Evaluating Speech-Driven Web Retrieval in the Third NTCIR Workshop

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

Abstract

Speech recognition has of late become a practical technology for real world applications. For the purpose of research and development in speech-driven retrieval, which facilitates retrieving information with spoken queries, we organized the speech-driven retrieval subtask in the NTCIR-3 Web retrieval task. Search topics for the Web retrieval main task were dictated by ten speakers and were recorded as collections of spoken queries. We used those queries to evaluate the performance of our speech-driven retrieval system, in which speech recognition and text retrieval modules were integrated. The text retrieval module, which is based on a probabilistic model, indexed only textual contents in documents (Web pages), but did not use HTML tags and hyperlink information in documents. Experimental results showed that a) the use of target documents for language modeling and b) enhancement of the vocabulary size in speech recognition were effective to improve the system performance.

Introduction

Automatic speech recognition, which decodes human voice to generate transcriptions, has of late become a practical technology. It is feasible that speech recognition is used in real world computer-based applications, specifically, those associated with human language. In fact, a number of speech-based methods have been explored in the information retrieval (IR) community, which can be classified into the following two fundamental categories:

- spoken document retrieval, in which written queries are used to search speech (e.g., broadcast news audio) archives for relevant speech information (Johnson *et al.* 1999; Jones *et al.* 1996; Sheridan, Wechsler, & Schäuble 1997; Singhal & Pereira 1999; Srinivasan & Petkovic 2000; Wechsler, Munteanu, & Schäuble 1998; Whittaker *et al.* 1999),
- speech-driven retrieval, in which spoken queries are used to retrieve relevant textual information (Barnett *et al.* 1997; Crestani 2000; Fujii, Itou, & Ishikawa 2002a; 2002b; Itou, Fujii, & Ishikawa 2001; Kupiec, Kimber, & Balasubramanian 1994).

Initiated partially by the TREC-6 spoken document retrieval (SDR) track (Garofolo *et al.* 1997), various methods have been proposed for spoken document retrieval. However, a relatively small number of methods have been explored for speech-driven text retrieval, although they are associated with numerous keyboard-less retrieval applications, such as telephone-based retrieval, car navigation systems, and user-friendly interfaces.

Barnett et al. (1997) performed comparative experiments related to speech-driven retrieval, in which the DRAGON speech recognition system was used as an input interface for the INQUERY text retrieval system. They used as test inputs 35 queries collected from the TREC topics and dictated by a single male speaker. Crestani (2000) also used the above 35 queries and showed that conventional relevance feedback techniques marginally improved the accuracy of speech-driven text retrieval.

These above cases focused solely on improving text retrieval methods and did not address problems in improving speech recognition accuracy. In fact, an existing speech recognition system was used with no enhancement. In other words, speech recognition and text retrieval modules were fundamentally independent and were simply connected by means of an input/output protocol.

However, since most speech recognition systems are trained based on specific domains, the accuracy of speech recognition across domains is not satisfactory. As can easily be predicted, in cases of Barnett et al. (1997) and Crestani (2000), the speech recognition error rate was relatively high and decreased the retrieval accuracy. Additionally, speech recognition with a high accuracy is important for interactive retrieval, such as dialog-based retrieval.

Kupiec et al. (1994) proposed a method based on *word* recognition, which accepts only a small number of keywords, derives multiple transcription hypotheses (i.e., possible word combinations), and uses a target collection to determine the most plausible word combination. In other words, word combinations that frequently appear in the target collection can be recognized with a high accuracy. However, for longer queries, such as phrases and sentences, the number of hypotheses increases, and thus the searching cost is prohibitive. Thus, their method cannot easily be used for *continuous* speech recognition methods.

Motivated by these problems, we integrated continuous speech recognition and text retrieval to improve both recognition and retrieval accuracy in speech-driven text retrieval (Fujii, Itou, & Ishikawa 2002a; 2002b; Itou, Fujii, & Ishikawa 2001). In brief, our method used target documents to adapt language models and to recognize out-of-vocabulary words for speech recognition. However, a number of issues still remain open questions before speech-driven retrieval can be used as a practical (real-world) application. For example, extensive experiments using large test collections have not been performed for speech-driven retrieval. This stimulated us to further explore this exciting research area.

In the NTCIR-3 Web retrieval task[1], the *main* task was organized to promote conventional text-based retrieval (Eguchi *et al.* 2002). Additionally, *optional* subtasks were also invited, in which a group of researchers voluntarily organized a subtask to promote their common research area. To make use of this opportunity, we organized the "speech-driven retrieval" subtask, and produced a reusable test collection for experiments of Web retrieval driven by spoken queries. Since we also participated in the main task, we performed comparative experiments to evaluate the performance of text-based and speech-driven retrieval systems.

Test Collection for Speech-Driven Retrieval

Overview

The purpose of the speech-driven retrieval subtask was to produce reusable test collections and tools available to the public, so that researchers in the information retrieval and speech processing communities can develop technologies and share the scientific knowledge inherent in speech-driven information retrieval.

In principle, as with conventional IR test collections, test collections for speech-driven retrieval are required to include test queries, target documents, and relevance assessment for each query. However, unlike conventional text-based IR, queries are speech data uttered by human speakers.

In practice, since producing the entire collection is prohibitive, we produced speech data related to the Web retrieval main task. Therefore, target documents and relevance assessment in the main task can be used for the purpose of speech-driven retrieval. It should be noted that in the main task, retrieval results driven by spoken queries were not used for pooling, which is a method to reduce the number of relevant document candidates by using retrieval results of multiple IR systems (Voorhees 1998).

However, participants for the NTCIR workshop are mainly researchers in the information retrieval and natural language processing communities, and are not necessarily experts in developing and operating speech recognition systems. Thus, we also produced dictionaries and language models that can be used with an existing speech recognition engine (decoder), which helps researchers to perform similar experiments described in this paper.

All above data are included in the NTCIR-3 Web retrieval test collection, which is available to the public.

Spoken Queries

For the NTCIR-3 Web retrieval main task, 105 search topics were manually produced, for each of which relevance assessment was manually performed with respect to two different document sets, i.e., 10GB and 100GB collections. The 10GB and 100GB collections translate approximately to 1M and 10M documents, respectively.

Each 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 (`<CONC>`), sample of relevant documents (`<RDOC>`), and brief profile of the user who produced the topic (`<USER>`).

Figure 1 depicts a translation of an example topic. Although Japanese topics were used in the main task, English translations are also included in the Web retrieval collection mainly for publication purposes.

```
<TOPIC>
<NUM>0010</NUM>
<TITLE CASE="b">Aurora, conditions, ob-
servation</TITLE>
<DESC>For observation purposes, I want
to know the conditions that give rise to
an aurora</DESC>
<NARR><BACK>I want to observe an aurora
so I want to know the conditions neces-
sary for its occurrence and the mecha-
nism behind it.</BACK><RELE>Aurora ob-
servation records, etc.  list the place
and time so only documents that pro-
vide additional information such as the
weather and temperature at the time of
occurrence are relevant.  </RELE></NARR>
<CONC>Aurora, occurrence, conditions,
observation, mechanism</CONC>
<RDOC>NW003201843, NW001129327,
NW002699585</RDOC>
<USER>1st year Master's student, female,
2.5 years search experience</USER>
</TOPIC>
```

Figure 1: An example topic in the Web retrieval collection.

Participants for the main task were allowed to submit more than one retrieval result using one or more fields. However, participants were required to submit results obtained with the title and description fields independently. Titles are a list of keywords, and descriptions are phrases and sentences.

From the viewpoint of speech recognition, titles and descriptions can be used to evaluate *word* and *continuous* recognition methods, respectively. Since the state-of-the-art speech recognition is based on a continuous recognition framework, we used only the description field. For the first speech-driven retrieval subtask, we focused on *dictated* (*read*) speech, although our ultimate goal is to recognize *spontaneous* speech. We asked ten speakers (five adult males/females) to dictate descriptions in the 105 topics.

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

The ten speakers also dictated 50 sentences in the ATR phonetic-balanced sentence set as reference data, which can potentially be used for speaker adaptation (however, we did not use this additional data for the purpose of experiments described in this paper).

These above spoken queries and sentences were recorded with the same close-talk microphone in a noiseless office. Speech waves were digitized at a 16KHz sampling frequency and were quantized at 16 bits. The resultant data are in the RIFF format.

Language Models

Unlike general-purpose speech recognition, in speech-driven text retrieval, users usually speak contents associated with a target collection, from which documents relevant to user needs are retrieved.

In a stochastic speech recognition framework, the accuracy depends primarily on acoustic and language models (Bahl, Jelinek, & Mercer 1983). While acoustic models are related to phonetic properties, language models, which represent linguistic contents to be spoken, are related to target collections. Thus, it is intuitively feasible that language models have to be produced based on target collections. To sum up, our belief is that by adapting a language model to a target IR collection, we can improve the speech recognition accuracy and consequently the retrieval accuracy.

Motivated by this background, we used target documents for the main task to produce language models. For this purpose, we used only the 100GB collection, because the 10GB collection is a subset of the 100GB collection.

State-of-the-art speech recognition systems still have to limit the vocabulary size (i.e., the number of words in a dictionary), due to problems in estimating statistical language models (Young 1996) and constraints associated with hardware, such as memory. In addition, computation time is crucial for a real-time usage, including speech-driven retrieval. Consequently, for many languages the vocabulary size is limited to a couple of ten thousands (Itou *et al.* 1999; Paul & Baker 1992; Steeneken & van Leeuwen 1995).

We produced two language models of different vocabulary sizes, for which 20,000 and 60,000 high frequency words were independently used to produce word-based trigram models, so that researchers can investigate the relation between the vocabulary size and system performance. We shall call these models "Web20K" and "Web60K", respectively. We used the ChaSen morphological analyzer[2] to extract words from the 100GB collection.

To resolve the data sparseness problem, we used a back-off smoothing method, in which the Witten-Bell discounting method was used to compute back-off coefficients. In addition, through preliminary experiments, cut-off thresholds were empirically set 20 and 10 for the Web20K and Web60K models, respectively. Trigrams whose frequency was above the threshold were used for language modeling. Language models and dictionaries are in the ARPA and HTK formats, respectively.

Table 1 shows statistics related to word tokens/types in the 100GB collection and ten years of "Mainichi Shimbun" newspaper articles in 1991–2000. We shall use the term "word token" to refer to occurrences of words, and the term "word type" to refer to vocabulary items. Roughly, the size of the 100G collection ("Web") is ten times that of ten years of newspaper articles ("News"), which was one of the largest Japanese corpora available for the purpose of research and development in language modeling. In other words, the Web is a vital, as yet untapped, corpus for language modeling.

Table 1: The number of words in source corpora for language modeling.

	Web (100GB)	News (10 years)
# of Word types	2.57M	0.32M
# of Word tokens	2.44G	0.26G

System Description

Overview

Figure 2 depicts the overall design of our speech-driven text retrieval system, which consists of speech recognition and text retrieval modules.

In the off-line process, a target IR collection is used to produce a language model, so that user speech related to the collection can be recognized with a high accuracy. However, an acoustic model was produced independent of the target collection.

In the on-line process, given an information need spoken by a user (i.e., a spoken query), the speech recognition module uses acoustic and language models to generate a transcription of the user speech. Then, the text retrieval module searches the target IR collection for documents relevant to the transcription, and outputs a specific number of top-ranked documents according to the degree of relevance in descending order. In the following two sections, we explain speech recognition and text retrieval modules, respectively.

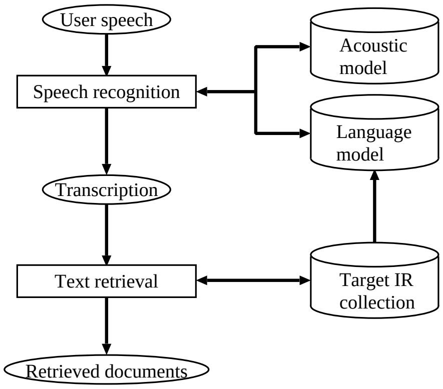

Figure 2: The overview of our speech-driven text retrieval system.

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

Speech Recognition

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

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

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

We used the Japanese dictation toolkit (Kawahara *et al.* 2000)[3], which includes the Julius decoder and acoustic/language models. Julius performs a two-pass (forward-backward) search using word-based forward bigrams and backward trigrams.

The acoustic model was produced from the ASJ speech database (Itou *et al.* 1998), which contains approximately 20,000 sentences uttered by 132 speakers including the both gender groups. A 16-mixture Gaussian distribution triphone Hidden Markov Model, in which states are clustered into 2,000 groups by a state-tying method, is used. The language model is a word-based trigram model produced from 60,000 high frequency words in ten years of Mainichi Shimbun newspaper articles.

This toolkit also includes development softwares so that acoustic and language models can be produced and replaced depending on the application. While we used the acoustic model provided in the toolkit, we used new language models produced from the 100GB collections, that is, the Web20K and Web60K models.

Text Retrieval

The retrieval module is based on an existing retrieval method (Robertson & Walker 1994), which computes the relevance score between the transcribed query and each document in the collection. The relevance score for document d is computed by Equation (2).

$$\sum_{t} f_{t,q} \cdot \frac{(K+1) \cdot f_{t,d}}{K \cdot \{(1-b) + \frac{dl_d}{b \cdot avgdl}\} + f_{t,d}} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5} \qquad (2)$$

Here, $f_{t,q}$ and $f_{t,d}$ denote the frequency that term t appears in query q and document d, respectively. N and n_t denote the total number of documents in the collection and the number of documents containing term t, respectively. dl_d denotes the length of document d, and $avgdl$ denotes the average length of documents in the collection. We empirically set $K = 2.0$ and $b = 0.8$, respectively.

Given transcriptions (i.e., speech recognition results for spoken queries), the retrieval module searches a target IR collection for relevant documents and sorts them according to the score in descending order.

We used content words, such as nouns, extracted from documents as index terms, and performed word-based indexing. We used the ChaSen morphological analyzer to extract content words. We also extracted terms from transcribed queries using the same method. We used words and bi-words (i.e., word-based bigrams) as index terms.

We used the same retrieval module to participate in other text retrieval workshops, such as NTCIR-2. However, the 10GB/100GB Web collections were different from existing Japanese test collections in the following two perspectives.

First, the Web collections are much larger than existing test collections. For example, the file size of the NTCIR-2 Japanese collection including 736,166 technical abstracts is approximately 900MB (NII 2001). Thus, tricks were needed to index larger document collections. Specifically, files of more than 2GB size were problematic for file systems and tools in existing operating systems.

To resolve this problem, we divided the 100GB collection into 20 smaller sub-collections so that each file size did not exceed 2GB, and indexed the 20 files independently. Given queries, we retrieved documents using the 20 indexes and sorted documents according to the relevance score. The relevance score of a document was computed with respect to the sub-collection from which the document was retrieved.

Second, target documents are Web pages, in which HTML (Hyper Text Markup Language) tags provide the textual information with a certain structure. However, the use of HTML tags are usually different depending on the author. Thus, we discarded HTML tags in documents, and indexed only textual contents. Additionally, we did not use hyperlink information among Web pages for retrieval purposes.

Experimentation

Evaluating Text-to-Text Retrieval

In the Web retrieval main task, different types of text retrieval were performed. The first type was “Topic Retrieval” resembling the TREC ad hoc retrieval. The second type was “Similarity Retrieval,” in which documents were used as queries instead of keywords and phrases. The third type was “Target Retrieval,” in which systems with a high precision were highly valued. This feature provided a salient contrast to the first two retrieval types, in which both recall and precision were equally used as evaluation measures.

Although the produced spoken queries can be used for the first and third task types, we focused solely on the Topic Retrieval for the sake of simplicity. In addition, our previous experiments (Fujii, Itou, & Ishikawa 2002a; 2002b; Itou, Fujii, & Ishikawa 2001), in which the IREX[4] and NTCIR[5] collections were used, were also a type of Target Retrieval. We used the 47 topics for the Topic Retrieval task to retrieve 1,000 top documents, and used the TREC evaluation software to calculate mean average precision (MAP) values (i.e., non-interpolated average precision values, averaged over the 47 topics).

Relevance assessment was performed based on four ranks of relevance, that is, highly relevant, relevant, partially relevant and irrelevant. In addition, unlike conventional retrieval tasks, documents hyperlinked from retrieved documents were optionally used for relevance assessment. To

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

[4] http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html

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

sum up, the following four assessment types were available to calculate the MAP values:

- (highly) relevant documents were regarded as correct answers, and hyperlink information was NOT used (RC),
- (highly) relevant documents were regarded as correct answers, and hyperlink information was used (RL),
- partially relevant documents were also regarded as correct answers, and hyperlink information was NOT used (PC),
- partially relevant documents were also regarded as correct answers, and hyperlink information was used (PL).

In the formal run for the main task, we submitted results obtained with different methods for the 10GB and 100GB collections, respectively. First, we used title (<TITLE>) and description (<DESC>) fields independently as queries. Second, we used as index terms either only words or a combination of words and bi-words. As a result, we investigated the MAP values for 32 cases as shown in Table 2.

By looking at Table 2, there was no significant difference among the four methods in performance. However, by comparing two indexing methods, the use of both words and bi-words generally improved the MAP values of that obtained with only words, irrespective of the collection size, topic field used, and relevance assessment type.

Evaluating Speech-Driven Retrieval

The purpose of experiments for speech-driven retrieval was two-fold. First, we investigated the extent to which a language model produced based on a target document collection contributes to improve the performance. Second, we investigated the impact of the vocabulary size for speech recognition to speech-driven retrieval. Thus, we compared the performance of the following four retrieval methods:

- text-to-text retrieval, which used written queries, and can be seen as the perfect speech-driven text retrieval (“Text”),
- speech-driven text retrieval, in which the Web60K model was used (“Web60K”),
- speech-driven text retrieval, in which a language model produced from 60,000 high frequency words in ten years of Mainichi Shimbun newspaper articles was used (“News60K”),
- speech-driven text retrieval, in which the Web20K model was used (“Web20K”).

For text-to-text retrieval, we used descriptions (<DESC>) as queries, because the spoken queries used for speech-driven retrieval methods were descriptions dictated by speakers. In addition, we used both bi-words and words for indexing, because the experimental results in Table 2 showed that the use of bi-words for indexing improved the performance of that obtained with only words.

For speech-driven text retrieval methods, queries dictated by the ten speakers were used independently, and the final result was obtained by averaging results for all speakers. Although the Julius decoder used in the speech recognition module generated more than one transcription candidate (hypothesis) for a single speech, we used only the one with the greatest probability score.

All language models were produced by means of the same softwares, but were different in terms of the vocabulary size and source documents.

Table 3 shows the MAP values with respect to the four relevance assessment types and the word error rate in speech recognition, for different retrieval methods targeting the 10GB and 100GB collections.

As with existing experiments for speech recognition, word error rate (WER) is the ratio between the number of word errors (i.e., deletion, insertion, and substitution) and the total number of words. In addition, we investigated error rate with respect to query terms (i.e., keywords used for retrieval), which we shall call term error rate (TER). It should be noted that unlike MAP, smaller values of WER and TER are obtained with better methods.

Table 3 also shows test-set out-of-vocabulary rate (OOV), which is the ratio between the number of words not included in the speech recognition dictionary and the total number of words in spoken queries. In addition, the column of “Time” denotes CPU time (sec.) required for speech recognition per query, for which we used a PC with two CPUs (AMD Athlon MP 1900+) and a memory size of 3GB.

Suggestions which can be derived from the results in Table 3 are as follows.

Looking at columns of WER and TER, News60K and Web20K were comparable in the speech recognition performance, but Web60K outperformed both cases. However, difference of News60K and Web20K in OOV did not affect WER and TER. In addition, TER was greater than WER, because in computing TER, functional words, which are generally recognized with a high accuracy, were excluded.

While the MAP values of News60K and Web20K were also comparable, the MAP values of Web60K, which were roughly 60-70% of those obtained with Text, were greater than those for News60K and Web20K, irrespective of the relevance assessment type. These results were observable for both the 10GB and 100GB collections.

The only difference between News60K and Web60K was the source corpus for language modeling in speech recognition, and therefore we can conclude that the use of target collections to produce a language model was effective for speech-driven retrieval. In addition, by comparing the MAP values of Web20K and Web60K, we can conclude that the vocabulary size for speech recognition was also influential for the performance of speech-driven retrieval.

CPU time for speech recognition did not significantly differ depending on the language model, despite the fact that the number of words and N-gram tuples in Web60K was larger than those in News60K and Web20K. In other words, Web60K did not decrease the time efficiency of News60K and Web20K, which is crucial for read-world usage. At the same time, response time also depends on various factors, such as the hardware and decoder program used, we do not pretend to draw any premature conclusions regarding the time efficiency.

We analyzed speech recognition errors, focusing mainly on those attributed to the out-of-vocabulary problem. Ta-

Table 2: MAP values for different text-to-text retrieval methods targeting the 10GB and 100GB collections.

		MAP (10GB)				MAP (100GB)			
Field	Index	RC	RL	PC	PL	RC	RL	PC	PL
`<DESC>`	word & bi-word	.1470	.1286	.1612	.1476	.0855	.0982	.1257	.1274
`<DESC>`	word	.1389	.1187	.1563	.1374	.0843	.0928	.1184	.1201
`<TITLE>`	word & bi-word	.1493	.1227	.1523	.1407	.0815	.0981	.1346	.1358
`<TITLE>`	word	.1402	.1150	.1437	.1323	.0808	.0938	.1280	.1299

Table 3: Experimental results for different retrieval methods targeting the 10GB and 100GB collections (OOV: test-set out-of-vocabulary rate, WER: word error rate, TER: term error rate, MAP: mean average precision).

					MAP (10GB)				MAP (100GB)			
Method	OOV	WER	TER	Time (sec.)	RC	RL	PC	PL	RC	RL	PC	PL
Text	—	—	—	—	.1470	.1286	.1612	.1476	.0855	.0982	.1257	.1274
Web60K	.0073	.1311	.2162	7.2	.0966	.0916	.0973	.1013	.0542	.0628	.0766	.0809
News60K	.0157	.1806	.2991	7.0	.0701	.0681	.0790	.0779	.0341	.0404	.0503	.0535
Web20K	.0423	.1642	.2757	6.7	.0616	.0628	.0571	.0653	.0315	.0378	.0456	.0485

ble 4 shows the ratio of the number of out-of-vocabulary words to the total number of misrecognized words (or terms) in transcriptions. However, it should be noted that the actual ratio of errors due to the OOV problem can potentially be higher than those figures, because non-OOV words collocating with OOV words are often misrecognized. Remaining reasons of speech recognition errors are associated with insufficient N-gram statistics and the acoustic model.

Table 4: The ratio of the number of OOV words/terms to the total number of misrecognized words/terms.

	Word	Term
Web60K	.0704	.1838
News60K	.0966	.2143
Web20K	.2855	.5049

As can be predicted, the ratio of OOV words (terms) in Web20K was much higher than those in Web60K and News60K. However, by comparing News60K and Web20K, WER and TER of News60K in Table 3 were higher than those of Web20K. This suggests that insufficient N-gram statistics were more problematic in News60K, when compared with Web20K.

Although we used only the top-ranked transcription hypotheses as queries, certain words can potentially be correctly transcribed in lower-ranked hypotheses. Thus, to investigate the effect of multiple hypotheses, we varied the number of hypotheses used as queries and evaluated its effect on the MAP value. Table 5 shows the result, in which we used the Web60K model for speech recognition and targeted the 100G collection. In the case of $H = 1$, the MAP values are the same as those in Table 3. According to this table, the MAP values marginally decreased when we increased the number of hypotheses used as queries, irrespective of the relevance assessment type.

Table 5: MAP values of the Web60K speech-driven retrieval method with different numbers of hypotheses (H), targeting the 100G collection.

	RC	RL	PC	PL
$H = 1$	.0542	.0628	.0766	.0809
$H = 3$	.0527	.0608	.0755	.0794
$H = 5$	.0529	.0609	.0754	.0794

Conclusion

In the NTCIR-3 Web retrieval task, we organized the speech-driven retrieval subtask and produced 105 spoken queries dictated by ten speakers. We also produced word-based trigram language models using approximately 10M documents in the 100GB collection used for the main task. We used those queries and language models to evaluate the performance of our speech-driven retrieval system. Experimental results showed that a) the use of target documents for language modeling and b) enhancement of the vocabulary size in speech recognition were effective to improve the system performance. As with the collection for the main task, all speech data and language models produced for this subtask are available to the public. Future work will include experiments using spontaneous spoken queries.

Acknowledgments

The authors would like to thank the organizers of the NTCIR-3 Web retrieval task for their support to the speech-driven retrieval subtask.

References

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

Barnett, J.; Anderson, S.; Broglio, J.; Singh, M.; Hudson, R.; and Kuo, S. W. 1997. Experiments in spoken queries for document retrieval. In *Proceedings of Eurospeech97*, 1323–1326.

Crestani, F. 2000. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems*, 267–281.

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

Fujii, A.; Itou, K.; and Ishikawa, T. 2002a. A method for open-vocabulary speech-driven text retrieval. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 188–195.

Fujii, A.; Itou, K.; and Ishikawa, T. 2002b. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In Coden, A. R.; Brown, E. W.; and Srinivasan, S., eds., *Information Retrieval Techniques for Speech Applications (LNCS 2273)*. Springer. 94–104.

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

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

Itou, K.; Yamamoto, M.; Takeda, K.; Takezawa, T.; Matsuoka, T.; Kobayashi, T.; and Shikano, K. 1999. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of Acoustic Society of Japan* 20(3):199–206.

Itou, K.; Fujii, A.; and Ishikawa, T. 2001. Language modeling for multi-domain speech-driven text retrieval. In *IEEE Automatic Speech Recognition and Understanding Workshop*.

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

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

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

Kupiec, J.; Kimber, D.; and Balasubramanian, V. 1994. Speech-based retrieval using semantic co-occurrence filtering. In *Proceedings of the ARPA Human Language Technology Workshop*, 373–377.

National Institute of Informatics. 2001. *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*.

Paul, D. B., and Baker, J. M. 1992. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of DARPA Speech & Natural Language Workshop*, 357–362.

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

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

Singhal, A., and Pereira, F. 1999. Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 34–41.

Srinivasan, S., and Petkovic, D. 2000. Phonetic confusion matrix based spoken document retrieval. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 81–87.

Steeneken, H. J. M., and van Leeuwen, D. A. 1995. Multilingual assessment of speaker independent large vocabulary speech-recognition systems: The SQALE-project. In *Proceedings of Eurospeech95*, 1271–1274.

Voorhees, E. M. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 315–323.

Wechsler, M.; Munteanu, E.; and Schäuble, P. 1998. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 20–27.

Whittaker, S.; Hirschberg, J.; Choi, J.; Hindle, D.; Pereira, F.; and Singhal, A. 1999. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 26–33.

Young, S. 1996. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine* 45–57.