SPEECH-ACTIVATED TEXT RETRIEVAL SYSTEM FOR MULTIMODAL CELLULAR PHONES

Shin-ya ISHIKAWA, Takahiro IKEDA, Kiyokazu MIKI, Fumihiro ADACHI, Ryosuke ISOTANI, Ken-ichi ISO, Akitoshi OKUMURA

Multimedia Research Laboratories, NEC
s-ishikawa@dg.jp.nec.com

ABSTRACT

This paper describes an on-line manual page retrieval system activated by spoken queries for multimodal cellular phones. The system recognizes user's naturally spoken queries by telephone LVCSR and searches an on-line manual with a retrieval module on a server. The user can view the retrieved data on the screen of the phone via Web access. The LVCSR module consists of a telephone acoustic model and an n-gram language model derived from a task query corpus. The adaptation method using the target manual is also presented. The retrieval module utilizes pairs of words with dependency relations and also distinguishes affirmative and negative expressions to improve precision. The proposed system gives 82.6% keyword recognition accuracy and 77.5% task achievement rate. The field trial of the system is now underway.

1. INTRODUCTION

Cellular phones are now widely used and those with Web browsing capability (Figure 1) are becoming very popular. Users browse and download contents from the Internet such as news, weather, and train timetable in mobile environment. However, with limited number of keys, process of entering data becomes cumbersome with complex keystrokes. Cellular phone users are finding that entering information on their small devices becomes quite a task, especially when searching for documents using information retrieval (IR). It would be convenient if users can voice-enter their search queries. As the recent electric appliances including cellular phones become more sophisticated in its functionality, their owner's manuals become complex in content. Speech input will be an effective interface especially in such data retrieval. For example, a spoken dialog system for appliance manual retrieval has been studied [1]. Such spoken dialogue systems might be available on cellular phones using IVR. IVR systems over the telephone offer speech recognizer and speech synthesizer allowing users to call in and speak to the system, but sometimes, users have to listen to lengthy presentations that are straight from the written manual, or the wrong presentations due to recognition errors.

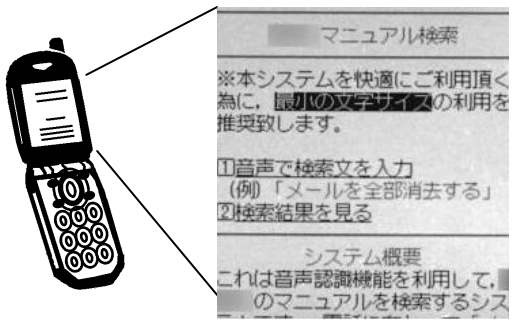

Figure 1: Cellular phone with Web browser. 15 lines of 30 characters displayed.

In this paper, we present a Japanese text retrieval system using spoken queries that allows users to voice-input the queries and view the retrieved result passage on the cellular phone screen. Users can input queries without complicated keystrokes and quickly check the retrieved data on their phone screen. This system is realized by the following techniques:

(1) Integration of Web and voice systems [2]
(2) Telephone LVCSR with a language model trained on a task-specific corpus
(3) Text retrieval which works well for short queries

The spoken queries are transmitted through telephone network from the cellular phone to a telephone server, and LVCSR module recognizes the input. Retrieval module searches the on-line manual using the query input. The user can view the retrieved results on the screen of the phone. To reduce recognition errors in LVCSR, users are recommended to speak query sentences in one breath. Training corpus of recognition language model is collected in the same manner. To improve precision for short queries, retrieval module utilizes pairs of words with dependency relations and also distinguishes affirmative and negative expressions.

We also propose a language model adaptation method that is useful in changing the target manual without re-collecting manual-specific queries for language model estimation. In the following sections, we describe the overview of the system, the details of each module, and system evaluation results.

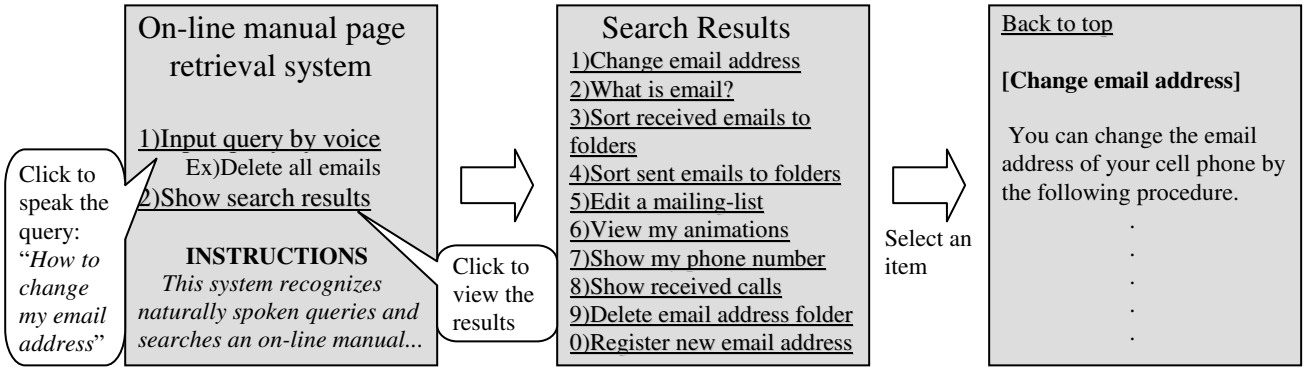

Figure 2. Top screen

Figure 3. 10 best-matched items

Figure 4. Manual page

2. SYSTEM OVERVIEW

The system configuration is shown in Figure 5. The cellular phone has a screen of 30 characters x 15 lines and a Web browser (Figure 1-4). The telephone server which contains LVCSR and retrieval modules and the Web server are implemented on Windows PCs. Interactions between the user and the system are as follows. Note that this system works in Japanese, but explanations and Figures 2-4 are translated into English.

1. The user accesses the Web server via the Internet to view the top screen (Figure 2). There are two hyperlinks on the page along with brief instructions and examples of query.

2. The user clicks the first link labeled "Input query by voice". It is linked to the telephone number of the telephone server and clicking it makes a call to the server.

3. The user inputs a query following the voice guidance from the system. In this example, user spoke "How to change my email address". The LVCSR module in the telephone server recognizes it and sends the result text to the retrieval module. The retrieval module searches the on-line manual text and uploads the retrieved results to the Web server. The system hangs up the phone and switches back to the top screen.

4. Then the user clicks the second link labeled "Show search results" which is linked to the URL of the Web server. The titles of 10 best-matched items are downloaded and listed on the screen (Figure 3).

5. By selecting an item, the corresponding body text is displayed (Figure 4). If no appropriate item is found, the user can go back to the top screen and re-enter a query by speech.

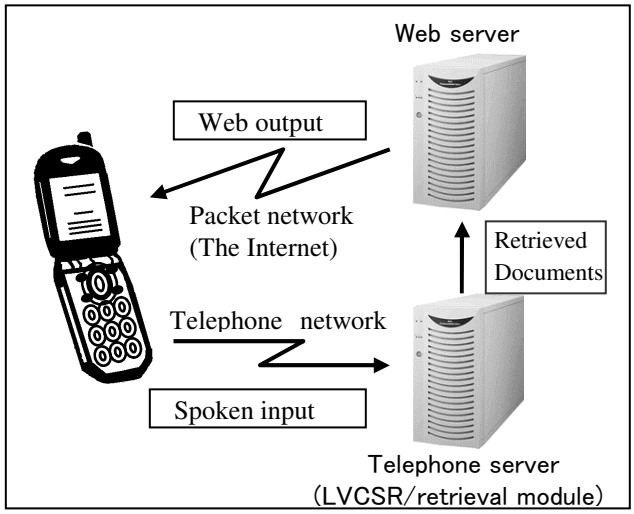

Figure 5. Speech-activated retrieval system for cellular phones

3. TELEPHONE LVCSR MODULE

3.1. Language model

A statistical language model (LM) with word and class n-gram estimates is used. Word 3-gram is backed off to word 2-gram, and word 2-gram is backed off to class 2-gram. Parts of speech are used as the classes of each word. The LM is trained on a text corpus of query samples for this on-line manual retrieval task. Nouns in the manual document are added to the recognition dictionary.

The text corpus was manually constructed. Several people were given a cellular phone and the owner's manual and asked to collect questions they might wish to ask. The instructions given to them are as follows:

1. Assume an on-line manual page retrieval system. It searches for a passage most-relevant to your query. It does not speculate nor always select the suitable answers for your questions.
2. Each query should be easily read in one breath.

3-a. Collect possible questions a user might ask while operating the phone.

3-b. Collect additional questions referring to the owner's manual.

A total of 15k queries were collected by this procedure, most of which were obtained in step 3-b. The LM trained on this text corpus has about 4k words in the recognition vocabulary, about 20k word 2-gram entries, and about 40k word 3-gram entries.

3.3. Acoustic model

A speech signal is sampled at 8kHz, with MFCC analysis frame rate of 10 ms. Spectral subtraction (SS) is applied to remove stationary additive noises. The feature set includes MFCC, pitch, and energy with their time derivatives. The LVCSR decoder supports triphone HMMs with tree based state clustering on phonetic contexts. The state emission probability is represented by Gaussian mixtures with diagonal covariance matrices. Gender-dependent acoustic models were trained on the speech corpus consisting of 200k sentences read by 1385 speakers collected through telephone line.

3.4. LVCSR decoder

The LVCSR decoder recognizes the query utterances with the triphone acoustic model, the statistical language model, and a tree-structured word dictionary. It performs two-stage processing. On the first stage, input speech is decoded by frame-synchronous beam search to generate a word candidate graph using the acoustic model, 2-gram language model, and the word dictionary. On the second stage, the graph is searched to find the optimal word sequence using the 3-gram language model.

Both male and female acoustic models are used and decoding is performed independently for each model except for the common beam pruning in every frame. Recognition results by male and female acoustic models are compared and the one with better score is used as the result. Gender-dependent models improve the recognition accuracy while curbing the increase of the computational amount by using common beam pruning.

3.5. LVCSR evaluation

The LVCSR module was evaluated using keyword accuracy as the measure. The keyword accuracy is calculated as follows. Each recognition result is transformed to the list of keywords by extracting content words from it preserving their order. The same transformation is applied to the corresponding correct answer, and the two lists are compared to calculate normal word accuracy assuming they are ordinary sentences. Percent correct of keywords (keyword P.C.) is also calculated. Table 1 shows the keyword accuracy evaluated on the newly collected 304 query utterances by 12 male speakers. The test set perplexity of the language model is 54. The average number of content words in each query is 3.0.

Table 1. Evaluation results of LVCSR module

Keyword P.C.	Keyword accuracy
88.6%	82.6%

3.6. LM adaptation to a new model using its manual

Newer cellular phone models are frequently released with additional functions like many other digital devices. For example, new model M2 can handle the Flash animations and an external memory card that were unavailable in older model M1. The LM training corpus described in section 3.1 is for phone M1. It would be convenient if the language model could be adapted to the new model M2 without collecting a new corpus. In this section, we propose several methods to train LM for new cellular phone M2 using its owner's manual document (CP2) and the LM training corpus for M1 (CP1):

(1) Addition of 1-gram counts of keywords in CP2: Each 1-gram count of nouns and verbs in CP2 is added to the counts in CP1, and the n-gram probabilities are calculated. This includes adding new words that are seen only in CP2 to the recognition dictionary. Word 2-gram, 3-gram, and class 2-gram counts are not affected by this addition, but when word 2-gram is backed off to class 2-gram, the probability is distributed by the modified 1-gram counts. New keywords are used only by back-off.

(2) Addition of counts of keyword 3,2,1-grams in CP2: Not only 1-gram counts but keyword 3,2-gram counts are added to corresponding counts in CP1. Here the keywords are nouns, verbs, adjectives, and function words that join those words.

(3) Mixing the CP1 and CP2 to train the LM: This is just for comparison.

Method (2) is expected to outperform method (3) because it adds only word sequences in the manual document that are thought to be common with queries.

Table 2 shows the results evaluated on newly collected query utterances by 15 male and 3 female speakers for M2 cellular phone. The LM for M1(Base model) does not cover M2 queries well enough. The methods (2) and (3) effectively adapt the LM to a new model, and the proposed method (2) gives the better performance than the conventional method (3), which shows the effectiveness of the proposed method.

Table 2. Evaluation of LM adaptation method

	Keyword P.C.	Keyword accuracy
Base model	80.4%	76.4%
(1) proposed-1	86.0%	81.5%
(2) proposed-2	87.3%	82.4%
(3) conventional	87.0%	82.2%

4. TEXT RETRIEVAL MODULE

The on-line manual for retrieval has about 600 pages including 11k sentences and 4k unique words. The retrieval module in this system retrieves relevant passages from the target on-line manual using a query entered by a user with speech recognition. In this system, a passage is a part of the document corresponding to a feature in the on-line manual. Our retrieval module adopts a word-based retrieval method. It generates indexes for content words in passages and obtains relevant passages from the words in the utterance based on Okapi BM25 probabilistic retrieval model [3] in principle. We extended the retrieval model on the following points to improve the retrieval quality customized for the manual.

1) Utilization of pairs of words with dependency relations: It is difficult to identify passages in an on-line manual based on an individual word since these manuals usually describe all the functions of the product exhaustively. For example, a word "mail" would be used in a passage explaining various functions such as sending mails, receiving mails, composing mails, and many others. Our system improves precision by assigning larger relevance scores for passages including the same pair of words with dependency relations as in the query.

2) Distinction between the negative and the affirmative phrases by auxiliary words: An on-line manual often includes a pair of features for each function: one activating and the other deactivating the function. In such cases, phrases with the same set of content words can denote two different features depending on whether the phrase is in the affirmative or in the negative (ex. "Sending the caller number" and "Not sending the caller number"). Our system improves precision by distinguishing the phrases in the negative from that in the affirmative based on their auxiliary words.

3) Aggregation of synonyms: The user often inputs a query using words which may not appear in the target manual. We developed a synonym dictionary for principal words used in the manual. Our system aggregates synonyms based on the developed dictionary and retains sufficient recall.

5. SYSTEM EVALUATION

Evaluation of the proposed system was conducted using the subset of test utterances as described in 3.5 that have relevant information in the on-line manual. The 1st, 5th, and 10th retrieval rates were calculated. We define the n-th retrieval rate as a percentage of the correct answers for the query included in n-best matched results. The results are shown in Table 3. The retrieval rates by the transcriptions of the evaluation utterances are also shown for comparison.

Table 3. System evaluation using spoken queries

	1st/5th/10th retrieval rates(%)	number of content words in average
Using LVCSR	35.6 / 67.8 / 77.5	3.0
(Transcription)	35.7 / 68.3 / 78.6	3.4

The retrieval rates by LVCSR are almost the same as those by transcription. Since the cellular phones used in this system can display about 10 feature lines on the average, the 10th retrieval rate represents the rate of successfully delivering the passage requested by the user. The results show that users can find answers to their queries in 77.5% of the time by this system, which we think is sufficient for practical use.

The system performance was evaluated for the queries that have relevant information in the on-line manual (89% of total). Further efforts need to be made to improve the system to detect the absence of relevant information and notify it to the user.

6. CONCLUSION

In this paper, we presented a Japanese text retrieval system using spoken queries for multimodal cellular phones, and applied it to an on-line manual of cellular phones. The system recognizes user's naturally spoken queries and retrieves the relevant passages in the on-line manual. The user can view the feature titles of the retrieved passages and their body text on the screen of the cellular phone. To realize this system, we developed a telephone LVCSR and text retrieval for short queries, and integrated voice and Web systems. We also proposed a method of adapting the language model to a new phone model and showed its effectiveness. The evaluation results show that users can reach the intended pages in 77.5% of the queries if the answers exist in the on-line manual. Field trial of this system is now underway.

7. REFERENCES

[1] Kawahara, T., Ito, R., Komatani, K., "Spoken dialogue system for queries on appliance manuals using hierarchical confirmation strategy", *Proc. of Eurospeech2003*, pp. 1701-1704.

[2] Yoshida, K., Hagane, H., Hatazaki, K., Iso, K., Hattori, H., "Human-Voice Interface", *NEC Res. & Develop.*, Vol.43, No. 1, January 2002.

[3] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., and Gatford, M., "Okapi at TREC-3," *Proc. of the 3rd Text Retrieval Conference*, pp.109-126, 1995.