reference	0.473	0.454	0.489
	10-best	0.413	0.400	0.428
	1-best	0.423	0.409	0.441

Tables V, VI, VII, and VIII lists all the evaluation results obtained by combining the four passage lengths (15, 30, 60 utterances, or a whole lecture), two degrees of relevance (**R** or **R+P**), three kinds of transcription (reference, 1-best or 10-best recognition candidates), and three kinds of indexing unit (word, character 2-gram, or a combination of the two).

Using words as the indexing unit is more effective than using character 2-grams. Using both words and character 2-grams slightly improves the retrieval performance, especially for longer target document lengths, i.e., using 60 utterances or a whole lecture as a document. **R+P** consistently gives better results than **R**, but the difference is not large.

Figure 6 summarizes the results using a word as the indexing unit and R degree for the relevancy, to compare the three kinds of representations of the target documents. It shows that using the 1-best automatically transcribed text decreases the IR performance by 10% to 15% compared with using the reference transcription. We also found that the use of 10-best candidates was effective for tasks with shorter passages, namely 15 and 30 utterances, but was less effective for those with longer passages, namely 60 utterances and whole lectures.

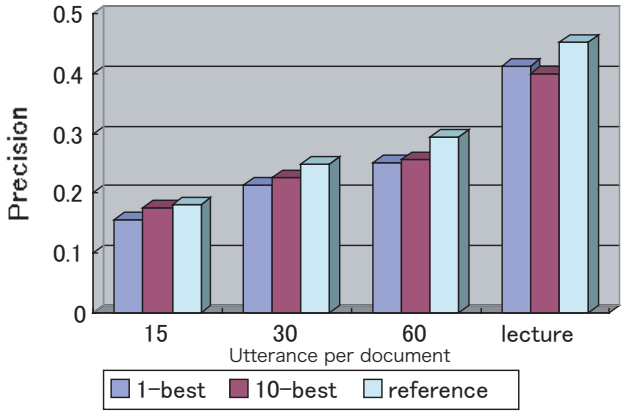

Fig. 6. 11-point average precision using 1-best, 10-best, and reference transcriptions for indexing documents.

Overall, the evaluation results show that the ad hoc retrieval task for lecture audio data is much more difficult than that for broadcast news, where the precision was reported to be around 0.45 for a task condition comparable to our 30-utterance condition [15]. The retrieval performance is very low except the case where the whole lecture is used as a passage. This is partly because a relevant passage often has its supporting segments separated from it in the same document, meaning that the relevant passage does not always have self-contained information.

We observed two other reasons why lectures are difficult to be retrieved. Firstly, the speaker of the lecture at an academic society tends to omit the basic explanation about his presentation as his audience has common background knowledge about his research topic. Secondly, presentation slides are used in the lecture at academic society, and the keywords written in them are not often uttered in the speech. For these reasons, the useful keywords for retrieval may not appear in the speech data, making the retrieval difficult.

IV. Research Activities using the Test Collection

Using the test collection presented in this paper, several SDR studies have been already reported.

Akiba and Yokota [20] proposed an SDR method using the word translation model. It filled the gap between the automatically transcribed text and the correctly transcribed text by using a statistical translation technique. They evaluated the method by the redefined SDR task described in Section III-B.

Hu and Kashioka [21] applied a dimension reduction technique to SDR. They used the non-negative matrix factorization (NMF) to retrieve spoken documents in semantic space rather than word-vector space, in order to deal with unmatching but synonymous keywords between the input query and a document. They evaluated the method by the task retrieving lectures in the test collection.

Sugimoto et al. [22] proposed a document expansion method for SDR. They expanded spoken documents by using the Web pages related to them, in order to overcome Out-Of-Vocabulary problem caused by the automatic transcription. The related Web pages were collected by querying the Web search engine using the transcribed documents as a query. The task of retrieving lectures in the test collection was also used to evaluate the method.

V. Conclusion and Future Work

A test collection for spoken lecture ad hoc retrieval was constructed. We chose the Corpus of Spontaneous Japanese (CSJ) as the target collection and constructed 39 queries designed to search the information described in a partial lecture rather than a whole lecture. Relevance judgments for these queries were conducted manually and performed against every variable length segment in the target collection. Automatic transcriptions of the target collection were also constructed by applying a large vocabulary continuous speech recognition (LVCSR) decoder, to support researchers in various fields.

To evaluate the test collection and assess the baseline retrieval performance obtained by applying a standard method for SDR, an ad hoc retrieval experiment targeting the test collection was conducted. It revealed that the ad hoc retrieval task for lecture audio data was much more difficult than that for broadcast news.

We are now constructing another test collection for the term detection task. We will also prepare another automatic transcription with moderate WER by using an acoustic model and a language model trained in open conditions.

References

[1] A. Fujii, K. Itou, and T. Ishikawa, “Lodem: A system for on-demand video lectures,” *Speech Communication*, vol. 48, No.5, pp.516-531, no. 5, pp. 516–531, 2006.

[2] H. Okamoto, W. Nakano, T. Kobayashi, S. Naoi, H. Yokota, K. Iwano, and S. Furui, “Presentation-content retrieval integrated with the speech information,” *IEICE Transactions on Information and Systems*, vol. J90-D, no. 2, pp. 209–222, 2007.

[3] S. Nakagawa, S. Togashi, M. Yamaguchi, Y. Fujii, and N. Kitaoka, “Useful contents of classroom lecture speech and a browsing system,” *IEICE Transactions on Information and Systems*, vol. J91-D, no. 2, pp. 238–249, 2008.

[4] S. yi Kong, M. ru Wu, C. kuang Lin, Y. sheng Fu, and L. shan Lee, “Learning on demand - course lecture distillation by information extraction and semantic structuring for spoken documents,” 2009, pp. 4709–4712.

[5] T. Kitani, Y. Ogawa, T. Ishikawa, H. Kimoto, I. Keshi, J. Toyoura, T. Fukushima, K. Matsui, Y. Ueda, T. Sakai, T. Tokunaga, H. Tsuruoka, H. Nakawatase, and T. Agata, “Lessons from BMIR-J2: A test collection for Japanese IR systems,” in *Proceedings of ACM SIGIR*, 1998, pp. 345–346.

[6] K. Oyama, M. Takaku, H. Ishikawa, A. Aizawa, and H. Yamana, “Overview of the NTCIR-5 WEB navigational retrieval subtask 2,” in *Proceedings of the Fifth NTCIR Workshop Meeting*, 2005, pp. 423–442.

[7] A. Fujii, M. Iwayama, and N. Kando, “Overview of patent retrieval task at NTCIR-5,” in *Proceedings of the Fifth NTCIR Workshop Meeting*, 2005, pp. 269–277.

[8] F. C. Gey and D. W. Oard, “The TREC-2001 cross-language information retrieval track: Searching arabic using english, french or arabic queries,” in *Proceedings of TREC-10*, 2001, pp. 16–25.

[9] K. Kishida, K. hua Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, and S. H. Myaeng, “Overview of CLIR task at the fifth NTCIR workshop,” in *Proceedings of the Fifth NTCIR Workshop Meeting*, 2005, pp. 1–38.

[10] E. M. Voorhees and D. M. Tice, "The TREC-8 question answering track evaluation," in *Proceedings of the 8th Text Retrieval Conference*, Gaithersburg, Maryland, 1999, pp. 83–106.
[11] T. Kato, J. Fukumoto, and F. Masui, "An overview of NTCIR-5 QAC3," in *Proceedings of the Fifth NTCIR Workshop Meeting*, 2005, pp. 361–372.
[12] T. Hirao, M. Okumura, T. Fukusima, and H. Nanba, "Text summarization challenge 3 – text summarization evaluation at NTCIR workshop 4," in *Proceedings of the Fourth NTCIR Workshop*, 2004.
[13] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proceedings of LREC*, 2000, pp. 947–952.
[14] K. Maekawa, *Overview of the Corpus of Spontaneous Japanese, Version 1.0*, the CSJ attached document.
[15] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of TREC-9*, 1999, pp. 107–129.
[16] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," in *Proceedings of European Conference on Speech Communication and Technology*, Sept. 2001, pp. 1691–1694.
[17] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the corpus of spontaneous Japanese," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 135–138.
[18] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proceedings of ACM SIGIR*, 1996, pp. 21–29.
[19] S. Teufel, "An overview of evaluation methods in TREC ad hoc information retrieval and TREC question answering," in *Evaluation of Text and Speech Systems*, ser. Text, Speech and Language Technology, L. Dybkjär, H. Hemsen, and W. Minker, Eds. Springer, 2007, no. 37, pp. 163–186.
[20] T. Akiba and Y. Yokota, "Spoken document retrieval by translating recognition candidates into correct transcriptions," in *Proceedings of International Conference on Speech Communication and Technology*, 2008, pp. 2166–2169.
[21] X. Hu and H. Kashioka, "Study on spoken document retrieval by using non-negative matrix factorization," in *Proceedings of 2009 Spring Meeting of Acoustical Society of Japan*, 2009, pp. 281–282.
[22] K. Sugimoto, S. Maezawa, H. Nishizaki, and Y. Sekiguchi, "Spoken document retrieval using web pages highly related to target documents (in Japanese)," in *Proceedings of the Third Spoken Document Processing Workshop*, 2009, pp. 33–38.