Journal of Information Processing Vol. 17 82–94 (Feb. 2009)

Regular Paper

Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data

Tomoyosi Akiba,[†1] Kiyoaki Aikawa,[†2]
Yoshiaki Itoh,[†3] Tatsuya Kawahara,[†4]
Hiroaki Nanjo,[†5] Hiromitsu Nishizaki,[†6]
Norihito Yasuda,[†7] Yoichi Yamashita[†8]
and Katunobu Itou[†9]

The lecture is one of the most valuable genres of audiovisual data. Though spoken document processing is a promising technology for utilizing the lecture in various ways, it is difficult to evaluate because the evaluation require a subjective judgment and/or the verification of large quantities of evaluation data. In this paper, a test collection for the evaluation of spoken lecture retrieval is reported. 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. This paper also reports 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.

1. Introduction

The lecture is one of the most valuable genres of audiovisual data. Previously, however, lectures have mostly been archived in the form of books or related papers. The main reason for this is that spoken lectures are difficult to reuse because browsing and efficient searching within spoken lectures is difficult.

Spoken document processing is a promising technology for solving these problems. 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)–3)]. 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 [4)], Web documents [5)], and patent documents [6)]. Test collections for cross-language retrieval [7),8)], open-domain question answering [9),10)], and text summarization [11)] 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 2 describes how we constructed the test collection for spoken document retrieval, targeting lecture audio data. In Section 3, we evaluate the test collection by investigating its baseline retrieval performance, which was obtained by applying a conventional document retrieval method.

2. Constructing a Test Collection for SDR

A test collection for text document retrieval comprises three elements: (1) a

†1 Toyohashi University of Technology
†2 Tokyo University of Technology
†3 Iwate Prefectural University
†4 Kyoto University
†5 Ryukoku University
†6 University of Yamanashi
†7 Nippon Telegraph and Telephone Corporation
†8 Ritsumeikan University
†9 Hosei University

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: (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.

2.1 Target Document Collection

We chose the Corpus of Spontaneous Japanese (CSJ) [12] 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: lectures at academic societies, and simulated lectures on a given subject. The collection contains 2,702 lectures and more than 600 hours of speech. **Table 1** summarizes the collection [13]. Because its size is comparable to the Text Retrieval Conference (TREC) SDR test collection [14], the size is sufficient for the purposes of retrieval research.

2.2 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."

Table 1 Summary of the target document collection from CSJ.

	Speakers	Lectures	Data size (hours)
Academic lectures	819	987	274.4
Simulated lectures	594	1,715	329.9

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 2,702 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.

2.3 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 informa-

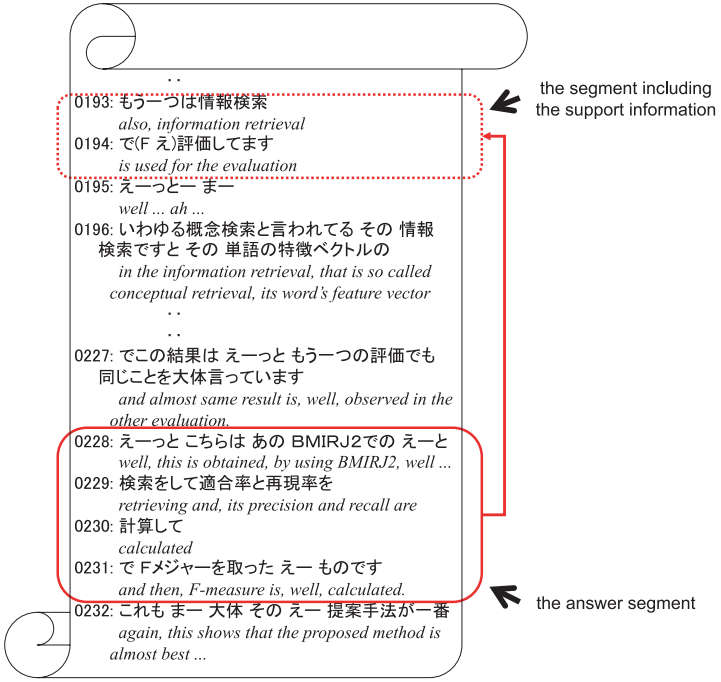

Fig. 1 An example of the answer and the supporting segment.

tion 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 Fig. 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 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.

Table 2 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

Finally, after we filtered out the queries that had no more than four relevant passages in the target collection, 39 queries, listed in Appendix A.1, were selected for our test collection. **Table 2** shows some statistics of the result. Appendix A.2 samples some queries and judgments of relevancy.

2.4 Automatic Transcription

A Japanese LVCSR decoder [15] 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 [16]. 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].

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

⋆1 More specifically, all lectures excluding ten *test-set* lectures. See Ref. 16) for more details.

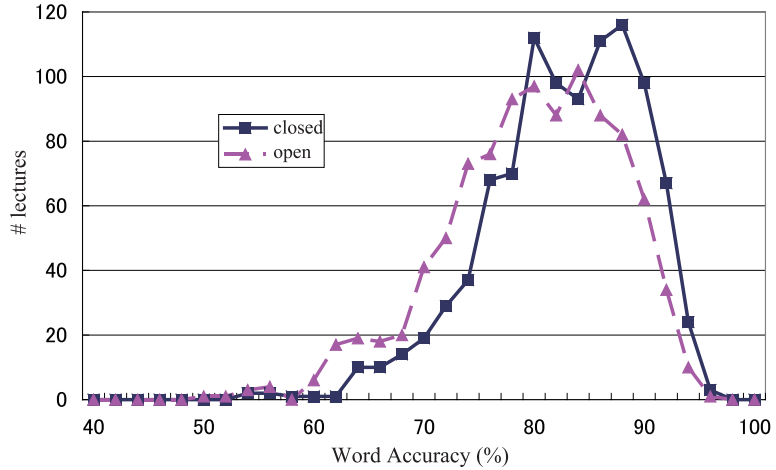

Fig. 2 Distribution of word error rates in CSJ lectures.

Table 3 A comparison between TREC-9 SDR and our CSJ SDR test collections.

	TREC9 SDR	CSJ
Language	English	Japanese
Target documents	Broadcast news	Lecture speech
Quantity	557 hours	604.3 hours
Documents	21,754	2,702 (30,762 seg. *)
Words per document	169	2,324.9 (204.2 per seg. *)
Queries	50	39
Transcription	Low grade (WER 10.3%)	High grade
WER	26.7%	21.4%

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

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.

2.5 Summary of the Test Collection

Table 3 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.

3. 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.

3.1 Alignment between Automatic and Manual Transcriptions

The relevance judgment described in Section 2.3 is performed against the CSJ transcriptions. On the contrary, the automatic transcription described in Section 2.4 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 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.

3.2 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.

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 4 lists the size of the target documents (the number of pseudopassages) and the number of relevant documents for each task. **Figure 3** shows the distribution of the relevant documents found in our redefined ad hoc retrieval task.

3.3 Ad hoc Retrieval Methods

All pseudopassages were then indexed by using either their words, their character 2-grams, or a combination of the two. The vector space model was used as the retrieval model, and TF–IDF (Term Frequency–Inverse Document Frequency) with pivoted normalization [17] 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.

Table 4 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

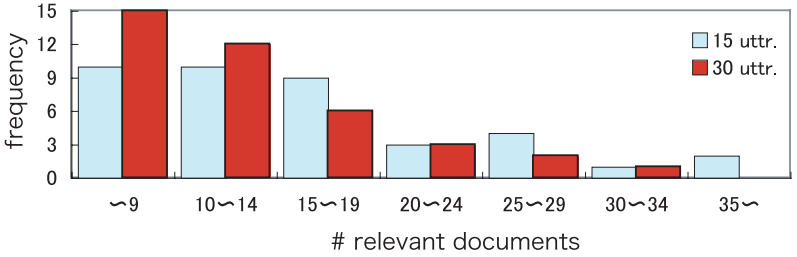

Fig. 3 The distribution of the relevant documents.

3.4 Evaluation Metric

We used 11-point average precision [18] as our evaluation metric, which is obtained by averaging the following AP over the queries.

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

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

where R_i and P_i are the recall and precision up to the i-th retrieved documents. In practice, we retrieved 1,000 documents for each query to calculate the AP.

3.5 Results

Figure 4 shows the 11-point average precision for each query, where 30 utterances were used as a pseudopassage, and the reference transcriptions were used for indexing. It indicates that the variance in difficulty is high. For example, the hardest query can find only one (**R** degree) relevant passage in the 100-best candidates. On the other hand, the easiest query can find eight (**R** degree) relevant passages in the 10-best candidates.

Tables 5, **6**, **7** and **8** 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-

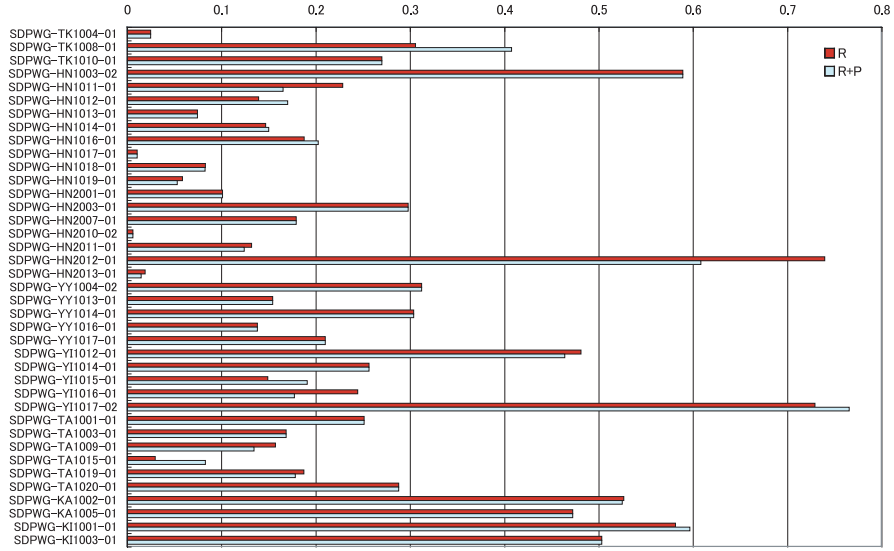

Fig. 4 11-point average precision for each query (using 30 utterances as a document, and manual transcription for the indexing).

Table 5 11-point 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

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.

Table 6 11-point 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

Table 7 11-point 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 8 11-point average precisions using the 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

Figure 5 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.

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 pre-

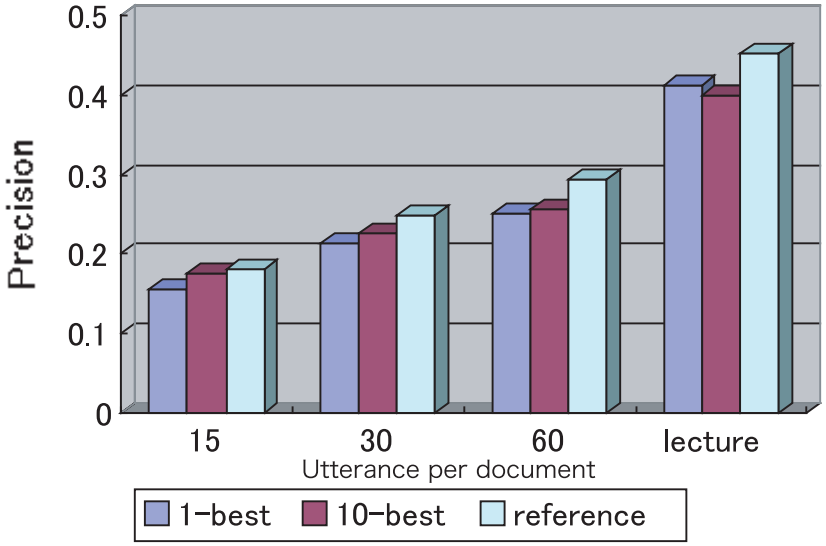

Fig. 5 11-point average precision using 1-best, 10-best, and reference transcriptions for indexing documents.

cision was reported to be around 0.45 for a task condition comparable to our 30-utterance condition [14]. 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.

4. 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) Fujii, A., Itou, K. and Ishikawa, T.: LODEM: A system for on-demand video lectures, *Speech Communication*, Vol.48, No.5, pp.516–531 (2006).
2) Okamoto, H., Nakano, W., Kobayashi, T., Naoi, S., Yokota, H., Iwano, K. and Furui, S.: Presentation-Content Retrieval Integrated with the Speech Information, *IEICE Trans. Inf. Syst.*, Vol.J90-D, No.2, pp.209–222 (2007).
3) Nakagawa, S., Togashi, S., Yamaguchi, M., Fujii, Y. and Kitaoka, N.: Useful Contents of Classroom Lecture Speech and a Browsing System, *IEICE Trans. Inf. Syst.*, Vol.J91-D, No.2, pp.238–249 (2008).
4) Kitani, T., Ogawa, Y., Ishikawa, T., Kimoto, H., Keshi, I., Toyoura, J., Fukushima, T., Matsui, K., Ueda, Y., Sakai, T., Tokunaga, T., Tsuruoka, H., Nakawatase, H. and Agata, T.: Lessons From BMIR-J2: A Test Collection for Japanese IR Systems, *Proc. ACM SIGIR*, pp.345–346 (1998).
5) Oyama, K., Takaku, M., Ishikawa, H., Aizawa, A. and Yamana, H.: Overview of the NTCIR-5 WEB Navigational Retrieval Subtask 2, *Proc. 5th NTCIR Workshop Meeting*, pp.423–442 (2005).
6) Fujii, A., Iwayama, M. and Kando, N.: Overview of Patent Retrieval task at NTCIR-5, *Proc. 5th NTCIR Workshop Meeting*, pp.269–277 (2005).

7) Gey, F.C. and Oard, D.W.: The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French or Arabic Queries, *Proc. TREC-10*, pp.16–25 (2001).
8) Kishida, K., Chen, K.-H., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H. and Myaeng, S.H.: Overview of CLIR Task at the Fifth NTCIR Workshop, *Proc. 5th NTCIR Workshop Meeting*, pp.1–38 (2005).
9) Voorhees, E.M. and Tice, D.M.: The TREC-8 Question Answering Track Evaluation, *Proc. 8th Text Retrieval Conference*, Gaithersburg, Maryland, pp.83–106 (1999).
10) Kato, T., Fukumoto, J. and Masui, F.: An Overview of NTCIR-5 QAC3, *Proc. 5th NTCIR Workshop Meeting*, pp.361–372 (2005).
11) Hirao, T., Okumura, M., Fukusima, T. and Nanba, H.: Text Summarization Challenge 3 — Text Summarization Evaluation at NTCIR Workshop 4, *Proc. 4th NTCIR Workshop* (2004).
12) Maekawa, K., Koiso, H., Furui, S. and Isahara, H.: Spontaneous Speech Corpus of Japanese, *Proc. LREC*, pp.947–952 (2000).
13) Maekawa, K.: *Overview of the Corpus of Spontaneous Japanese, Version 1.0*, the CSJ attached document.
14) Garofolo, J.S., Auzanne, C.G.P. and Voorhees, E.M.: The TREC Spoken Document Retrieval Track: A Success Story, *Proc. TREC-9*, pp.107–129 (1999).
15) Lee, A., Kawahara, T. and Shikano, K.: Julius — An Open Source Real-Time Large Vocabulary Recognition Engine, *Proc. European Conference on Speech Communication and Technology*, pp.1691–1694 (2001).
16) Kawahara, T., Nanjo, H., Shinozaki, T. and Furui, S.: Benchmark test for speech recognition using the Corpus of Spontaneous Japanese, *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp.135–138 (2003).
17) Singhal, A., Buckley, C. and Mitra, M.: Pivoted document length normalization, *Proc. ACM SIGIR*, pp.21–29 (1996).
18) Teufel, S.: An overview of evaluation methods in TREC ad hoc information retrieval and TREC question answering, *Evaluation of Text and Speech Systems*, Dybkjär, L., Hemsen, H. and Minker, W. (Eds.), Text, Speech and Language Technology, No.37, pp.163–186, Springer (2007).

Appendix

A.1 The 39 Constructed Queries

TK1004-01 話者認識の学習データのサイズが知りたい
Describe the size of training data for speaker recognition systems.

TK1008-01 ディズニーランドに行った話し
Stories of personal visits to Disneyland.

TK1010-01 ペットの犬の名前のリストアップ
List the names of pet dogs.

HN1003-02 オークションにおける自動入札戦略を知りたい
What strategy exists for an automatic tender in an auction?

HN1011-01 中国語の特徴を知りたい
Tell me the characteristics of the Chinese language?

HN1012-01 フランス語の特徴を知りたい
Tell me the characteristics of French?

HN1013-01 翻訳手法にはどのようなものがあるか
What methods of automatic translation are there?

HN1014-01 講演音声の特徴について知りたい
Tell me the characteristics of lecture speech?

HN1016-01 ＯＳの役割または種類についての解説を見たい
Tell me some types of operating systems or their functions.

HN1017-01 植物の効用について述べている箇所を探したい
Search for a description of the beneficial effects of plants.

HN1018-01 ペットを飼うことの効用または目的について述べている箇所を探したい
Search for a description of the beneficial effects of pets or their purpose.

HN1019-01 電車乗車時のマナーについて意見を述べている箇所をみつけたい
Search for an opinion about manners when getting on a train.

HN2001-01 音声認識技術，音声処理技術を利用，応用しているアプリケーションにはどのようなものがあるか，どのような処理技術を用いているかを知りたい．
What is an application using speech recognition or speech processing techniques? In addition, what technique is used in it?

HN2003-01 非言語情報とパラ言語情報の違いを知りたい．具体的な例があると嬉しい．
What is the difference between nonverbal and paralanguage information? I prefer concrete examples.

HN2007-01 機械学習システムにはどのようなものがあるのかを知りたい．できればどんな研究テーマに使われているのかも知りたい．
I would like to know about various types of machine learning systems. I preferably want to know which research areas machine learning is applied in.

HN2010-02 煙草が体に及ぼす影響，有害性にはどのようなものがあるか？
How does smoking influence our health and what hazards does smoking have?

HN2011-01 悪いマナーの例にはどのようなものが挙げられるか？
I want to know some examples of bad manners.

HN2012-01 ワインの産地を知りたい．有名もしくは個人的に好まれている地方のワインについて，特に知りたい．
Where are some wine production areas? I especially want to know about very famous or personally preferred areas.

HN2013-01 有名もしくは個人的評価の高い温泉地について知りたい．どこの地域・都道府県にあるのかが分かれば尚良い．
Where is a hot-spring resort that is very famous or highly rated? I also would like some information about the area or the name of the prefecture where the resort is located.

YY1004-02 予稿集や資料の訂正に関する内容を知りたい。
Tell me the information of correction in the proceedings and the handouts.

YY1013-01 登山をする場合の心構えについて知りたい。
Tell me the preparations for mountain climbing.

YY1014-01 毎日日課としてやっていることについて知りたい。
Tell me what everybody's daily tasks are.

YY1016-01 尊敬されている人やものについて知りたい
Tell me about someone or something to be respected.

YY1017-01 趣味になっていることについて知りたい
Tell me what everybody's hobbies are.

YI1012-01 日本語話し言葉コーパスを用いている研究を教えてください。
Tell me about some research using the Corpus of Spontaneous Japanese.

YI1014-01 DP マッチングを用いた研究を探したい
I want to search for some research using DP matching.

YI1015-01 音声認識システムを利用した実用的なアプリケーションを紹介している研究を教えてください。
Tell me about some research introducing practical applications that use a speech recognition system.

YI1016-01 音声認識にニューラルネットワークを導入した研究を知りたい。
I want to get some examples of research introducing neural networks to speech recognition.

YI1017-01 ニュース番組を音声認識して字幕化する研究を知りたい。
I want to know about some research for superimposing captions on a news program by using speech recognition.

TA1001-01 情報検索性能を評価するにはどのような方法があるか知りたい。
How can we evaluate the performance of information retrieval?

TA1003-01 日本語話し言葉コーパスにはどのような種類の講演が含まれているのか。
What kinds of lectures are included in the Corpus of Spontaneous Japanese?

TA1009-01 音声認識を応用したシステムにはどのようなものがあるか知りたい。
I want to know about some applied systems of speech recognition.

TA1015-01 機械翻訳の手法にはどのようなものがあるか。
Tell me some methods of machine translation.

TA1019-01 日本の都道府県庁所在地にはどのようなところがあるか。
Tell me some prefectural capitals in Japan.

TA1020-01 世界遺産にはどのようなところがあるか。
List some World Heritage sites.

KA1002-01 ベクトルによる言語処理を用いた自然文検索関係の発表にはどんなものがあるか？
Tell me some presentations related to natural language retrieval using vector space.

KA1005-01 音声認識率の有意差について言及している論文、または、有意差判定の方法について述べた論文を教えてください。
Please find papers that discuss the significant differences between speech recognition rates, or the method for determining the significant differences.

KI1001-01 マガーク効果とは何ですか？
What is the McGurk effect?

KI1003-01 基本周波数の抽出方法にはどのようなものがありますか？
Tell me about some methods for extracting the fundamental frequency.

A.2 Examples of the queries and the judgments of their relevancy

The slash "/" in the tables represents the boundary of the utterrances.

Utterances	Relevancy	Supporting Information
SDPWG-HN1014-01: 講演音声の特徴について知りたい (Tell me the characteristic of lecture speech?)		
それから講演音声は読み上げ音声のモデルよりも えーと 対話音声のモデルに近い発話スタイルに/なっていると/いうことも まー 言えると思います (Moreover, from the viewpoint of speaking style, lecture speech is more close to dialogue speech than read speech.)	Relevant	
えー その一方で ま 講演音声というものの特徴を考えていきますと えー ま 話し言葉の冗長的な表現というものを多く含みまして (Meanwhile, we have investigated the characteristics of lecture speech, and found that it contains redundant expressions deriving from spontaneous speech,)	Relevant	
講演である為に丁寧な/口調で話されておりますので丁寧語が/各所入っております (Since they talked at an academic conference, utterances were made in a polite manner and contained some polite expressions.)	Partially Relevant	

Utterances	Relevancy	Supporting Information
SDPWG-TA1001-01: "情報検索性能を評価するにはどのような方法があるか知りたい。" (How can we evaluate the performance of information retrieval?)		
他方が あーの 犠牲になるというような関係に基本的になりますでしたがって評価尺度の再現率と精度っていうのも普通は (... basically the relation is like that one improves at the cost of another. Therefore, the evaluation metrics, recall and precision, are usually ...)	Relevant	ですから ん えーっと いい検索システムというのは 両方の尺度ができるだけ高いと/いうことになります (So it can be said that a good retrieval system has high values for both the metrics.)
通常の情報検索システムの/出力 と/でよく使われる え 平均精度/で えー ランキングを評価する方法そして (The conventional output of an information retrieval system, ranking, is evaluated in terms of the average precision, and ...)	Relevant	検索結果を評価する基準ですが/でこれに関しても二通り (Talking about the criterions for evaluating the retrieval results, again, two kinds of methods ...)
でその日英検索の十一点平均適合率を/取ると (and when calculating the 11-point average precision of its Japanese to English Retrieval ...)	Irrelevant	*No supporting information.*

Utterances	Relevancy	Supporting Information
SDPWG-HN2010-02: "煙草が体に及ぼす影響，有害性にはどのようなものがあるか？" (How does smoking influence our health and what hazards does smoking have?)		
などの炭水化物の取り過ぎによってビタミン ビーワン 欠乏状態になる例が最近あるらしいので (There are examples of vitamin B1-deficiency caused by an excessive carbohydrate intake.)	Relevant	副腎皮質ホルモンの分泌を盛んにさせるストレスの増加や喫煙は/ビタミン シーをより多く消耗させるということなんですね (Increase of stress that activates adrenal cortex hormone secretion and smoking deplate more vitamin-C.)
たばこは/肺癌の七十二パーセント/喉頭癌の九十六パーセント/膀胱癌でさえ三十一パーセントの原因があると言われています (72% of lung cancer, 96% of larynx cancer, and 31% of bladder cancer are caused by smoking.)	Relevant	
ニコチンは/血管を収縮させ/血の巡りを悪くします (Nicotine constricts blood vessels and becomes be blockheaded.)	Relevant	煙には/ニコチン/さまざまな発癌物質/発癌促進物質/一酸化炭素/さまざまな線毛障害物質/その他/四千種以上の化学物質が含まれ/そのうち有害物質は/確認されただけでも/二百七十種あります (Cigarette smoke includes more than 4,000 kinds of chemical material such as nicotine, various cancerous substances, cancer promoter, carbon monoxide, fimbriae disorder substance and so on. The number of sorts of hazrdous substances of them is 270 at least.)

Utterances	Relevancy	Supporting Information
SDPWG-HN2012-01: "ワインの産地を知りたい．有名もしくは個人的に好まれている地方のワインについて，特に知りたい." (Where are some wine production area? I especially want to know about very famous or personally preferred areas.)		
フランスのシャンパーニュ地方で造られた/発泡性のワインてことであります (It is a sort of sparkling wine brewed in the area of Champagne in France.)	Relevant	次は東の横綱フランスワイン (Next, French wine, eastern king of wine ...)
あ 更に ワ ワイン以外で私大好きなのが あのシャンパンなんですけれどもこれはフランスはシャンパーニュ地方の (I love also champagne, as well as wine. The area of Champagne in France ...)	Relevant	
あのー 南フランスのバイヨンヌっていうところなんですけども/あの 非常に田舎町でして (Bayonne in southern France. It is a very rural town.)	Partially Relevant	いうことでですね あのー ま 結構ワイン/がおいしいと (That's it. Wine is very delicious.)

Utterances	Relevancy	Supporting Information
SDPWG-YY1016-01: 尊敬されている人やものについて知りたい (Tell me about someone or something to be respected.)		
多分父親本当に尊敬してる人は父親だけだと思うんですけど (My father, I think. The only person who I really respect is my father.)	Relevant	
合掌造りの里とか曲がり屋とか (Japanese traditional villages that consists of houses roofed with thatch grass (Gassho-Zukuri or Magariya).)	Relevant	.../あー 私達が見て/とても尊敬 し/に値すると思います (I think they are of great worth to be respected.)
ま ケー 理事長 (Probably a director.)	Relevant	あ 尊敬する二人のトップを横軸に/話をしてみたい/と/思います/あ まず (I am going to begin my talk focusing on two directors I respect.)

Utterances	Relevancy	Supporting Information
SDPWG-YI1014-01: DP マッチングを用いた研究を探したい (I want to search for some research using DP matching.)		
でこれは二つのモジュールからなってまして 第一段階で統合モジュール/これにより ディーピー マッチングを行ないまして各システムが出す単語列というものの対応を取ります (Then, this is composed of two modules, and at the first step, an integration module/ DP matching is performed by this module, and an alignment is obtained between word sequences generated by each system.)	Relevant	
と 本日の発表で/< 雑音 >/えー/主眼を置いているのはこの/ディーピー マッチングを おく 連続 ディーピー を行なう際の距離尺度なんですがここを/色々と/< 雑音 >/変えてみようと考えています (and what is the focus, well, in today's presentation, is a distance measure in performing DP matching, Continuous DP, and we are going to try the various measures.)	Relevant	キーワードを/えー ディーピー マッチング連続 ディーピー を行なった結果 き/えー 得られたパスというものはこのように も (As the result of DP matching, Continuous DP, well, for a keyword, a path is obtained like this and ...)
システムの方で音声区間抽出/え ディーピー マッチングを行ない/整合経路の表示を行ないます/これが/その ディーピー マッチングをした時の結果の例/です (By the system, voice activity detection, well, and DP matching are performed, a consistent path is displayed. This is an example of the result of DP matching.)	Relevant	

(Received June 4, 2008)
(Accepted November 5, 2008)
(Released February 4, 2009)

Tomoyosi Akiba received the B.E. degree in 1990 in information science, the M.E. degree in 1992, and the Ph.D. degree in 1995 in system science from Tokyo University of Technology, Tokyo, Japan. In 1995, he became a Researcher in the Electrotechnical Laboratory, MITI, Japan. In 2004, he became an Associate Professor in Toyohashi University of Technology. His research interests include natural language processing and spoken language processing. He is a member of ISCA, Institute of Electronics, Information and Communication Engineers (IEICE), Japanese Society of Artificial Intelligence (JSAI), Acoustical Society of Japan (ASJ), and the Association for Natural Language Processing (NLP).

Kiyoaki Aikawa received his Ph.D. from the University of Tokyo in 1980. He engaged in the NTT Basic Research Laboratory in 1980. He was a visiting scientist of Carnegie Mellon University in 1990. From 1992 to 1995 he was a senior researcher in Advanced Telecommunications Research Laboratories. He stayed in NTT Laboratories from 1996 to 2002. He is a professor of School of Media Science and the director of Media Center at Tokyo University of Technology. He received the Sato Award from the Acoustical Society of Japan. He received the Telecom-System Technology Award form the Electrical Communication Foundation. He is a member of IEICE, ASJ, ASA, and IEEE.

Yoshiaki Itoh received the B.E. degree, M.E. degree and the Dr.Eng. from the University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1999, respectively. From 1989 to 2001, he was a researcher and a staff of Kawasaki Steel Corporation, Tokyo and Okayama. From 1992 to 1994, he transferred as a researcher to Real World Computing Partnership, Tsukuba, Japan. He has been an Associate Professor in the Faculty of Software and Information Science at Iwate Prefectural University, Iwate, Japan, from 2001. He is a member of IEEE, ISCA, Acoustical Society of Japan (ASJ), Institute of Electronics, Information and Communication Engineers (IEICE), and Japan Society of Artificial Intelligence (JSAI).

Tatsuya Kawahara received the B.E. degree in 1987, the M.E. degree in 1989, and the Ph.D. degree in 1995, all in information science, from Kyoto University, Kyoto, Japan. In 1990, he became a Research Associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Adjunct Professor in the School of Informatics, Kyoto University. He has published more than 150 technical papers covering speech recognition, spoken language processing, and spoken dialogue systems.

Hiroaki Nanjo received the B.E. degree in 1999, the M.E. degree in 2001, and the Ph.D. degree in 2004 from Kyoto University, Kyoto, Japan. In 2004, he became a Research Associate at Department of Media Informatics, Faculty of Science and Technology, Ryukoku University. Currently he is an Assistant Professor at Ryukoku University. He has been working on speech recognition and understanding. He is a member of Acoustical Society of Japan (ASJ), Institute of Electronics, Information and Communication Engineers (IEICE), and Institute of Electrical and Electronics Engineers (IEEE).

Hiromitsu Nishizaki was born in 1975. He received his B.E., M.E., and Dr.Eng. degrees in information and computer sciences from Toyohashi University of Technology in 1998, 2000, and 2003. He is now an assistant professor in the Dept. of Research Interdisciplinary Graduate School of Medicine and Engineering at University of Yamanashi. His research interests include spoken language processing.

Norihito Yasuda received the bachelor's degree in integrated human studies and the master's degree in human and environmental studies from Kyoto University in 1997, and 1999. In 1999, he joined Nippon Telegraph and Telephone Corporation (NTT), Japan. His research interests include natural language processing, information retrieval and spoken language processing. He is a member of Acoustical Society of Japan (ASJ) and Japanese Society of Artificial Intelligence (JSAI).

Yoichi Yamashita received the B.E., M.E. and Dr.Eng. degrees from Osaka University in 1982, 1984 and 1993, respectively. He has worked for the Institute of Scientific and Industrial Research of Osaka University as a Technical Official, a Research Associate, and an Assistant Professor from 1984 to 1997. In 1997, he joined Ritsumeikan University as an Associate Professor in the College of Science and Engineering. He is currently a Professor in the College of Information Science and Engineering. His research interests include speech understanding, speech synthesis and spoken dialog processing. He is a member of IEICE, ASJ, JSAI, ISCA and IEEE.

Katunobu Itou received the B.E., M.E. and Ph.D. degress in computer science from Tokyo Institute of Technology in 1988, 1990 and 1993 respectively. From 2003 to 2006, he was an associate professor at Graduate School of Information Science of the Nagoya University. In 2006, he joined the Faculty of Computer and Information Sciences at Hosei University, Japan, as a Professor. His current research interest is spoken language processing. He is a member of the Acoustical Society of Japan.