

Experiments on Web Retrieval Driven by Spontaneously Spoken Queries

Tomoyosi Akiba

Department of Information and Computer Sciences, Toyohashi University of Technology
1-1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, 441-8580, JAPAN
akiba@c1.ics.tut.ac.jp

Atsushi Fujii, Tetsuya Ishikawa

Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, JAPAN

Katunobu Itou

Graduate School of Information Science, Nagoya University
1 Furo-cho, Nagoya, 464-8603, JAPAN

Abstract

Motivated to realize the speech-driven information retrieval systems that accept spontaneously spoken queries, we developed a method to collect such speech data derived from the pre-defined search topics that had been systematically constructed for IR research. In order to evaluate both our method and the performance of the document retrieval by using the spontaneously spoken queries, we took place two experiments of collecting the speech data by our method using publicly available test collections of evaluating document retrieval. The first preliminary experiment took place with relatively small number of search topics selected from the NTCIR-3 Web retrieval collection, in order to test our method. The second experiment took place with all of the search topics released from the NTCIR-4 Web task to participate the formal run of the evaluation. The information about the collected data and the result of the evaluation with respect to both the speech recognition accuracy and the precision of document retrieval by using the collected data are presented in this paper.

1 Introduction

This paper describes our speech-driven information retrieval system participated in the NTCIR-4 Web task. We previously evaluated a Web retrieval system driven by read (not spontaneously spoken) queries [4]. We are enhancing our system for spoken queries, which are more realistic than read speech.

Automatic speech recognition has recently become a practical technology. A number of speech-based methods have been explored in the information retrieval (IR) community, which can be classified into the two fundamental categories. The first category is spoken document retrieval (SDR), in which text queries are used to search speech archives for relevant information, and the second category is spoken query retrieval (SQR), in which spoken queries are used to retrieve relevant text information. Initiated partially by the TREC-97 SDR track [6], various methods have been proposed for spoken document retrieval. However, a relatively small number of methods have been explored for speech-driven text retrieval [3, 5]. Furthermore, none of the existing methods use the spontaneously spoken queries as inputs for IR systems.

In this paper, we mean a spontaneously spoken query, or a query in spontaneous speech, the speech uttered before/during thinking what-to-say and how-to-say. An advantage of the use of spontaneously spoken queries is that it enables users to easily submit long queries to provide IR systems rich clues for retrieval. Unconstrained speeches are commonly used in daily life. Another advantage is that spontaneously spoken queries allow users to start searching even if they cannot clearly express their needs. Taylor [8] categorized information need in four levels, which are visceral, conscious, formalized and compromised needs. Ideally both the conventional keyboard-based retrieval and the speech-driven IR systems should be queried by the visceral need. However, existing IR systems are intended for the compromised or, at best, the formalized need. Our IR system queried by spontaneous speech

can also target the conscious need, because users can start speaking and searching based on their unclear need and make the need more concrete progressively.

Section 2 describes our method to collect spontaneously spoken queries from subjects using the pre-defined search topics for document retrieval. Section 3 describes our experiments of collecting the queries by using our method.

2 Collecting Spontaneously Spoken Queries

For research and development purpose, in which we enhance our retrieval system progressively by means of experiments, a large collection of spontaneously spoken queries are needed. To collect read speech, human subjects are requested to speak prepared scripts, as performed in our previous work [4, 5]. However, collecting spontaneous speech is more difficult than collecting read speech, because by definition it is impossible to prepare scripts for spontaneous speech in advance.

In addition, to make use of relevance judgments performed for text search topics, user speeches must be associated with those topics. In this sense, user utterances must be controlled to a certain extent.

Our solution is that, instead of the literal word sequence, we make users understand the meaning of search topics and then make them freely speak their own expression about the topics. In order to avoid users to memorize the word sequence of search topics literally, we used relatively long and rich explanation of search topics and placed an interval between the stage of understanding and that of speaking in our experiment.

The steps of our experiment is as follows:

1. Provide a subject with a written search topic (a script),
2. Give them 30 seconds to understand the content,
3. Take the script away,
4. Give them another 30 seconds, in which they were allowed to recall the content and think about what-to-say and how-to-say,
5. Make them speak a query about the topic,
6. Make them utter the phrase “that’s all”, when they think sufficient information is provided.

Because our main target was collecting “spontaneous” speech, we carefully designed the protocol not to restrict what the subjects speak. In the experiment, we told the subjects that what-to-say and how-to-say are up to them as long as the content is associated with the script. The subjects were allowed to speak as

```
<TOPIC>
<NUM>0008</NUM>
<TITLE CASE="b">Salsa, learn, methods
</TITLE>
<DESC>I want to find out about methods for
learning how to dance the salsa</DESC>
<NARR><BACK>I would like to find out in
detail how best to learn how to dance the
salsa, which is currently very popular.
For example, if I should go to dance
classes, I need detailed information such
as where I should go and what the class
would be like.</BACK>
<RELE>Documents simply saying that it
is popular without giving any detailed
information are irrelevant.</RELE></NARR>
<CONC>Salsa, learn, methods, place,
curriculum</CONC>
<RDOC>NW011992774, NW011992731, NW011992734
</RDOC>
<USER>1st year Master's student, female,
2.5 years search experience </USER>
</TOPIC>
```

Figure 1. An example search topic in the NTCIR-3 Web collection.

many contents as they like and repeat/modify the same content at step 5. They were also allowed to pauses queries. The subjects were encouraged to speak as informative content as possible to improve the retrieval accuracy.

3 Experiments

3.1 Search Topics

We used search topics produced for the Web tasks at NTCIR-3 and NTCIR-4. Each search topic is in SGML-style form and consists of the topic ID (<NUM>), title of the topic (<TITLE>), description (<DESC>), narrative (<NARR>), list of synonyms related to the topic (<CONS>), sample of relevant documents (<RDOC>), and a brief profile of the user who produced the topic (<USER>). Figure 1 depicts an English translation of an example Japanese topic. Although Japanese topics were used in the main task, English translations are also included in the Web retrieval collection mainly for publication purposes.

In our previous work [4], we collected the read speech by using the NTCIR-3 Web retrieval collection, in which the subjects read only the description field. However to collect spontaneously spoken queries, we used both the description and narrative fields as scripts (see Section 2). We performed two experiments, in which NTCIR-3 and NTCIR-4 Web collections were used, respectively.

Table 1. Statistics of spoken queries for the 12 selected topics using NTCIR-3 Web collection and arbitrary topics.

Subject ID	12 selected topics			Arbitrary topic (sec.)
	Min	Max	Mean	
Female#1	52.1	98.4	73.3	115.0
Female#2	14.0	44.5	29.4	50.7
Male#1	14.1	36.6	20.3	23.9
Male#2	38.3	86.7	58.2	37.9

Table 2. OOV and WER of spoken queries for the NTCIR-3 Web collection.

Subject ID	12 Selected topics		Arbitrary topic	
	OOV (%)	WER (%)	OOV (%)	WER (%)
Female#1	4.8	49.9	5.7	51.3
Female#2	1.3	21.7	2.0	32.9
Male#1	1.6	25.9	2.6	25.0
Male#2	5.8	57.1	3.8	48.8
Average	3.4	48.7	3.5	39.5

3.2 Preliminary Experiment using NTCIR-3 Web collection

For a preliminary experiment, we collected the spontaneously spoken queries using search topics in the NTCIR-3 Web retrieval test collection. The subjects of this experiment were four (two males and two females) university students. Out of the 105 search topics, 12 topics were selected and used to collect spontaneous speech. We also collected spoken queries for an arbitrary topic for each subject, in order to investigate the difference between the utterances corrected by controlled and no-controlled manner.

The statistics of the resultant spoken queries are shown in Table 1.

To transcribe read and spontaneous speech automatically, we used an existing Japanese speech recognition system [7]. The language model used was produced from the 10M Web pages in the NTCIR Web test collection [4].

Both out-of-vocabulary rate (OOV), which is the ratio of query words not included in the language model and the total number of query words, and word error rate (WER), which is the ratio of errors and the total number of query words, are shown in Table 2.

Compared with our previous results obtained with the read queries [4], in which OOV and WER were 0.73 % and 13.1 %, respectively, the recognition of spontaneous speech was harder. In addition, OOV and WER varied significantly depending on the human subject and the search topic. We did not find significant differences between the results from selected topics and that from an arbitrary topic.

We used the document retrieval system [4] to investigate the retrieval accuracy for the following input types.

- (a). written search topics, for which the description field tagged with <DESC> were used,
- (b). read speech transcribed by speech recognition,
- (c). spontaneous speech transcribed manually,
- (d). spontaneous speech transcribed by speech recognition.

The retrieval results were evaluated by mean average precision (MAP), which were non-interpolated average precision averaged over the 12 search topics. Note that the MAP value of the read speech (b) was obtained by averaging the four results by two females and two males who were different from the subjects of collecting spontaneous speech. Figure 2 shows the MAP values for the different input types above.

The MAP values for (b) and (c) were two thirds of that for (a). The MAP value for (d) was one third of that for (a). A reason why the results of (c) and (d) were inferior to that of (a) is that we did not participated in the pooling with the result obtained by (c) and (d).

3.3 Experiment using the NTCIR-4 Web collection

We collected spontaneously spoken queries for all 153 search topics in the NTCIR-4 Web retrieval task. Using the manually transcribed spoken queries as the

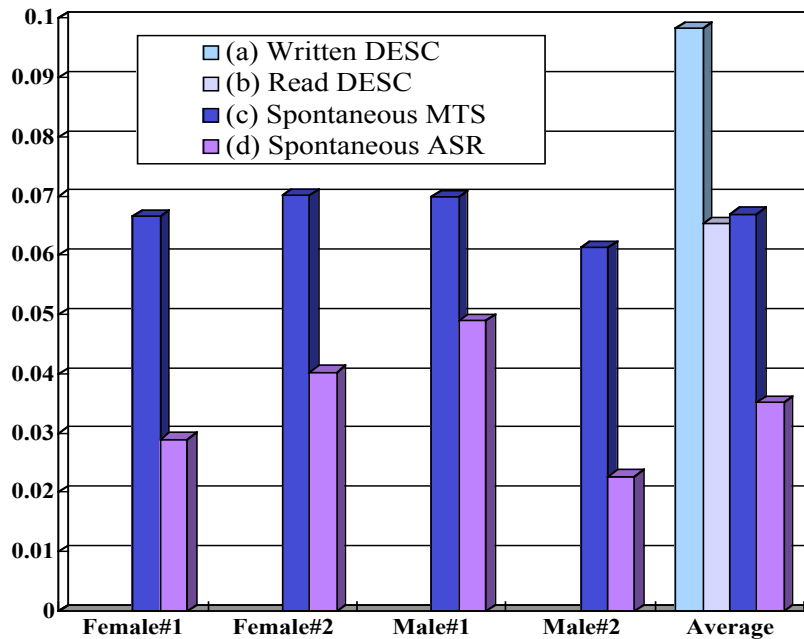


Figure 2. Mean Average Precision (MAP) for the NTCIR-3 Web collection.

inputs, we participated in the NTCIR-4 Web task (as an optional run of using an interactive system). Unlike the experiments in Section 3.2, our submitted documents were used for the pooling the relevance judgment.

In order to participate in the formal evaluation task, we force the subject's queries more consistent with the topics than that of the preliminary experiment. We told the subjects to divide their information need into the part that was faithful to, and did not include any excessive need out of, the search topic shown, and the additional part that could include additional needs they want to know about the topic. The subjects were told to speak the two parts separately; firstly, they should speak faithful need to an indicated topic, then keyword "that's all", in succession the additional need, and finally the keyword again, in this order.

The subjects were eight (four males and four females), each of who was set (not always same) 20 topics that was exhaustively divided from all of the 153 search topics of the NTCIR-4 Web collection. The total amount of collected speech data was about 178 minutes. The statistics of the speech data are shown in Table 3. Both OOV and WER are shown in table 4.

The manual transcriptions of the faithful parts were used as queries for the document retrieval system and the results were submitted to the formal evaluation in the NTCIR-4 Web task. Because the schedule of the evaluation result release in the NTCIR-4 Web task had been postponed, we could obtain the results of results of relevance judgment only for 35 search topics. For the 35 topics, we investigated the retrieval accuracy for the following input types.

Written-TITLE written three keywords, for which the title field tagged with <TITLE> was used,

Written-DESC written search topics, for which the description field tagged with <DESC> was used,

Written-DESC&NARR written search topics, for which the description and narrative fields tagged with <DESC> and <NARR> were used,

Spontaneous-MTS-F spontaneous speech corresponding to the faithful part transcribed manually,

Spontaneous-MTS-F&A spontaneous speech corresponding to the faithful and additional part transcribed manually,

Spontaneous-ASR-F spontaneous speech corresponding to the faithful part transcribed by speech recognition,

Spontaneous-ASR-F&A spontaneous speech corresponding to the faithful and additional part transcribed by speech recognition,

In the NTCIR-4 Web task, the relevance of each document with a search topic is classified in four grades, which are highly relevant, fairly relevant, partially relevant and irrelevant. We made two types of relevance judgment; rigid judgment (referred to as *rigid*), in which documents classified as highly relevant or fairly relevant are judged relevant, and relaxed judgment (referred to as *relaxed*), in which documents classified as partially relevant are also judged relevant.

Table 3. Statistics of spoken queries for all topics using NTCIR-4 Web collection.

Subject ID	Faithful part (sec.)			Additional part (sec.)		
	MIN	MAX	MEAN	MIN	MAX	MEAN
Female#3	9.8	38.2	19.2	9.4	37.7	21.0
Female#4	6.0	25.7	14.1	18.8	182.0	104.7
Female#5	10.8	105.5	60.2	13.3	60.1	35.5
Female#6	6.5	62.0	21.7	7.3	72.2	43.7
Male#3	5.7	16.9	8.7	12.3	40.6	26.5
Male#4	12.9	33.1	20.2	17.4	43.2	30.5
Male#5	11.8	38.2	26.6	31.1	95.9	55.8
Male#6	4.3	11.3	8.1	23.2	51.4	35.9

Table 4. OOV and WER of spoken queries for the NTCIR-4 Web collection.

Subject ID	Faithful part		The other part	
	OOV	WER	OOV	WER
Female#3	1.9	38.4	1.7	40.7
Female#4	0.6	38.1	1.9	59.6
Female#5	1.1	83.4	1.0	76.7
Female#6	0.9	35.7	1.6	45.9
Male#3	1.3	51.5	0.7	69.1
Male#4	1.1	79.5	0.8	67.2
Male#5	1.5	86.2	1.0	85.7
Male#6	1.0	66.4	1.9	60.1
Average	1.2	59.9	1.3	63.1

The two document retrieval systems were used for the experiments. One of them was the same system used in the previous experiment (referred to as *BASE*). The other was the extended system so as using character bi-grams that are used for the indexes for document retrieval in addition to the word-based indexes (referred to as *CBG*).

The results are shown in Table 5.

With respect to the word-based document retrieval (*BASE*), the result by the spontaneous inputs (Spontaneous-MTS-F) was as good as that by the written inputs (Written-DESC). Both results were improved by enlarging the input queries (Written-DESC&NARR and Spontaneous-MTS-F&A). It indicated that one of the features of spontaneously spoken queries that enables users to submit long queries easily was advantageous for document retrieval.

With respect to the extended system (*CBG*), while the results by the written inputs were improved, the results by the spontaneous inputs were degraded. One of the reason why the additional use of character bi-grams decrease the precision of the search results by spontaneous inputs seemed that it was influenced for the worse by the difference between the written language, which was used in both target documents and the written inputs, and the spoken language, which was used in the spontaneously spoken queries.

4 Conclusion

Motivated to realize the speech-driven information retrieval systems that accept spontaneously spoken queries, a method was presented to collect such speech data derived from the pre-defined search topics that had been systematically constructed for IR research. Because by definition it was impossible to prepare scripts for spontaneous speech in advance, we made subjects understand the meaning of search topics and made them freely speak their own expression about the topics. In order to evaluate both our method and the performance of the document retrieval by using the spontaneously spoken queries, we took place two experiments of collecting the speech data by our method using publicly available test collections of evaluating document retrieval. The first preliminary experiment took place with relatively small number of search topics selected from the NTCIR-3 Web retrieval collection, in order to test our method. The second experiment took place with all of the search topics released from the NTCIR-4 Web task to participate the formal run of the evaluation. The information about the collected data and the result of the evaluation with respect to both the speech recognition accuracy and the precision of document retrieval by using the collected data were presented. The results indicated that one of the features of spontaneously spoken queries that enables

Table 5. Mean Average Precision (MAP) for the NTCIR-4 Web collection.

	BASE		CBG	
	rigid	relaxed	rigid	relaxed
Written-TITLE	0.1255	0.1635	0.1542	0.1859
Written-DESC	0.1011	0.1351	0.1195	0.1419
Written-DESC&NARR	0.1347	0.1682	0.1345	0.1551
Spontaneous-MTS-F	0.1037	0.1210	0.0731	0.0873
Spontaneous-MTS-F&A	0.1184	0.1478	0.0742	0.0934
Spontaneous-ASR-F	0.0464	0.0643	0.0470	0.0576
Spontaneous-ASR-F&A	0.0566	0.0777	0.0480	0.0670

users to submit long queries easily was advantageous for document retrieval. We are also going to use the method in this paper to collect the spoken queries submitted to speech-driven question answering systems [2, 1].

5 Acknowledgements

This work was partly supported by Grant-in-Aid for Scientific Research (KAKENHI) (A) 14208033 from Japan Society for the Promotion of Science.

References

- [1] T. Akiba, K. Itou, and A. Fujii. Adapting language models for frequent fixed phrases by emphasizing n-gram subsets. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1469–1472, Geneva, Switzerland, Sept. 2003.
- [2] T. Akiba, K. Itou, A. Fujii, and T. Ishikawa. Selective back-off smoothing for incorporating grammatical constraints into the n-gram language model. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 881–884, Denver, Colorado, Sept. 2002.
- [3] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. Experiments in spoken queries for document retrieval. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1323–1326, Rhodes, Greece, Sept. 1997.
- [4] A. Fujii and K. Itou. Building a test collection for speech-driven web retrieval. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1153–1156, Geneva, Switzerland, Sept. 2003.
- [5] A. Fujii, K. Itou, and T. Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In A. R. Coden, E. W. Brown, and S. Srinivasan, editors, *Information Retrieval Techniques for Speech Applications (LNCS 2273)*, pages 94–104. Springer, 2002.
- [6] J. S. Garofolo, E. M. Voorhees, V. M. Stanford, and K. S. Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text Retrieval Conference*, pages 83–91, Gaithersburg, Maryland, Nov. 1997.
- [7] A. Lee and K. S. Tatsuya Kawahara. Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1691–1694, Aalborg, Denmark, Sept. 2001.
- [8] R. S. Taylor. The process of asking questions. *American Documentation*, 13(4):391–396, 1962.