Speech-Activated Text Retrieval System for Cellular Phones with Web Browsing Capability

Takahiro Ikeda **Shin-ya Ishikawa** **Kiyokazu Miki** **Fumihiro Adachi**
Ryosuke Isotani **Kenji Satoh** **Akitoshi Okumura**

Media and Information Research Laboratories, NEC Corporation
1753 Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa 211-8666, Japan

t-ikeda@di.jp.nec.com s-ishikawa@dg.jp.ne.com k-miki@bq.jp.nec.com f-adachi@aj.jp.nec.com
r-isotani@bp.jp.nec.com k-satoh@da.jp.nec.com a-okumura@bx.jp.nec.com

Abstract

This paper describes a text retrieval system for cellular phones with Web browsing capability, which accepts spoken queries over the cellular phone and provides the search result on the cellular phone screen. This system recognizes spoken queries by large vocabulary continuous speech recognition (LVCSR), retrieves relevant document by text retrieval, and provides the search result on the World Wide Web by the integration of the Web and the voice systems. The text retrieval in this system improves the performance for spoken short queries by: 1) utilizing word pairs with dependency relations, 2) distinguishing affirmative and negative expressions, and 3) converging synonyms. The LVCSR in this system shows enough performance level for speech over the cellular phone with acoustic and language models derived from a query corpus with target contents. The system constructed for user's manual for a cellular phone navigates users to relevant passages for 81.4% of spoken queries.

1. Introduction

Cellular phones are now widely used and those with Web browsing capability are becoming very popular. Users can easily browse information provided on the World Wide Web such as news, weather, and traffic report with the cellular phone screen in mobile environment. However, obtaining necessary information from large database such as user's manual or travelers' guide is quite a task for users since searching for appropriate information from seas of data requires cumbersome key operations. In most cases, users have to carefully navigate through deep hierarchical structures of menus or have to type in complex combination of keys to enter some keywords.

Text retrieval by voice input is one of the solutions for this problem. This paper presents a telephone-based voice query retrieval system in Japanese which enables cellular phone users to search through the user's manual. This system accepts spoken queries over the cellular phone with large vocabulary continuous speech recognition (LVCSR) and retrieves relevant parts from the user's manual with text retrieval. The results are provided to the user as a Web page by synchronously activating the Web and the voice systems (Yoshida et al., 2002). Users can input queries without complicated keystrokes and can view the list of results on the cellular phone screen.

With respect to voice input systems, a large number of interactive voice responses (IVR) systems and spoken dialogue systems has been designed and developed over the years (Zue, 1997). As for user's manual retrieval systems which accept voice input, Kawahara et al. (2003) has developed a spoken dialogue system for appliance manuals. However, they mainly focus on the dialogue strategy to select the appropriate result on screen-less systems such as VTR and FAX. On the other hand, retrieval methods for voice input have been examined on a TREC query set (Barnett et al., 1997; Crestani, 2000).

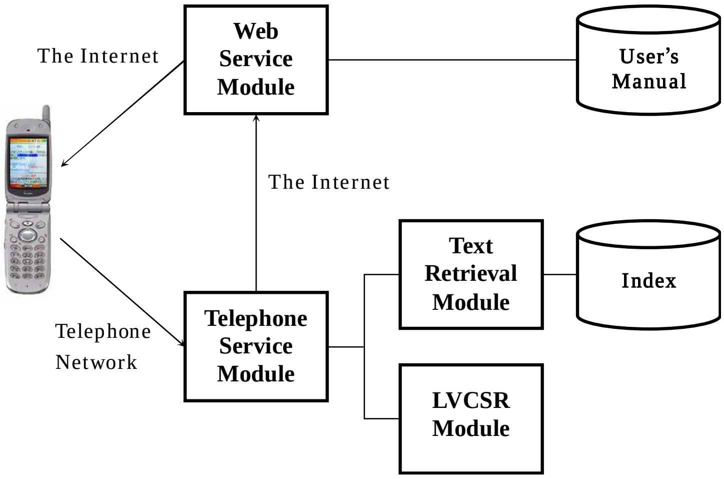


Figure 1: The configuration of the prototype system.

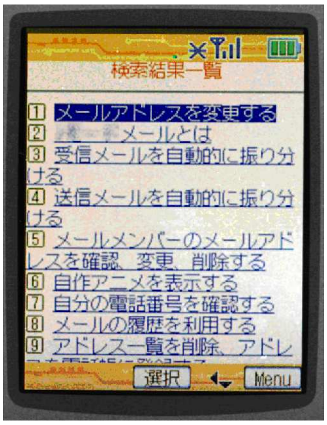

検索結果一覧
1 メールアドレスを変更する
2 [redacted]メールとは
3 受信メールを自動的に振り分ける
4 送信メールを自動的に振り分ける
5 メールメンバーのメールアドレスを確認、変更、削除する
6 自作アニメを表示する
7 自分の電話番号を確認する
8 メールの履歴を利用する
9 アドレス一覧を削除、アドレ

Figure 2: The screen of the cellular phone displaying the search result.

. However, text retrieval in TREC mainly aims to search open domain documents from long queries, while our system is required to search closed domain documents such as user's manuals based on short queries spoken over the cellular phone.

In order to apply text retrieval technique to speech-activated user's manual retrieval, we have investigated queries for searching manuals in addition to the text of the manuals from a linguistic viewpoint. We found that text retrieval for a user's manual has the following three difficulties.

1) The difficulty of identifying passages in a user's manual based on an individual word.

2) The difficulty of distinguishing affirmative and negative sentences which mean two different features in the manual.

3) The difficulty of retrieving appropriate passages for a query using words not appearing in the manual.

This paper presents how we overcome these difficulties using three techniques: 1) utilizing word pairs with dependency relations, 2) distinguishing affirmative and negative expressions by auxiliary verbs, and 3) converging synonyms with synonym dictionary.

The rest of the paper is organized as follows. Section 2 describes the system configuration of our speech-activated text retrieval system and how it works. Section 3 discusses the difficulties in text retrieval in our system and presents our proposed techniques in detail. Section 4 shows the developed prototype system and Section 5 reports its evaluation results. Finally Section 6 concludes the paper.

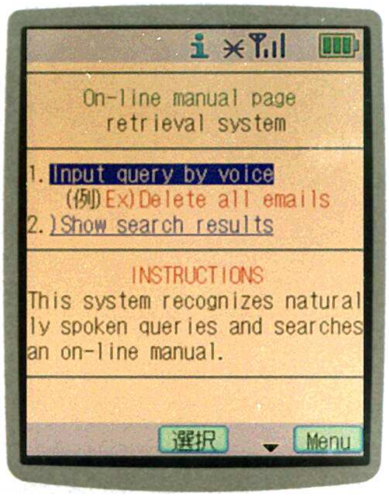
2. Speech-Activated Text Retrieval System

Our system receives spoken queries on the usage of the cellular phone and provides the list of relevant passages in the user's manual. In this paper, a passage denotes a part of the document corresponding to a feature in the user's manual.

2.1. System Configuration

Figure 1 shows the configuration of our retrieval system.

The telephone service module receives a phone call from the user. This module prepares the search operation by calling the LVCSR module, which recognizes the query spoken over the phone, and the text retrieval module, which provides the search result for the query.

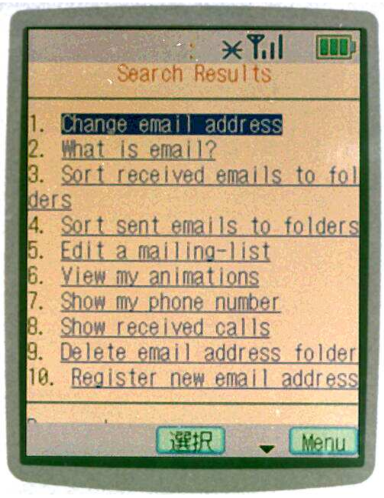


On-line manual page retrieval system

1. Input query by voice
(例)Ex)Delete all emails
2.)Show search results

INSTRUCTIONS
This system recognizes naturally spoken queries and searches an on-line manual.

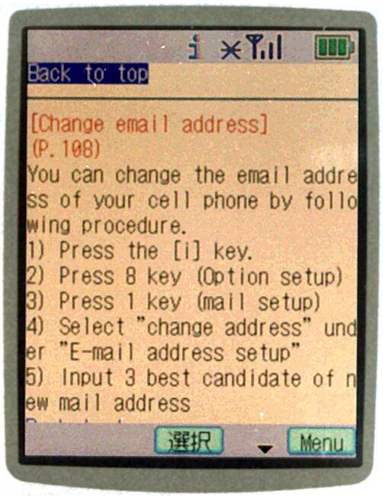
Figure 3: The main page of our system.



Search Results

1. Change email address
2. What is email?
3. Sort received emails to folders
4. Sort sent emails to folders
5. Edit a mailing-list
6. View my animations
7. Show my phone number
8. Show received calls
9. Delete email address folder
10. Register new email address

Figure 4: The result page displaying the title list of top ten results for the query.



Back to top

[Change email address]
(P.108)
You can change the email address of your cell phone by following procedure.
1) Press the [i] key.
2) Press 8 key (Option setup)
3) Press 1 key (mail setup)
4) Select "change address" under "E-mail address setup"
5) Input 3 best candidate of new mail address

Figure 5: The body of the passage displayed when the user selects the title in Figure 4.

The telephone service module sends the list of the relevant passages to the Web service module, and then hangs up the phone. The Web service module provides the result to the user according to the user's request via the internet.

We assume that the cellular phone screen displays about 30 letters per line and 15 lines of text according to the specifications of recent popular cellular phones in Japan. We assign top ten potential passages as the search result and display the title of them in order for the user to see with ease.

Figure 2 shows the screen of the cellular phone displaying the search result.

2.2. Example of Using the System

This section describes how our system works. Our system works in Japanese, but in the following section, English translation is provided for the reader's convenience. In our system, the user obtains the relevant passage in the user's manual with the voice query according to the following steps.

Step 1: The user first accesses the system's main page of our system with the cellular phone (Figure 3). The page contains two hyperlinks along with brief instructions and query examples.

Step 2: The user follows the first link labeled "Input query by voice." It is linked to the telephone service module, allowing the user to call the telephone service module.

Step 3: The user inputs a query following the voice guidance from the system. The LVCSR module recognizes it and outputs the result text. The text retrieval module searches the user's manual from recognized text and outputs the top ten results. The user goes back to the main page after the telephone service module hangs up the phone.

Step 4: The user follows the second link labeled "Show search results," which is linked to our Web service module. Then the user views the result page which contains the title list of top ten results (each passage consists of a title and a body). Figure 4 shows the example of the result page responding to the voice query "How to change my email address."

Step 5: By selecting a title of a passage from the result list, the user retrieves the corresponding body of the passage (Figure 5). If the result list contains no relevant passages, the user can go back to the homepage and re-enter a query by speech.

3. Text Retrieval for a User's Manual

3.1. The Problems on User's Manual Retrieval

In general, user's manual of equipment explains all functions extensively. Since the phrasing used in a user's manual is often similar, expressions with small difference might appear in completely different entries. We have investigated queries for searching manuals in addition to the text of the manuals from a linguistic viewpoint and found that text retrieval for user's manual has the following three difficulties.

1) It is difficult to identify passages in a user's manual based on an individual word. For example, a word "mail" shows up in passages explaining various functions such as sending mails, receiving mails, composing mails, and many others. In order to overcome this difficulty, we need to use relations between words.

2) It is difficult to distinguish affirmative and negative sentences based on independent words. Sentences with the same set of content words can mean two different features depending on whether the sentence is in the affirmative or in the negative. This is often true in manual writings where each function is described in pair: one activating and the other deactivating the function (ex. "Sending the caller number" and "Not sending the caller number"). In order to overcome this difficulty, we need to handle polarity indicated by auxiliary verbs.

3) It is difficult to retrieve appropriate passages for a query using words not appearing in the manual. While the expression denoting an object is generally standardized in a user's manual, users often indicate the object with other expressions. In order to overcome this difficulty, we need to assimilate difference of various synonymous expressions.

3.2. The Approaches for User's Manual Retrieval

The system retrieves relevant passages from the user's manual with a word-based text retrieval method. The system generates indexes for content words in passages and obtains relevant passages from the words in the query based on Okapi BM25 probabilistic retrieval model without relevance feedback in principle (Robertson et al., 1993). In this model, the weight *W* of a passage *P* for a query *Q* is defined as follows:

$$W = \sum_{T \in Q} TW(T)$$

$$TW(T) = w \cdot \frac{(k_1 + 1) \cdot tf}{k_1 \cdot K + tf} \cdot \frac{(k_2 + 1) \cdot qtf}{k_2 + qtf}$$

$$w = \log \frac{N - n + 0.5}{n + 0.5}$$

$$K = (1 - b) + b \cdot \frac{PL}{AVPL}$$

Here *T* denotes a term in the query *Q*, *N* denotes the number of passages in the whole text, *n* denotes the number of passages containing the term *T*, *tf* denotes the frequency of occurrence of the term *T* within the passage *P*, *qtf* denotes the frequency of occurrence of the term *T* within the query *Q*, *PL* denotes the length of the passage *P*, and *AVPL* denotes the average length of all passages. k_1, k_2, and *b* are predefined constants.

Table 1: An example of a synonym dictionary

Standard Expression	Synonymous Expressions	
saito (site)	*webu* (web)	*hômupêji* (homepage)
chakushin'on (ringtone)	*chakushinmerodî* (ring melody)	*yobidashion* (phone beep)
ridaiaru (redial)	*môichido* → *kakeru* (again → call)	

In order to overcome the difficulties stated previously, we have expanded the retrieval model with the following three techniques.

1) Utilization of word pairs with dependency relations

This technique assigns larger weight for passages including the same word pairs with dependency relations as in the query. The system uses the following weight W_{wp}, which is simple extension of W:

$$W_{\mathrm{wp}} = {k_{\mathrm{wp}}}^{NP} \cdot W$$

where NP denotes the number of word pairs which appear both in the passage P and the query Q with dependency relations. k_{wp} is predefined constants.
We detect the dependency between words by shallow dependency analysis without parsing. The system assigns depend-to and depend-from attributes to each word based on its part of speech and connects them according to the surrounding relationship (Satoh et al., 2003).

2) Distinction between the negative and the affirmative phrases by auxiliary verbs

This technique assigns the different weight on the term according to the condition whether an auxiliary verb indicating negative polarity follows after the term. The system adds this condition to each word after morphological analysis, and distinguishes words with different conditions. The system uses the following weight W_{aux} instead of W:

$$\begin{aligned} W_{\mathrm{aux}} = & \sum_{T^{+} \in Q} \left(TW(T^{+}) + k_{\mathrm{aux}} \cdot TW(T^{-}) \right) \\ & + \sum_{T^{-} \in Q} \left(TW(T^{-}) + k_{\mathrm{aux}} \cdot TW(T^{+}) \right) \end{aligned}$$

where T^{+} denotes the term T with this condition and T^{-}denotes the term T without this condition. k_{aux} is predefined constants.

3) Converging synonyms

This technique assumes the occurrence of synonymous expressions for a word as the occurrence of the word itself in calculating the weight. The system converges various synonymous expressions into the standard expression by using predefined synonym dictionary. The system accepts a set of words with dependency relations as a synonymous expression in order to converge complex synonymous expressions.

Table 1 shows an example of a synonym dictionary. An arrow sign denotes a dependency relation between words.

4. Prototype System

We have constructed a prototype system to search through the manuals for cellular phone users (Ishikawa et al., 2004). The user's manual contains about 14,000 passages and consists of about 4,000 unique words. The prototype system works in real time according to the user's operation.

4.1. LVCSR Module

4.1.1. Language Model

A statistical language model (LM) with word and class n-gram estimates is used in our system. Word 3-gram is backed off to word 2-gram, and word 2-gram is backed off to class 2-gram. Part-of-speech patterns are used as the classes of each word. The LM is trained on a text corpus of query samples for our target user's manual. Nouns in the manual document are added to the recognition dictionary apart from the training.

A total of 15,000 queries were manually constructed and used for training the LM. The final LM for the prototype system has about 4,000 words in the recognition vocabulary, about 20,000 word 2-gram entries, and about 40,000 word 3-gram entries.

4.1.2. Acoustic Model

A speech signal is sampled at 8kHz, with MFCC analysis frame rate of 10ms. 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.

For the prototype system, Gender-dependent acoustic models were prepared by the training on the speech corpus with 200,000 sentences read by 1,385 speakers collected through telephone line.

4.1.3. 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 common beam pruning.

4.2. Text Retrieval Module

All the techniques described in Section 3.2 are implemented on the text retrieval module in the system. We fixed the constants as follows according to the preliminary experiments using query samples developed for training the LM:

$$k_1 = 100,\ k_2 = 1000,\ b = 0.3,\ k_{\mathrm{wp}} = 1.3,\ k_{\mathrm{aux}} = 0.3$$

We developed the synonym dictionary with about 500 entries to converge synonymous expressions used to describe cellular phone functions.

Table 2: Examples of queries used for evaluation.

Shashin-o mêru-de okuritai (I want to send a picture via email)
Aikon-o desukutoppu-ni tôroku shitai (I want to register a new icon on the desktop)
Jushin-shita mêru-o minagara henshin mêru-o sakusei-suru hôhô (How to write a reply mail while looking at the incoming mail)

Table 3: The retrieval success rate for the transcriptions of queries.

Number of Result Passages	Retrieval Success Rate for Transcriptions			
	BL	WP	WP+AUX	ALL
1	40.0%	42.7%	44.5%	49.1%
5	65.5%	69.1%	70.0%	77.3%
10	73.6%	73.6%	74.5%	87.3%

Table 4: The retrieval success rate for the utterances of queries.

Number of Result Passages	Retrieval Success Rate for Utterances
1	44.3%
5	72.5%
10	81.4%

5. Evaluation

In order to evaluate the usefulness of our system, we have composed 150 new queries independently of the query corpus used for configuring the system. We have used 110 queries for evaluation, eliminating 40 queries without relevant passages in the manual. Table 2 shows some examples of the queries used for the evaluation. Each query contains 3.8 words in average.

The retrieval success rate, which we adopted as a criterion, measures how well the system is able to provide a relevant passage within the top predefined number of result passages. We have calculated the retrieval success rates at 1, 5, and 10 passages for several conditions.

In order to discuss the effect of each technique presented in Section 3.2, we first present the result for transcriptions of the queries among the following text retrieval methods.

Method BL: This is the baseline method with no techniques applied.

Method WP: This method utilizes word pairs with dependency relations.

Method WP+AUX: This method distinguishes between the negative and the affirmative phrases by auxiliary verbs in addition to the method WP.

Method ALL: This method converges synonyms in addition to the method WP+AUX. This is the same condition as the prototype system.

Table 3 summarizes the result. The result shows each of the three techniques has contributed to the improvement of the retrieval success rate. Especially, converging synonyms enhances the performance as derived from the difference between methods WP+AUX and ALL.

Next we present the performance of the total system. Table 4 shows the result for 660 utterances of the queries by 18 speakers where the LVCSR module and the text retrieval module in the prototype system are used. The retrieval success rates for utterances are almost the same as those for transcription. Since the cellular phones used in this system can display about 10 lines on the average, the 10th retrieval rate represents the rate of successfully delivering the passage requested by the user. The result shows that the system designed for cellular phone user's manual was able to direct user to appropriate information at 81.4%, which is sufficient for practical use.

6. Conclusions

In this paper, we presented a voice query retrieval system in Japanese applied to document search on user's manual for cellular phones with Web access capability. The system recognizes user's naturally spoken queries over the cellular phone by LVCSR and retrieves the relevant passages by text retrieval and then provides the output on the cellular phone screen. In order to improve the performance for spoken short queries, we apply three techniques into text retrieval: 1) utilizing word pairs with dependency relations, 2) distinguishing affirmative and negative expressions, and 3) converging synonyms. With respect to LVCSR for speech over the cellular phone, we adopt acoustic and language models derived from a query corpus for the target user's manual. The evaluation on the system designed for cellular phone user's manual shows that the system is able to direct users to appropriate data at 81.4% of the time, if the matching passage exists in the manual.

Our next step is to apply this system to different contents such as travelers' guide and customer surveys. We plan to clarify the problems for different contents and to enhance the portability of this system.

References

Barnett, J., S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. 1997. *Experiments in Spoken Queries for Document Retrieval.* Proceedings of Eurospeech'97, pp.1323–1326.

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

Ishikawa, S., T. Ikeda, K. Miki, F. Adachi, R. Isotani, K. Iso, and A. Okumura. 2004. *Speech-activated Text Retrieval System for Multimodal Cellular Phones.* Proceedings of ICASSP 2004.

Kawahara, T., R. Ito, and K. Komatani. 2003. *Spoken Dialogue System for Queries on Appliance Manuals using Hierarchical Confirmation Strategy.* Proceedings of Eurospeech 2003, pp.1701–1704.

Robertson, S. E., S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. 1995. *Okapi at TREC-3.* Proceedings of the 3rd Text Retrieval Conference (TREC-3), pp.109–126.

Satoh, K., T. Ikeda, T. Nakata, and S. Osada. 2003. *Design and Development of Japanese Processing Middleware for Customer Relationship Management.* Proceedings of the 9th Annual Meeting of The Association for Natural Language Processing, pp.109–112 (in Japanese).

Yoshida, K., H. Hagane, K. Hatazaki, K. Iso, and H. Hattori. 2002. *Human-Voice Interface.* NEC Research & Development, 43(1), pp.33–36.

Zue, V. 1997. *Conversational Interfaces: Advances and Challenges.* Proceedings of Eurospeech '97, KN9–18.