# Toward Intelligent Spoken Language Interface Technology

*Hirohito Inagaki, Takaaki Hasegawa, Satoshi Takahashi, and Yoshihiro Matsuo*

**Abstract**

The spoken language interface has become much more prominent in recent years, as services for smartphones that take speech input for search functions and provide answers in the form of synthesized speech have grown in popularity. This article describes work being done at the NTT Media Intelligence Laboratories aimed towards implementing intelligent spoken language interface technology to support a variety of services.

## 1. Introduction

People are most accustomed to speaking their thoughts and intentions. With the rising popularity of applications and services for smartphones in recent years, various services that use speech as an interface for operating smartphones have been developed to supplement manual operation. Web search services in which speech can be used to input keywords to the Google search service or to smartphones that use the Google Android OS (operating system) are now available. Apple's iPhone is equipped with an interactive speech agent known as Siri. The interactive speech agent Shabette-Concier* provided by NTT DOCOMO for smartphones also has many users. Cell phones and smartphones are mobile devices that do not have keyboards, so it is not easy to input text quickly. Using a speech interface rather than a soft keyboard, touch panel, or other such manual input device also creates new value by making it possible to obtain various kinds of information from network databases while moving around outdoors or while engaged in indoor activities such as cooking that require the use of both hands. The naturalness and convenience of a speech interface is appealing because it is similar to the way people interact with each other. Nevertheless, easy access to networks and to various kinds of data and knowledge requires more than simple speech processing; it now also requires new knowledge processing, natural language processing, and other back-end developments from Internet search technology. Furthermore, because a speech interface does not provide a visual medium, it is not suited to the immediate exchange of a large amount of information in the way that the screen of a personal computer is suitable for. Thus, accurately grasping what information the user is requesting, and selecting information on the basis of that understood request is even more important for a speech interface than when the information is output to an ordinary Internet browser on the screen of a personal computer.

The NTT laboratories are working to implement a personalized user interface and user experience (UI/UX) that are simpler and easier to use for the development of advanced services by linking information processing functions for various kinds of existing databases and Internet information in addition to developing front-end functions for highly accurate processing of speech input and output.

This article explains the most recent research on the intelligent spoken language interface technology that is needed to support the various increasingly advanced services of the future.

* "Shabette-Concier" is a registered trademark of the NTT DOCOMO Corporation.

## 2. Evolution of the intelligent spoken language interface

We begin by looking back on the history of dialogue systems, focusing particularly on their configuration, and then we consider the directions of intelligent spoken language interface development and the conditions it requires. Research on technology for the interaction of humans and computers has a long history. About half a century ago, the interactive system consisting of typing on a keyboard to communicate with a computer appeared. That was followed by research and development on speech interaction systems that understand spoken words. Around 2000 in the U.S., there was a boom in the development of practical voice portals and interactive voice response (IVR) systems, in which speech interaction processing was performed on a remote server. Telephones have now become smartphones, and speech interaction with computers anytime and anywhere has become possible through both wired and wireless connections. Furthermore, dialogue system servers can be accessed via the Internet, enabling the acquisition of knowledge from the immense collection of documents that are available on the Internet. That has made it possible to answer even broad questions for which it is difficult to prepare responses in advance. Such documents are continuously being added, and using them as a source of knowledge enables real-time handling of information in finer detail.

The continuing development of the spoken language interface must include further development of knowledge processing and natural language processing (the back-end processing) as well as development of speech recognition and synthesis processing that is suited to the user's context (the front-end processing).

We can view the spoken language interface as having two directions:

(1) Human-agent spoken language interface (second-person interface)

This is an extension of current interactive systems in which an agent supports user thought and behavior by anticipating the user's intention on the basis of the user's present situation and behavioral history, gathering the precise information from the large amount of various kinds of data that are stored on the Internet, and composing an answer that is as complete as necessary.

(2) Human-human spoken language interface (third-person interface)

This is an agent that can invigorate communication between humans, whether face-to-face or over the Internet, by autonomously gathering information that is relevant to the topic of conversation and information from the communication history and profiles of the participants, and spontaneously offering useful information.

Both of these directions require personalization that involves strong awareness of the user at the front end. The NTT laboratories are researching spoken language processing as part of their efforts to realize such an intelligent spoken language interface.

## 3. Elemental technology for constructing an intelligent spoken language interface

The elemental technology needed to configure an intelligent spoken language interface is illustrated in **Fig. 1**. First, the user's speech is input to a speech recognizer and recognized on the basis of an acoustic model and a language model. The result is then converted to text. Next, the converted input is sent to a problem solver, which performs some processing and generates a response according to the result. This special issue specifically concerns the processing for generating an answer in response to a question and the processing for extracting the knowledge that is required to form an answer from a large collection of documents. These types of processing are positioned as applied natural language processing technology. In the final step of the intelligent spoken language interface, a speech synthesizer converts the textual output of the problem solver to speech using a speech database and prosody generation model.

The following sections describe the research on these technologies.

### 3.1 Speech recognition

The accuracy of speech recognition is extremely important in the intelligent spoken language interface because it strongly affects the overall performance of the system. We have therefore attempted to increase the accuracy of speech recognition in various ways. One approach is to achieve high recognition accuracy for unspecified speakers rather than particular individuals. The accuracy of speech recognition for particular users can be low, so it is necessary to improve recognition so that it is accurate for any user.

Another approach is to achieve good recognition accuracy when background noise is present. Use environments are diverse, and ambient noise is particularly problematic for mobile devices. At a train station platform or bus stop, for example, the sound

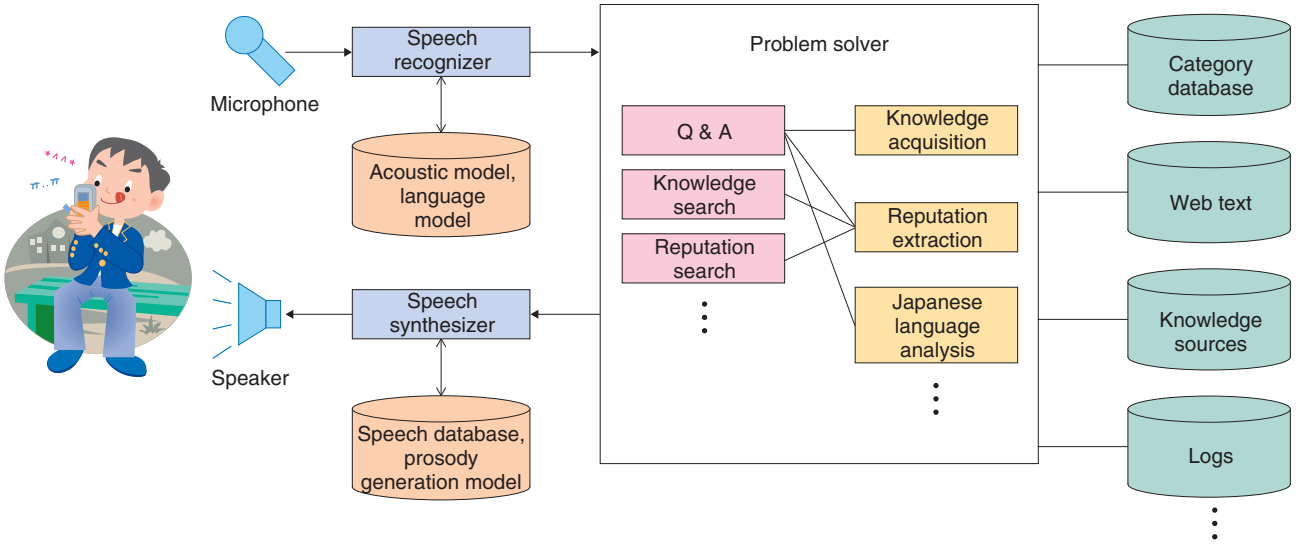


Fig. 1. Overview of an intelligent spoken language interface.

of passing trains or vehicles is present, and in restaurants, coffee shops, or other such locations where people gather, there is the noise of other people talking. There is thus a need for accurate recognition of user speech even under such noisy conditions.

A third approach to accurate speech recognition is the ability to recognize new words that enter the lexicon day by day. Generally, words that are not in dictionaries (referred to as out of vocabulary (OOV)) are a problem in speech recognition. The accuracy of a service can decrease greatly when OOV words cannot be recognized. It is therefore necessary to look for new words that should be recognized in past utterances of the user and other places.

New speech recognition technology for the three approaches described above is explained in detail in the article entitled "Speech Recognition Technology That Can Adapt to Changes in Service and Environment" [1] in these Feature Articles.

### 3.2 Knowledge extraction

The knowledge extraction performed by the problem solver extracts the information required to accurately answer a wide variety of user questions from a large set of Web pages. For example, if a user asks a question for which the answer is the name of a person or a mountain, the solver collects the names of persons and mountains from a relevant set of documents. Objects that have specific names to be referred to are called named entities (e.g., the mountain known as *Everest* is a named entity.). Examples of named entities include places, organizations, time expressions, numerical expressions, and monetary amounts. In implementing an intelligent spoken language interface, *knowledge* is required in order to *understand* the meaning of user utterances and to generate responses. Named entities are an important element in constructing such knowledge. For example, named entities can be used to construct a particular individual's information. For example, if it is possible to extract the name of a certain person, then attributes such as that person's date of birth (time expression) and place of birth (place name) could be included in the response. If it were also possible to extract the relations among named entities, then correct answers to even more complex questions would be possible. Technology for extracting named entities and the relations among them from a large set of Web pages is described in the article entitled "Knowledge Extraction from Text for Intelligent Responses" [2].

### 3.3 Question answering

In the question answering process performed by the problem solver, the user's question is first analyzed to determine what the user is asking and what type of question it is (who, what, where, or when) in order to determine what specific kind of named entity the question concerns. For example, is the user asking for a person's name, the name of a mountain, or the name of a food, etc.? Then, various calculations are performed to select the most suitable candidate for the content of the question from among the named

entities extracted from Web pages in order to form the answer. The major difference between question answering and document retrieval is that the response to question answering is a pinpointed answer rather than a list of documents, so very high accuracy is expected in the answer. The question answering technology, which is the heart of the intelligent spoken language interface, is explained in the article entitled "Question Answering Technology for Pinpointing Answers to a Wide Range of Questions" [3].

### 3.4 Speech synthesis

Speech synthesis technology generates speech with a synthetic voice of good quality by reading text correctly and with the appropriate intonation and pauses. We are also researching various aspects of speech synthesis for implementing an intelligent spoken language interface. In one line of research, the user is allowed to freely select the voice of a person they would like to hear, such as a family member or friend. Being able to easily synthesize voices in that way could lead to the implementation of services that can be adapted to individuals. Another area of investigation involves the generation of clearly articulated speech that can be heard clearly even in the presence of background noise in crowded user surroundings. Speech synthesis technology that implements these features is explained in the article entitled "Speech Synthesis Technology to Produce Diverse and Expressive Speech" [4] in these Feature Articles.

## 4. Future work

We have presented an overview of speech recognition, speech synthesis, and natural language processing (knowledge extraction and question answering), all of which constitute the intelligent spoken language interface. Further progress requires audio signal processing for advanced sound acquisition and reproduction technology for mobile devices in addition to development of the technologies we have described. We will continue with the research and development of audio, speech, and language media technology that integrates these technologies with the objective of realizing a personalized UI/UX.

## References

[1] H. Masataki, T. Asami, S. Yamahata, and M. Fujimoto, "Speech Recognition Technology That Can Adapt to Changes in Service and Environment," NTT Technical Review, Vol. 11, No. 7, 2013.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa2.html

[2] K. Sadamitsu, R. Higashinaka, T. Hirano, and T. Izumi, "Knowledge Extraction from Text for Intelligent Responses," NTT Technical Review, Vol. 11, No. 7, 2013.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa3.html

[3] R. Higashinaka, K. Sadamitsu, K. Saito, and N. Kobayashi, "Question Answering Technology for Pinpointing Answers to a Wide Range of Questions," NTT Technical Review, Vol. 11, No. 7, 2013.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa4.html

[4] H. Mizuno, H. Nakajima, Y. Ijima, H. Kamiyama, and H. Muto, "Speech Synthesis Technology to Produce Diverse and Expressive Speech," NTT Technical Review, Vol. 11, No. 7, 2013.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa5.html

**Hirohito Inagaki**
Vice President, Director, NTT Media Intelligence Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Keio University, Kanagawa, in 1984 and 1986, respectively. Since joining NTT Electrical Communication Laboratories in 1986, he has been engaged in R&D of natural language processing and its applications. From 1994 through 1997, he was on transfer to NTT Intelligent Technology Co., Ltd., where he was engaged in developing information security systems and multimedia systems. He moved to NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2006. His research interest is R&D of audio, video, and language interfaces and their applications. He is a member of the Institute of Electronics, Information and Communications Engineers (IEICE).

**Takaaki Hasegawa**
Senior Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E. and M.E. degrees from Keio University, Kanagawa, in 1992 and 1994, respectively, and the Dr.Eng degree from Tokyo Institute of Technology in 2010. Since joining NTT in 1994, he has been engaged in the research of natural language processing and intelligent information access. He was a visiting researcher at New York University from 2003 to 2004. He is a member of the Information Processing Society of Japan (IPSJ), the Japanese Society for Artificial Intelligence, and the Association for Natural Language Processing (NLP).

**Satoshi Takahashi**
Executive Manager, Executive Research Engineer, Supervisor, Audio, Speech and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E., M.E., and Ph.D. degrees in information science from Waseda University, Tokyo, in 1987, 1989, and 2002, respectively. Since joining NTT in 1989, he has been engaged in speech recognition, spoken dialog systems, and pattern recognition. He is a member of the

**Yoshihiro Matsuo**
Group Leader, Senior Research Engineer, Supervisor, Audio, Speech and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.S. and M.S. degrees in physics from Osaka University in 1988 and 1990, respectively. He joined NTT Communications and Information Processing Laboratories in 1990. He moved to NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2001. His research interests include multimedia indexing, information extraction, and opinion analysis. He is a member of IPSJ and NLP.