Constructing Japanese Test Collections for Spoken Term Detection

Yoshiaki Itoh[1], Hiromitsu Nishizaki[2], Xinhui Hu[3], Hiroaki Nanjo[4], Tomoyosi Akiba[5], Tatsuya Kawahara[6], Seiichi Nakagawa[5], Tomoko Matsui[7], Yoichi Yamashita[8], and Kiyoaki Aikawa[9]

[1] Iwate Prefectural University, Japan
[2] University of Yamanashi, Japan
[3] NICT, Japan
[4] Ryukoku University, Japan
[5] Toyohashi University of Technology, Japan
[6] Kyoto University, Japan
[7] The Institute of Statistical Mathematics, Japan
[8] Ritsumeikan University, Japan
[9] Tokyo University of Technology, Japan
[1] y-itoh@iwate-pu.ac.jp

Abstract

Spoken Document Retrieval (SDR) and Spoken Term Detection (STD) have been two of the most intensively investigated topics in spoken document processing research according to the establishment of the SDR and STD test collections by the Text REtrieval Conference (TREC) and NIST. Because Japanese spoken document processing researchers also requires such test collections for SDR and STD, we have established a working group to develop these collections in Special Interest Group -Spoken Language Processing (SIG-SLP) of the Information Processing Society of Japan. The working group has constructed and made available a test collection for SDR, and is now constructing new test collections for STD that will be open to researchers. The present paper introduces the policies, outline, and schedule of the new test collections. Then, the new test collections are compared with the NIST STD test collections.

Index Terms: spoken term detection, test collection

1. Introduction

In recent years, the amount of video content available on PCs has increased rapidly, owing to the extremely large capacity of video recorders as well as to video sites available via high-speed Internet. For example, remote video lectures, television program, and video blogs are now available. Faced with this increase in content, users must be able to search efficiently for desired content from a vast amount video data. Spoken Document Retrieval (SDR) and Spoken Term Detection (STD) are the most promising approaches for this purpose. The Text REtrieval Conference (TREC) has dealt with SDR since 1996 [1]. The task of SDR is to identify a target spoken document from among a large number of spoken documents, where the target is often defined as containing a particular topic, for example, the content of a news clip. However, SDR has limitations: even if the target spoken document is identified, the user must browse the entire spoken document to confirm its content, even if the user would rather browse only the section containing the keyword of interest. Therefore, STD functionality to detect the target section that contains a spoken query term is needed. NIST has set up STD test collections and collected the results of conference attendees [2]-[4]. Research and development of SDR and STD has been actively carried out, owing to the construction of these test collections.

In SDR and STD research, a standard test collection is indispensable because the performance of SDR and STD depends on the query set, the size and category of the spoken documents, and the correct documents or sections, and the occurrences in spoken documents that are included in such a test collection. Test collections for Japanese are highly desired by the spoken documents processing community in Japan.

For this purpose, a working group to construct a test collection for SDR was established in the SIG-SLP of the Information Processing Society of Japan. The working group constructed and made available a test collection for SDR [5] in 2008. The test collection was designed for an ad hoc retrieval task where the relevant sections from about 2700 presentation speeches in academic conferences are identified when a text query sentence is given. We are now constructing new test collections for STD which will also be open to researches. Such presentation speeches include special terms that rarely occur in other domain, such as news broadcasts [6], and unique language models cannot be trained from ordinal text corpora for such speeches. We established realistic tasks for STD by selecting several sets of such a meaningful term as a query term. In this paper, we describe the basic policies of this construction, discuss the outline of the design in detail, and report the schedule of the new test collections that will contain three types of test collections including query and correct section sets. In the following section, the Japanese test collections for STD are described in detail and then compared with the NIST STD test collections.

2. Japanese Test Collections for STD

2.1. Basic policy

The basic policy for constructing the Japanese test collections is “to supply plural and simple queries and relevant section sets under the assumption that various researchers and developers use them for SDR and STD,” so that the test collections can be used by not only current STD researchers but also researchers who begin STD research in the future. Therefore, we are planning to prepare the following Japanese test collections that mainly consist of sets of query terms and their relevant sections.

- Query term sets
- Indication of range of spoken documents
- Relevant section sets
- Transcripts generated by an automatic speech recognizer (ASR) and the acoustic and language models in the ASR
- Baseline results

NIST test collections were made available to researchers after submission of their STD systems; therefore, whether a query term is out of vocabulary (OOV) or in vocabulary (IV) is unknown. The performance largely depends on the language models of the researchers, especially the vocabulary. The total STD performance was obtained without difficulty in the NIST evaluation. To evaluate the difference between each methodology clearly, such as the difference between acoustical models, subword language models, or models/ methods for OOV terms, we make available simple test collections to researchers beforehand.

2.2. Constructing Japanese test collections

For spoken documents, the “Corpus of Spontaneous Japanese (CSJ)” [7] is assumed to be used. CSJ is basically a corpus of monologues and includes 2702 speeches (604 hours) in total that consist of speech from actual academic presentations and simulated public speech. The speech from academic presentations was recorded live at proceedings of nine academic societies. In this paper, each presentation speech is regarded as a spoken document. The subset of CSJ, the so-called “CORE”, includes 177 spoken documents and detailed annotations are given in addition to transcription. Therefore, annotations in the CORE data allow other approaches for STD such as a method using intonation and so on. We use “CORE” and “ALL” to denote the 177 spoken documents and the entire collection of 2702 spoken documents, respectively, in the following sections. We assume CSJ spoken documents are obtained separately by the user.

All speeches except CORE parts were divided into two groups according to the speech ID: an odd group and an even group. We constructed two sets of acoustic models and language models, and performed automatic speech recognition using the acoustic and language models trained by the other group. We used Julius [8] as a decoder, with a dictionary containing 27k vocabulary, and obtained N-best speech recognition results for all spoken documents. The followings can be made available to researchers.

-N-best transcripts generated by the ASR
-Odd acoustic models, language models
-Even acoustic models, language models
-a dictionary of the ASR

2.3. Query and correct sets

Currently, we have constructed 5 sets of query and correct section pairs, as follows. When selecting query terms, we mainly take into account of the length, occurrence, meaning of a query term, and IV or OOV of the ASR.

(1) In-vocabulary (IV) query term set for ALL: 100 query terms included in the ASR dictionary for all spoken documents.
(2) IV query term set for CORE: 50 query terms included in the ASR dictionary for CORE spoken documents.
(3) Low-frequency Out-of-vocabulary (OOV) query term set for ALL: 50 query terms that are not included in the ASR dictionary for all spoken documents.
(4) Low-frequency OOV query term set for CORE: 50 query terms that are not included in the ASR dictionary for CORE spoken documents.
(5) Noun set for performance evaluation: 50 query terms.

Below, the principle for selecting each query and correct section set is described in detail.

(1) IV query term set for ALL

For all 2702 spoken documents, we extracted 100 query terms that are assumed to correspond to query terms in actual situations. We take the Term Frequency (tf) in a document and the Inverse Document Frequency (idf) into account. Each query term consists of one or more words, and each word in a query term is included in the ASR dictionary. Because the STD performance depends on the length of the query terms, we selected queries of differing length. All candidates are categorized into a group by the length in morae, and about 12 query terms (length: 4 to 12 morae) are selected from each group. Some examples are shown in Table 1. In the column of the query term, the slash “/” denotes the boundary between words, and capitalized words denote a proper noun. Underlined words indicate foreign words (non-Japanese).

Table 1. *Example of query terms for IV query term set for ALL spoken documents.*

Length (morae)	Query term (meaning)
12	dai/goi/oNsei/niNshiki (large vocabulary speech recognition)
11	oNsei/taiwa/shisutemu (spoken dialog system)
10	juuyou/buN/chuushutu (important sentence extraction)
9	keitai/so/kaiseki (morphological Analysis)
8	kihoN/shuha/suu (pitch frequency)
7	iNtone:shoN (intonation)
6	shouteN/gai (shopping center)
5	obaachaN (grandmother)
4	chika/tetsu (subway)

Table 2. *Example of query terms for low-frequency OOV query term set for ALL spoken documents.*

Length (morae)	Query term (meaning)
12	HoNkomagome/HakusaN/chiku (Place-name place-name district)
11	Asuka/shiNnou/iQkou(Surname imperial prince’s group)
10	Sugisawa/Yotaro/shi (Mr. Given-name Surname)
9	FuteNma/kichi/isetsu(place-name base transfer)
8	jakuniku/kyoushoku (survival of the fittest)
7	zeNdai/mimoN (unparalleled in history)
6	Fujiko/Fujio (Given-name Surname)
5	ToranomoN (Place-name)
4	Sachiyo (Given-name)

(2) IV query term set for CORE

For CORE 177 spoken documents, we extracted 50 query terms included in set (1). This is a smaller task than (1). About 8 query terms are selected for each group with length from 4 to 10 morae. Several query terms with 11 and 12morae were included.

(3) Low-frequency IV query set for ALL

(4) Low-frequency IV query set for CORE

Because the cutoff number was 3 in the ASR described in (1), the words that appear less than 4 times in the odd or even spoken documents were not included in the ASR dictionary. These query terms are so-called unknown OOV words. An OOV query term is a word sequence, in which at least one word is OOV. We extracted 50 query terms including unknown words for both the 177 CORE and ALL 2702 spoken documents. Query terms are mainly proper nouns, such as location names, personal names, foreign expression, and new terms, as shown in Table 2.

(5) Noun set for performance evaluation

Some experiments for 50 noun query terms were conducted, and researchers can easily compare results obtained by their methods with these results. The 50 query terms were manually selected among 49 spoken documents (about 13 hours), taking into consideration meaningful terms that are assumed to be query terms.

Table 3. *Term frequency, document frequency, tf-idf, and performance of each query term in IV query term set for ALL spoken documents.*

Query term	tf	df	tf idf	hit/ output	rcl	Prc sn
12 morae						
1.kokuritu/kokugo/keNkyuu/sho	35	19	174	16/16	46	100
2. toukei/suuri/keNkyu/sho	10	8	87	5/5	50	100
4. dai/goi/oNsei/niNshiki	6	5	37	4/4	67	100
11 morae						
1. oNsei/taiwa/sisutemu	89	34	389	71/71	80	100
6. kikai/hoNyaku/sisutemu	24	7	143	12/13	50	92
12. kaNkyo/oN/no/shikibetu	8	1	63	0/0	0	0
10 morae						
1. tii/efu/ai/dii/efu	79	18	396	48/50	61	96
6. Sidonii/oriNpiQku	29	16	149	15/16	52	94
12. DorutoN/no/geNshi/setu	7	1	55	0/0	0	0
9 morae						
1. keitai/so/kaiseki	159	76	568	134/13	84	100
6. waakiNgu/horidee	27	10	165	14/14	52	100
12. Eberesuto/kaidou	6	1	47	2/2	33	100
8 morae						
1.kihoN/shuuha/suu	287	61	1088	225/22	78	100
6. iNtaarakushoN	61	27	281	48/52	79	92
12. chuuou/riNkaN	12	4	78	1/1	8	100
7 morae						
1. iNtoneeshoN	199	50	114	186/19	93	98
6. tounaN/Azia	44	28	201	4/4	9	100
12. oshuutome/saN	15	6	92	7/7	47	100
6 morae						
1.shouteN/gai	208	75	742	133/13	64	96
6. peQto/botoru	36	19	178	22/23	61	96
12. chouoN/chizu	12	1	95	4/4	33	100
5 morae						
1. obaachaN	244	100	804	177/20	73	88
6. koNkuuru	39	17	205	29/30	74	97
12. reNsou/go	16	1	126	0/0	0	0
4 morae						
1. chika/tetsu	134	70	486	83/85	62	98
6. Aoyama	58	16	292	25/26	43	96
12. yashi/no/ki	18	12	98	15/17	83	88

2.4. Sample STD performance

We conducted a simple evaluation experiment for the IV query terms mentioned in (1) and (2). As described in Section 2.2.1, all spoken documents were transformed into text transcripts by an ASR. The 1-best result was used for this evaluation.

Because all the query terms are included in the ASR dictionary, the Unix command "grep" can be used to detect the sections containing the query term. The information from time stamps at the utterance level can be utilized here. The time stamps for each utterance are tagged in CSJ; this is discussed in more detail in the Section 3. We simply regarded a "grepped" section as correct when the detected section was included in the utterance where the query term was spoken.

The tf, document frequency (df), tf-idf value and the performance of each query term in a known query term set for all spoken documents are shown in Table 3. The terms "hit" and "output" correspond to "correctly detected query term" and "occurrence in spoken documents as a result of ASR", respectively. There are 4 query terms of 12 morae, and 12 query terms of other lengths. Table 3 shows only information on 3 terms, namely, the highest, median and lowest tf for each length of query term. Underlined words indicate foreign (non-Japanese) words that are included in 44 query terms among the 100 query terms.

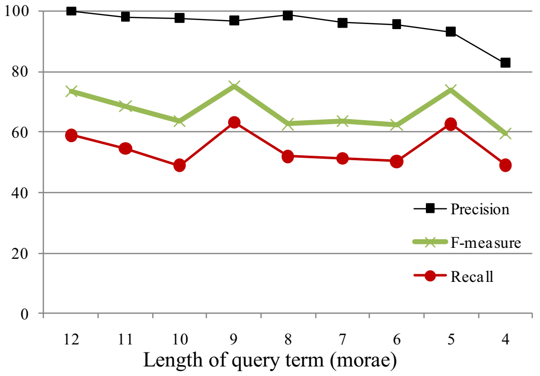

Figure 1. STD performance for each morae length of IV query terms for ALL spoken documents.

The average precision rate (prcsn), recall rate (rcl), and F-measure for each morae length are computed, as shown in Fig. 1. Here, we excluded zero values when averaging precision and F-measure. Although the STD performance in the recall and the precision rates was expected to decline when the length of a query term was short, a difference could not be observed in the recall rate. As the length of the query terms was decreased, a greater number of misrecognized short query terms were observed and the precision rate slightly declined, as shown in the figure. The total STD performance for IV query terms is as follows.

(1) IV query term set for ALL
  Recall: 49.8 (0.0-93.5), Precision: 87.6 (22.7-100.0)
  F-measure: 63.5 (12.5-95.6)
  ASR performance: word accuracy, 74.05%;
  percentage correct, 69.29%.

(2) IV query term set for CORE
  Recall: 54.1 (0.0-93.8), Precision: 93.6 (66.7-100.0)
  F-measure: 68.6 (20.0-96.8)
  ASR performance: word accuracy, 76.68%;
  percentage correct, 71.93%.

As shown above, the precision rates for query terms with more than 7 morae were almost 100.0% (no false alarms), because there were very few cases where a series of words was misrecognized as query terms. Due to the high precision rate and the long spoken documents whose total length is greater than 2 million seconds, the term of false alarms in Actual Term Weighted Value (ATWV) became negligible and the ATWV value became almost the same as the recall rate because β/(T-Ntrue) is less than 1/2000 for all data and less than 1/140 for CORE data. As a result, ATWV is equal to the recall rate: 49.8 for case (1) and 54.1 for case (2).

3. Discussions and Schedule

3.1. Evaluation metrics

For the processing resources [4], the following metrics are considered adequate.

Indexing time
Indexing memory consumption
Index size
Search time for each term
Searching memory consumption

To evaluate the detection performance of the entire system, the followings are used as metrics of STD performance: precision rate, recall rate, F-measure, average precision for each query term, MAP (Mean Average Precision), DET (Detection Error Tradeoff), ATWV, MTWV (Maximum Term Weighted Value) and so on. The metrics ATWV and MTWV have been

Table 4. Comparison of recent test collections for SDR and STD

Name	What kind of retrieval	Query type Num. of query	Spoken documents (SD)	Evaluation metric
TREC-6	Known word retrieval	47 topics	Mainly news speech 1451 spoken documents (50 hours)	Correct rate for the best candidate
TREC-7	Ad-hoc retrieval	23 topics (14.7 words/topic)	Mainly news speech 2866 spoken documents (87 hours)	MAP(Mean Average precision)
TREC-8	Large scale ad-hoc SDR	49 topics	Mainly news speech , no boundaries 21,754 spoken documents (557 hours)	MAP
NIST STD	STD	About 1000 1～4words/query	Broadcasted news 3 hours, Tel. conversation 3 hours, Roundtable meetings 2 hours for English (Arabic and Chinese sets are prepared)	DET ATWV
JPN SDR	Large scale ad-hoc SDR	39 topic	Mainly presentation speech, 2702 spoken documents (604 hours) Relevant passages are regarded as correct	11 points MAP F-measure, et al.
JPN STD	STD	5 sets of 50 or 100 query terms	Mainly presentation speech 2702 spoken documents (604 hours) and 177 spoken documents (44 hours)	MAP, F-measure, ATWV et al.

introduced recently, and the following problems remain with respect to the evaluation of STD performance, as mentioned in [8].

- Fair comparison between a query term with many occurrences and a term with few occurrences in spoken documents
- Fair comparison between correct hits and false alarms
- Consistent evaluation of the size of spoken documents

Accordingly, new and more appropriate evaluation metrics are necessary.

3.2. Length of spoken documents and occurrence for OOV query terms

The STD of OOV terms is a critical task because such query terms are likely to be OOV terms [9]. The performance should be simple to evaluate using test collections only for unknown query terms, and we prepared two sets of test collections for all spoken documents and CORE spoken documents. Query terms that appear less than 3 times are regarded as OOV terms. Searching for these low-frequency terms from 2702 spoken documents (604 hours) appeared to be too difficult. Table 4 shows a comparison between the present test collections and those of NIST for SDR and STD. Because the spoken documents for Japanese STD are much larger than those for NIST STD, we are planning to set tasks to search query terms for one spoken document that lasts about 12 minutes on average, or longer spoken documents, assuming that the spoken document is identified by the SDR method. Here, the spoken document that includes at least one query term is called the target document. We add non-target documents that do not include a query term to a target document and can provide 1-hour or 10-hour spoken documents for another task.

3.3. Judgment of true occurrences

In the NIST evaluation design, true occurrences are judged according to the following rule: the gap between adjacent words in a query must be less than 0.5 seconds in the corresponding speech. In CSJ, there are time stamps at the utterance level but not the word level. The length of utterances is not long. We can simply regard a detected section as correct when the detected section is included in the utterance where the query term is spoken. For more accurate evaluation, we are planning to obtain time stamps at the word level by performing forced alignment using ASR. The accuracy of the time stamps must be checked and the usability of the time stamps confirmed.

3.4. Schedule for opening test collections

The setup of the query terms is nearly complete. We are now preparing the baseline results and time stamps, as mentioned above, and are going to open the test collections to researchers in sequence. All test collections will be completed by July 2010 at http://www.cl.ics.tut.ac.jp/~sdpwg/.

4. Conclusions

The paper described the Japanese test collections for Spoken Term Detection that are now completed by the SIG-SLP working group of the Information Processing Society of Japan. These test collections are the second test collections successfully developed for SDR. The policies, the outline, and the schedule of the new test collections were explained. The new test collections for STD enable researchers to use large spoken documents, in comparison with NIST STD test collections, and will be open to researchers. We expect the Japanese test collections to facilitate new research into STD and SDR.

5. References

[1] Garofolo, J. S., et.al, V. M. and Jones, K. S., "TREC-6 1997 Spoken Document Retrieval Track Overview and Results", NIST Special Publication, vol. 500, no. 240, pp. 83–92, 1998.

[2] Garofolo, J. S., Auzanne, C.G.P. and Voorhees, E.M., "The TREC spoken document retrieval track: A success story," Ninth Text Retrieval Conference (TREC-9), NIST, 2000.

[3] NIST, "The Spoken Term Detection (STD) 2006 Evaluation Plan," http://www.nist.gov/speech/tests/std/docs/std06evalplan-v10.pdf, 2006.

[4] 2006 Spoken Term Detection Evaluation: http://www.itl.nist.gov/iad/mig/tests/std/2006/index.html.

[5] Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita Y. and Itou, K., "Construction of a test collection for spoken document retrieval from lecture audio data," IPSJ Journal Vol.50 No.2 1234-1245, 2009.

[6] Glass, J., et al., "Recent progress in the MIT spoken lecture processing project," Proc. of Interspeech, pp.2253-2256, 2007.

[7] Maekawa, K., "Corpus of Spontaneous Japanese", Its design and evaluation. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. 7–12, 2003.

[8] Miller, D. et al., "Rapid and Accurate Spoken Term Detection", Proc. of Interspeech, pp. 314–317, 2007.

[9] Logan, B., et al., "Confusion-based query expansion for OOV words in spoken document retrieval," Proc. ICSLP, 2002.