Evaluating Speech-Driven IR in the NTCIR-3 Web Retrieval Task

Atsushi Fujii[†,†††] and Katunobu Itou[††,†††]
[†] University of Library and Information Science
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@ulis.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 recorded as collections of spoken queries. We used those queries to evaluate the performance of our speech-driven retrieval system, where 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.

Keywords: *Web retrieval, speech recognition, language modeling, spoken queries, test collections*

1 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 [10, 11, 18, 19, 20, 22, 23],
- speech-driven retrieval, in which spoken queries are used to retrieve relevant textual information [2, 3, 4, 5, 7, 14].

Initiated partially by the TREC-6 spoken document retrieval (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, 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. [2] performed comparative experiments related to speech-driven retrieval, where 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 [3] also used the above 35 queries and showed that conventional relevance feedback techniques marginally improved the accuracy for 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 way 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. Thus, as can easily be predicted, in cases of Barnett et al. [2] and Crestani [3], a speech recognition error rate was relatively high and considerably decreased the retrieval accuracy.

Kupiec [14] proposed a method based on *word* recognition, which accepts only a small number of keywords, derives multiple transcription candidates (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 target collections can be recognized with a high accuracy. However, in the case of longer queries, such as phrases and sentences, the number of candidates increases, and thus the searching cost is prohibitive. In other words, 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 [4, 5, 7]. 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, which stimulates us to further explore this exciting research area.

In the NTCIR-3 Web retrieval task, the *main* task was organized to promote conventional text-based retrieval. Additionally, *optional* subtasks were also invited, where 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.

Section 2 describes the test collection produced for the speech-driven retrieval subtask. Section 3 describes our speech-driven retrieval system, and Section 4 elaborates on comparative experiments, in which we evaluated our system in terms of the speech recognition and retrieval accuracy.

2 Test Collection for Speech-Driven Retrieval Subtask

2.1 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 knowledge related to speech-driven information retrieval.

In principle, as with conventional IR test collections, test collections for speech-driven retrieval must include test queries, target documents, relevance assessment for each query. However, unlike the case of conventional text-based IR, queries must be speech data uttered by human speakers.

In practice, since producing the entire collection is prohibitive, we produced speech data which can be used as queries for the Web retrieval main task. Therefore, target documents and relevance assessment in the main task can also be used for the purpose of speech-driven retrieval. It should be noted that in the main task no retrieval results driven by spoken queries were not used for pooling, which reduces the number of candidates of relevant documents.

However, participants for the NTCIR workshop are mainly researchers in the information retrieval and natural language processing communities, and thus are not necessarily familiar with developing and operating speech recognition systems. In view of this problem, we also produced dictionaries and language models that can be used for an existing speech recognition engine (decoder), which helps researchers to perform similar experiments described in this paper (see Section 4 for details).

All of the above data will be available to the public in the NTCIR-3 Web retrieval test collection.

2.2 Spoken Queries

For the NTCIR-3 Web retrieval main task, 105 search topics (queries) were manually produced, for each of which relevance assessment was manually performed with respect to two different document sets, i.e., the 10GB and 100GB collections. The 10GB and 100GB collections include approximately 1,000,000 and 10,000,000 documents, respectively.

Each topic, which is organized in the SGML format, 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>).

Participants for the main task were allowed to submit more than one retrieval result using different fields. However, results obtained with the title and description fields independently must be submitted. 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* (or *read*) speech, although one of our ultimate goals is to recognize *spontaneous* speech. We asked ten speakers (five adult males/females) to dictate descriptions in the 105 topics.

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 were saved in the RIFF format.

2.3 Language Models

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

In a stochastic speech recognition framework, the accuracy depends primarily on acoustic and language models [1]. 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 based on a target IR collection, we can improve the speech recognition accuracy. Consequently, the retrieval accuracy can also be improved.

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 [24] 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 [8, 16, 21].

We produced two language models of different vocabulary sizes, for which 20,000 and 60,000 high-frequent 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[1] to extract words from the 100GB collection.

To resolve the data sparseness problem, we used a back-off smoothing method, where 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. Roughly, the 100G collection ("Web") is ten times the size of ten years of newspaper articles ("News"), which is (or "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

3 System Description

3.1 Overview

Figure 1 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 is 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 a 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.

3.2 Speech Recognition

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

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

Here, $P(X|W)$ models a probability that word sequence W is transformed into phone sequence X, and

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

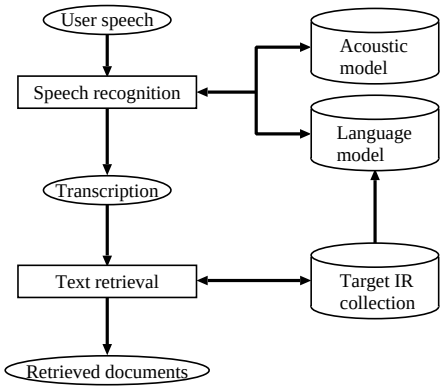

Figure 1. The design of our speech-driven text retrieval system.

$P(W)$ models a probability that W is linguistically acceptable. These factors are usually called acoustic and language models, respectively.

For the speech recognition module, we used the Japanese dictation toolkit [13][2], which includes the "Julius" recognition engine and acoustic/language models. Julius uses word-based forward bigrams and backward trigrams to perform a two-pass (forward-backward) search. Julius also uses a 16-mixture Gaussian distribution triphone Hidden Markov Model, where states are clustered into 2,000 groups by a state-tying method.

The acoustic model was produced by way of the ASJ speech database (ASJ-JNAS) [8, 9], which contains approximately 20,000 sentences uttered by 132 speakers including the both gender groups. The language model is a word-based trigram model produced from 60,000 high-frequent 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 (see Section 2.3 for details).

3.3 Text Retrieval

The retrieval module is based on an existing probabilistic retrieval method [17], which computes the relevance score between the translated query and each document in the collection. The relevance score for document d is computed based on Equation (2).

$$\sum_t \left(\frac{TF_{t,d}}{\frac{DL_d}{avglen} + TF_{t,d}} \cdot \log \frac{N}{DF_t} \right) \qquad (2)$$

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

Here, $TF_{t,d}$ denotes the frequency that term t appears in document d. DF_t and N denote the number of documents containing term t and the total number of documents in the collection. DL_d denotes the length of document d (i.e., the number of characters contained in d), and $avglen$ denotes the average length of documents in the collection.

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 extracted from documents as index terms, and performed a word-based indexing. For this purpose, 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. 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, where HTML (Hyper Text Markup Language) tags provide the textual information with a certain structure. However, the use of HTML tags are usually different depending the author. Thus, we discarded HTML tags in documents, and indexed only textual contents. Additionally, for the purpose of retrieval, we did not use hyperlink information among Web pages.

4 Experimentation

4.1 Evaluating Text-to-Text Retrieval

In the Web retrieval main task, different types of text retrieval were organized. 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, where both recall and precision were used as evaluation measures.

Although spoken queries described in Section 2.2 can be used for any of the above three retrieval types, we focused solely on the Topic Retrieval for the sake of simplicity. In addition, our previous experiments [4, 5, 7], where the IREX[3] and NTCIR[4] collections were used, were also a type of Target Retrieval. Thus, we used 47 topics to retrieve 1,000 top documents, and used the TREC evaluation software to calculate non-interpolated average precision values.

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 sum up, the following four assessment types were available to calculate average precision 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, for each of the above four relevance assessment types, we investigated non-interpolated average precision values of four different methods, 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 performance of that obtained with only words, irrespective of the collection size, topic field used, and relevance assessment type.

[3]http://cs.nyu.edu/cs/projects/proteus/irex/index-e.html
[4]http://research.nii.ac.jp/ntcir/index-en.html

4.2 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 in 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-frequent words in ten years of "Mainichi Shimbun" newspaper articles was used ("News60K"),
- speech-driven text retrieval, in which the Web20K model was used ("Web20K").

In the case of text-to-text retrieval, we used descriptions (<DESC>) as queries, because 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 experimental results in Section 4.1 showed that the use of bi-words for indexing improved the performance of that obtained with only words (see Table 2 for details).

In cases of 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 the speakers. Although the Julius decoder used in the speech recognition module generated more than one transcription candidates for a single speech, we used only the one with the greatest probability score.

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

Table 3 shows the non-interpolated average precision values of each relevance assessment and 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 noted that unlike the case of average precision, smaller WER (TER) values are obtained with better methods.

Table 2. Non-interpolated average precision values for different text-to-text retrieval methods targeting the 10GB and 100GB collections.

		Avg. precision (10GB)				Avg. precision (100GB)			
Field	Indexing	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 methods targeting the 10GB and 100GB collections (OOV: test-set out-of-vocabulary rate, WER: word error rate, TER: term error rate).

					Avg. precision (10GB)				Avg. precision (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	.1625	.2647	7.2	.0839	.0811	.0864	.0912	.0474	.0552	.0676	.0717
News60K	.0157	.2081	.3472	7.0	.0625	.0610	.0715	.0705	.0309	.0369	.0450	.0484
Web20K	.0423	.1973	.3272	6.7	.0550	.0565	.0524	.0593	.0281	.0339	.0410	.0438

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 these results are as follows.

Looking at columns of WER and TER, News60K and Web20K were quite comparable in 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 the case of computing TER, functional words, which are generally recognized with a high accuracy, were excluded.

While average precision values of News60K and Web20K were also comparable, average precision values of Web60K, which were roughly 57% of that obtained with Text, were greater than those for News60K and Web20K, irrespective of the relevance assessment type. These results were observable in cases of 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 concluded that the use of target collections to produce a language model was effective for speech-driven retrieval. In addition, by comparing average precision values of Web20K and Web60K, we concluded that the vocabulary size for speech recognition was also influential for the performance of speech-driven retrieval.

Finally, 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 depends on various factors, such as a hardware used, we do not pretend to draw any premature conclusions regarding the time efficiency.

4.3 Experiments by the other participants

We also report experimental results obtained by the TUT (Toyohashi University of Technology) group[5], which used the same test collection described in this paper. Details of techniques employed here as well as individual LVCSR models are found in a paper by Matsushita et al. [15].

The TUT group evaluated eight LVCSR models against the recognition of the spoken queries and then combined outputs of the eight models so as to improve word recognition rates. They employed the decision list classifier (a kind of machine learning technique) to the task of choosing a sequence of the most confident words, where, as features, the list of the models which output the word, part-of-speech of the word, and the syllable length of the word were used. Individual eight models differed in their decoders as well as their acoustic models, while their language models were the same as the "Web20K" model (see Section 4.2). Out of the eight LVCSR models, four models

[5]The members of the TUT group were Mr. M. Matsushita, Mr. H. Nishizaki, Dr. T. Utsuro, and Prof. S. Nakagawa.

Table 4. Experimental results obtained by the TUT group.

Method	OOV	WER	TER	Avg. precision (100GB)			
				RC	RL	PC	PL
Text	—	—	—	.0855	.0982	.1257	.1274
Julius (triphone)	.0435	.2688	.3961	.0249	.0288	.0337	.0367
SPOJUS (syllable)	.0435	.2410	.4355	.0274	.0311	.0358	.0386
Comb-8-ML	.0435	.1501	.2950	.0279	.0322	.0380	.0409
Union-8-Correct	.0435	.0838	.1861	.0294	.0348	.0411	.0446

were with the decoder Julius [13], while the other four were with the decoder SPOJUS [12]. As the acoustic models for the decoder Julius, three phoneme-based HMM models (triphone, PTM, monophone), and a syllable-based HMM model were used. The four acoustic models of the decoder SPOJUS were all based on syllable HMMs but differed in feature parameters (segment-based or frame-based) and/or in conventional HMM with self loop transition or HMM with duration control. The TUT group used the gender-dependent (male) acoustic models, and thus used only queries spoken by the five male subjects. For the purpose of text retrieval, they used the same method as in Section 4.2, but targeted only the 100GB collection.

Table 4 shows the results[6], where "Julius(triphone)" and "SPOJUS(syllable)" give results with spoken queries recognized by two individual models. Among the results with the eight individual LVCSR models, the one we show as "SPOJUS(syllable)" was the best in terms of the retrieval performance. "Comb-8-ML" gives the performance with the results of combining outputs of the eight LVCSR models. "Union-8-Correct" gives the performance with the results of keeping correctly recognized words and filtering out recognition error words, from the union of the outputs of the eight LVCSR models. This performance can be regarded as an upper bound of the model combination approach. As can be seen from these results, model combination by the machine learning technique contributed to the improvement of the retrieval performance.

5 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. 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. We also showed the effectiveness of model combination methods in speech recognition through experiments performed by the TUT group. All of the speech data and language models produced for this subtask will be available to the public in the NTCIR-3 Web retrieval test collection.

[6] Even with the same decoder, acoustic model, and language model, the values of OOV, WER, and TER in Table 4 could be different from those in Table 3. One of the major reasons for this difference could be in the differences of the specific procedures for calculating these values as well as the difference of the stop word lists. In addition, the TUT group did not use queries dictated by female speakers, for the purpose of experiments.

Acknowledgments

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

References

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

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

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

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

[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, 1997.

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

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

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

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

[11] G. Jones, J. Foote, K. S. Jones, and S. Young. 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*, pages 30–38, 1996.

[12] A. Kai, Y. Hirose, and S. Nakagawa. Dealing with out-of-vocabulary words and speech disfluencies in an n-gram based speech understanding system. In *Proceedings of the 5th International Conference on Spoken Language Processing*.

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

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

[15] M. Masahiko, H. Nishizaki, Y. Kodama, T. Utsuro, and S. Nakagawa. Evaluating LVCSR model combination in NTCIR-3 speech-driven Web retrieval task. In *Proceedings of the 2003 Spring Meeting of the Acoustical Society of Japan*, volume I, 2003. (In Japanese).

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

[17] S. Robertson and S. Walker. 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*, pages 232–241, 1994.

[18] P. Sheridan, M. Wechsler, and P. Schäuble. 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*, pages 99–108, 1997.

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

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

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

[22] M. Wechsler, E. Munteanu, and P. Schäuble. 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*, pages 20–27, 1998.

[23] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal. 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*, pages 26–33, 1999.

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