Speech Recognition Smartphone UI

Technology Reports

Speech Recognition Technology and Applications for Improving Terminal Functionality and Service Usability

User interfaces that utilize voice input on compact devices such as mobile terminals have been attracting attention as a means of performing complex operations in a more intuitive and less burdensome way. NTT DOCOMO is also using voice input to provide text input and initiate functions on terminals, and is working to realize a user interface in the future that will be able to respond to any spoken input. This research implements a large vocabulary based on a large data set in order to improve speech recognition performance and demonstrates the effectiveness of this approach. We also apply language processing to the results of speech recognition and introduce a sample application developed to support user operations.

Research Laboratories — *Shinya Iizuka*, *Kosuke Tsujino*
Service and Solution Development Department — *Shin Oguri*
Communication Device Development Department — *Hirotaka Furukawa*

1. Introduction

In recent years, it has become possible to use mobile terminals for a variety of services, beyond the basic communication tool functions such as voice calling and e-mail, through added functionality and applications. For example, users can make reservations and pay for holiday accommodation and transportation, check site-seeing maps and share their vacation photos on the Web, all from a single mobile terminal.

On the other hand, as such functionality and services become richer and more sophisticated, users are required to have more skill and perform more complex operations in order to use them effectively. Users must first understand how to access the desired information or service, and then perform the appropriate detailed settings and operations according to their specific situation.

We are conducting research and development on speech input with the goal of providing a less burdensome user interface on compact devices such as mobile terminals for performing these increasingly complex operations. With speech input, direct instructions can be given, even for complex hierarchical or compound conditions, so it is attracting attention as a means of reducing the burden of such operations.

Smartphones have become more common in recent years and many provide applications using speech input. NTT DOCOMO is also providing text input and operation of terminal functions using speech input, and we are working to implement a user interface that can respond when users simply says what they want, as shown in **Figure 1**. In the envisioned implementation, the mobile terminal will understand the intention of the utterance, and will provide a one-stop solution suitable to the need. This eliminates the need for complex opera-

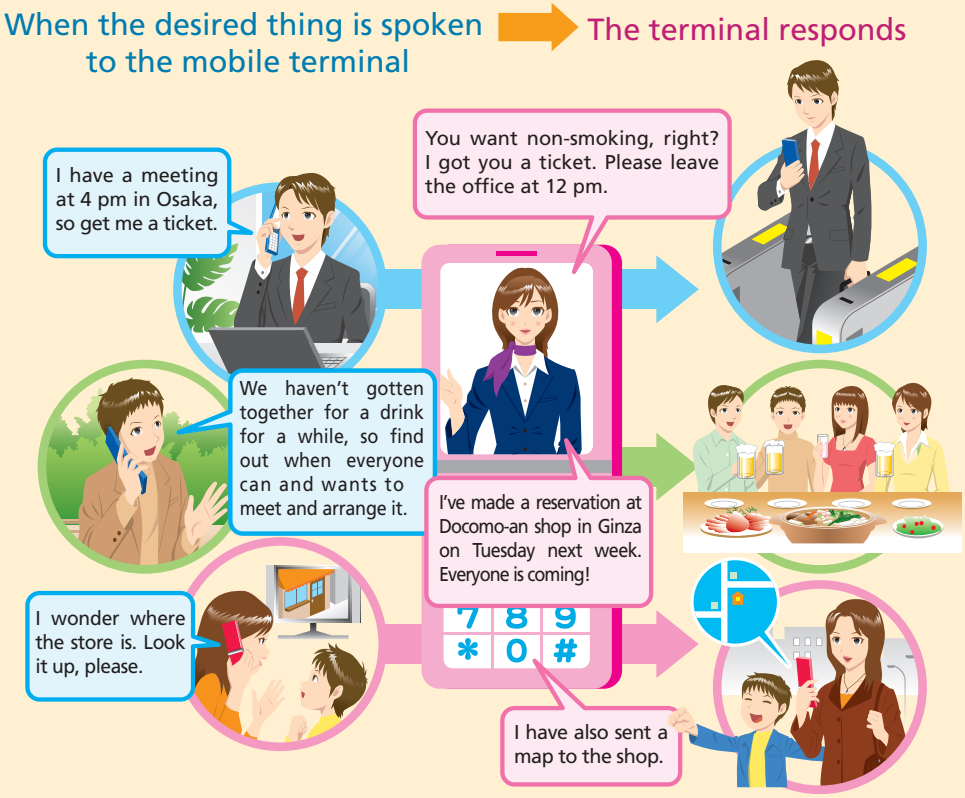


Figure 1 Envisioned system

tions by the user. For such an implementation, technology is needed in the areas of speech recognition, to convert the sound of the user's utterance into text, and language processing, to understand the meaning of this text and decide on an appropriate response to be taken by the application.

In this article, we give an overview of speech recognition technology and describe a sample application developed using speech recognition and language processing technology.

2. Speech Recognition Technology

2.1 Speech Recognition Methods and Characteristics

Some speech recognition methods operate directly on the terminal, while others operate on a server. Those that operate on the terminal have a speech recognition engine on the terminal. Those that use a server transmit the sound signal or sound feature values[*1] to the server, which runs a speech recognition engine and returns the text result back to the terminal.

Speech recognition on a terminal is restricted to relatively small vocabularies due to limitations in processing and power consumption, but it is not affected by communication conditions such as delay or being out of range. It is applied to applications such as terminal operations, which are limited but must be available at all times. On the other hand, server-based speech recognition is affected by the communications state but can use techniques that require relatively more processing. This makes it suitable for applications such as search or text input, which must support larger vocabularies.

*1 **Feature value**: The result of extracting only the information needed for speech recognition from the voice waveform. For feature values, Mel-Frequency Cepstrum Coefficients (MFCC) are often used, obtained by processing the voice waveform with a Fast Fourier Transform (FFT), aural characteristic filters and other processes.

2.2 Increasing Sophistication of Speech Recognition

The structure of a typical speech recognition system is shown in **Figure 2**. The input speech signal is processed to extract feature values using frequency response analysis, and these feature values are input to a speech recognition engine. The speech recognition engine compares and collates the input feature values with acoustic and language models trained using previously accumulated data, determines a list of the most likely morphemes[*2] and outputs this as the result. The acoustic model expresses the correspondence between speech feature values and phonemes[*3] (individual vowels and consonants), and the language model expresses the likelihood that a morpheme would precede or follow a given morpheme.

The accuracy of speech recognition depends on the conditions that above mentioned models are trained in how close to actual input environment. In other words, it is important to reflect the features of the actual user when training the acoustic model. On the other hand, it is necessary to include a large vocabulary in order to recognize a wide range of utterances when training the language model. Thus, building a language model with a large vocabulary requires training with a large text data set.

The authors built a language model with a vocabulary of several hundred thousand words and verified that the large vocabulary improved recognition performance. In building this language model, a large text data set with a diversity of expressions was used, but structuring it accurately as language was an issue. Thus, we increased the accuracy using processes such as screening the text data automatically, optimizing boundaries for morpheme analysis, and attaching pronunciations (yomigana) to morphemes.

3. Applications

3.1 Speech Function Call Application

For the summer 2011 mobile terminal models, we developed the technology used in the autumn 2010 i-mode models for the voice function calling application [1], for Android™[*4] smartphones. We also expanded the functionality, focusing mainly on being able to launch user-downloaded applications and to call the functions intended by the user. Operation of the terminal function calling application, "Voice Action," is shown in

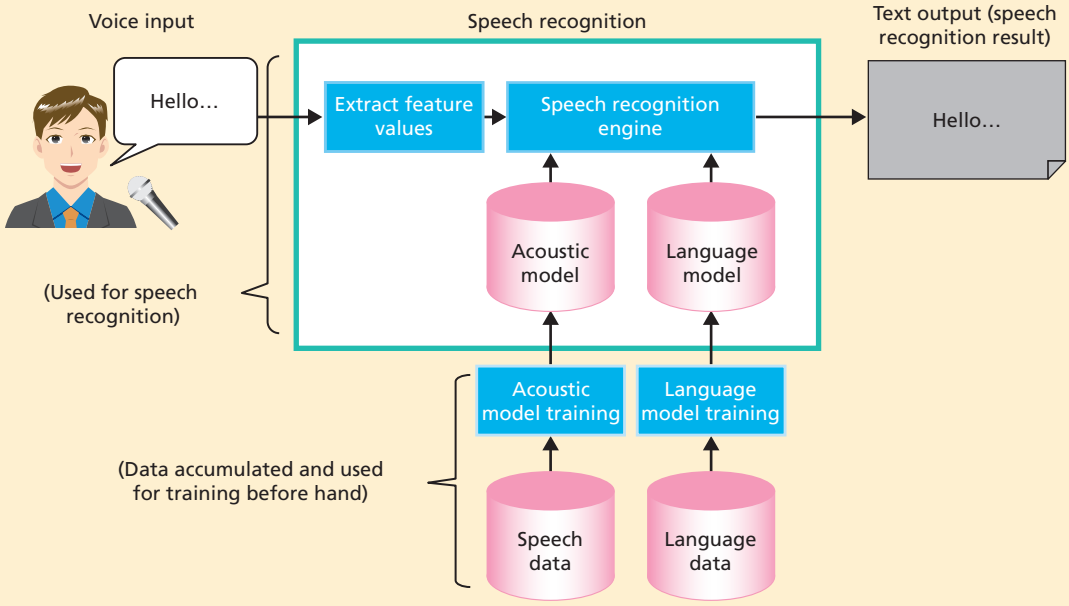


Figure 2 Structure of a speech recognition system

*2 **Morpheme**: In a given language, the smallest units of meaning that cannot be divide further. In addition to words, prefixes such as the honorific "oh-" in Japanese as well as suffixes are included. Automatic division of a sentence into morphemes is called morphological analysis.

*3 **Phoneme**: A smallest unit of sound used for discriminating meaning in a language.

*4 **Android™**: An open source platform targeted mainly at mobile terminals or promoted by Google Inc., in the United States. Android™ is a trademark or registered trademark of Google Inc., in the United States.

Figure 3.

The speech function-calling application associates the pronunciations of predetermined utterance keywords with the names of menu items and applications in the terminal. It can then substitute the result of speech recognition in the terminal for the function ID associated with the utterance keyword best matching the result, and launch the corresponding function.

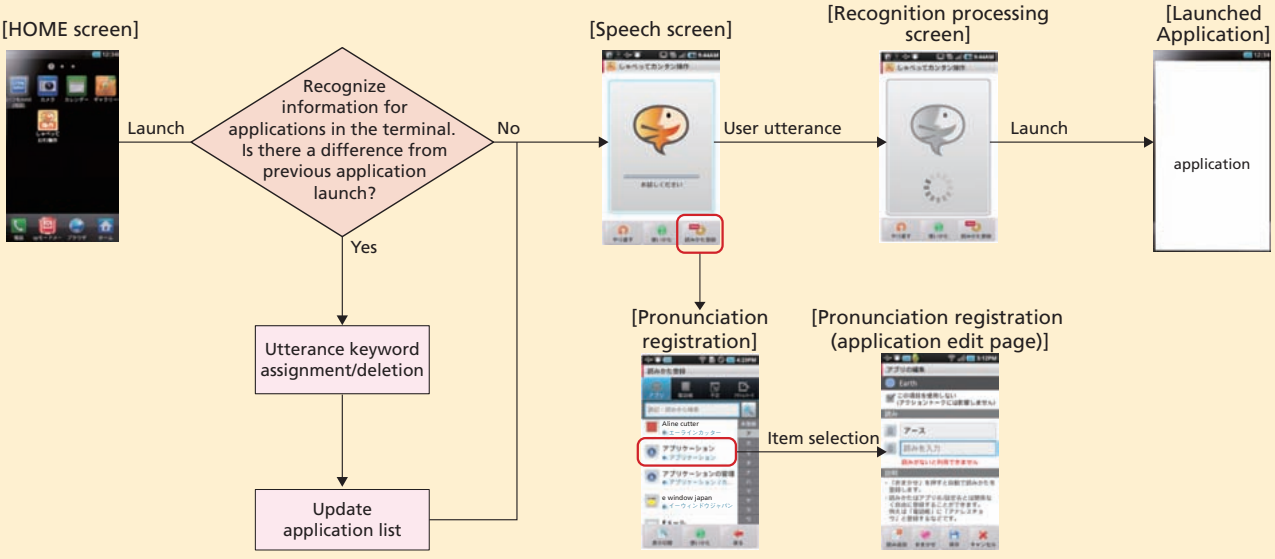
With smartphones, it is assumed that users will load arbitrary applications onto the device, so it is no longer possible to decide and pre-register unique names for launching all menus and applications that might be on the device.

Thus, as an extension for smartphones, we developed a mechanism to allow launching of applications downloaded later. This mechanism maintains an application list representing the current state of the device by detecting information about applications on the device (application and package names) whenever the voice function call application is launched, and adding or deleting them from the list.

Actually, it would be possible, to some extent, for the system to attach pronunciations to application names by using morpheme analysis on the name string obtained from the application. In the characteristics of morpheme analysis, use of English, numbers and word play, it is not possible to select the correct pronunciation. Such a system would also prevent users from calling applications other names that they may be accustomed to.

To resolve these issues, a pronunciation registration list screen for associating utterance keywords to application names was added to the speech function-calling application, providing a mechanism for the user to edit pronunciations. The pronunciation field not only allows the associated utterance keyword to be edited, but multiple pronunciations can be registered for a single application. This provides for a wider range of utterance keywords and more flexibility.

This development enables users to call applications in their smartphone freely, by voice, and according to how they use their device.



*This is provided only in Japanese at present.

Figure 3 Operation of "Voice Action"

3.2 Integrated User Interface Application

In order to implement an integrated user interface that enables seamless access across all types of Web services and mobile terminal operations, we developed a language processing technology that assumes categories for terminal functions and services that are associated with utterances. We also created a prototype application called "VOICE IT!" using the new technology, for Android OS smartphones. The application was provided as a trial in May, 2011.

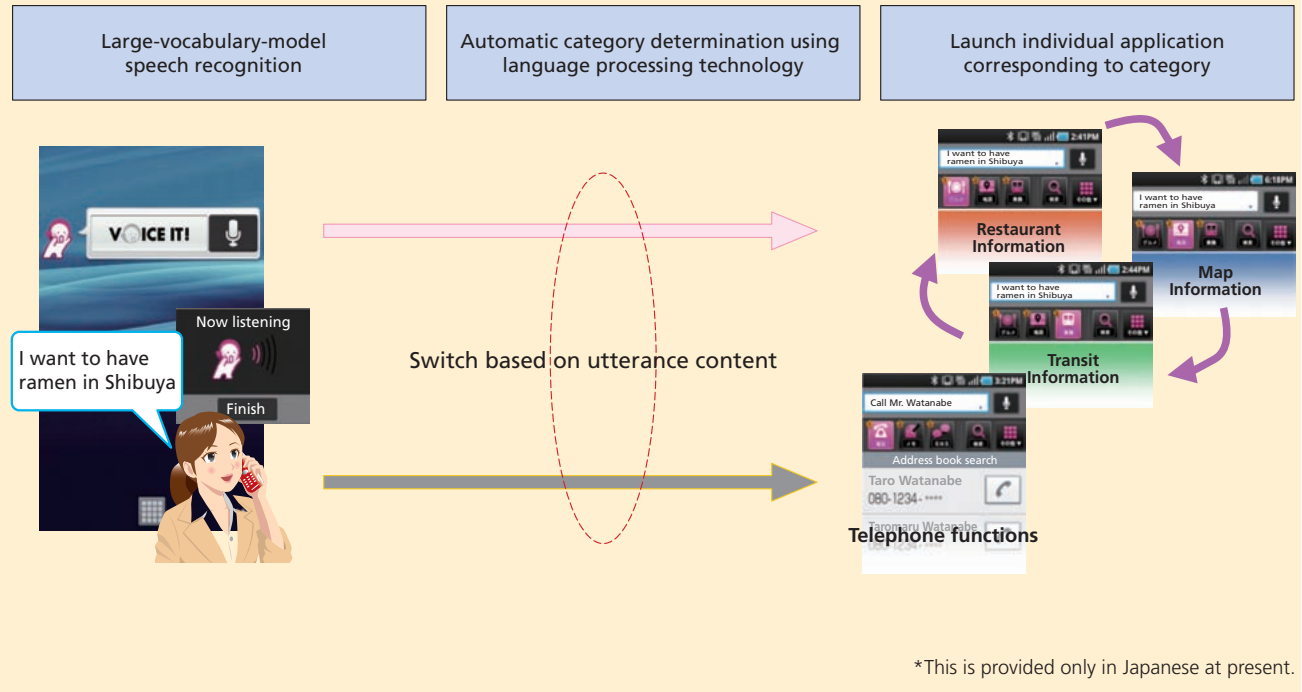
Ordinary Web search interfaces provide access to all kinds of information and services on the Internet, but the user must search through a list of results for the desired result. This can be a burden on operation when using devices that have a relatively small screen, such as mobile terminals.

Thus, in developing VOICE IT!, we used server-based speech recognition with a large vocabulary and incorporated the following into the design of the application.

- Enable calling of specialized applications for each category of terminal function or Web service.
- Use language processing technology and a ranking formula to automatically decide which category the user's utterance belongs to, and then suggest an appropriate application.
- Incorporate a screen that provides easy access to other applications in other categories that may be related to the utterance.

These measures enable the user to simply say what they want to do or know, and they can quickly access the desired application, without having to search for the terminal function or Web service they need to launch.

Figure 4 shows screen transitions for the VOICE IT! application. For example, if the user says "I want to eat ramen in Shibuya," information from a restaurant service about ramen shops in the Shibuya area is presented. Then, if users need a map of the Shibuya area or



*This is provided only in Japanese at present.

Figure 4 VOICE IT! application screen transitions

directions to Shibuya, they can select a corresponding icon to switch to that information. In the prototype, individual applications in the genres in **Table 1** were provided.

In the future we will work to implement user interfaces that are even more usable, based on the knowledge gained from the VOICE IT! trial.

4. Conclusion

In this article, we have described speech recognition technology and application development toward implementing a user interface that can respond to anything spoken by the user. In the future, we will improve the performance of speech recognition according to the service being provided, and study language processing technology to enable utterances converted to text to be understood more flexibly and at a higher level.

Table 1 Major specifications

Category	Major specifications
Web service	Web search
	Restaurant
	Travel
	Recipe
	Video
	Music
	Book
	Shopping
	Application
	Blog
	Transit
	Maps
	Weather
	News article search
	Dictionary
	Encyclopedia
	Q&A search
Terminal function	Telephone
	e-mail
	Memo
	Camera

REFERENCE

[1] H. Furukawa et al.: "Application Functions for Winter/Spring 2010-2011 Models —Evolving Mobile Terminal Applications —," NTT DOCOMO Technical Journal, Vol. 12, No. 4, pp. 15-23. Mar. 2011.