Cyber Attendant System with Spontaneous Speech Interface

Jun-ichi Hirasawa†, Tetsuo Amakasu, Shunichiro Yamamoto, Yoshikazu Yamaguchi, and Akihiro Imamura

Abstract

This article introduces technologies needed for spoken dialog services that can process the spontaneous utterances heard in everyday life. Up to now, most speech services have been based upon isolated word speech recognition. However, we need to be able to process common spontaneous utterances including fillers and hesitations to achieve really usable interfaces. Our new speech recognition technology, VoiceRex2003, can handle such utterances. We applied it to a cyber attendant system, CASYS2003, which offers multi-modal services.

1. Introduction

A common dream is to interact with computers in the same way as our friends. Science fiction novels and movies often contain machines that can communicate with humans via natural, everyday, spoken dialog. They accept our questions and requests, provide information, and do as they are told. We guess that the reason people would like to communicate with computers via spoken dialog is that speech is the easiest and most natural form of communication for us. Using the keyboard to input messages is too slow and laborious; speech sets us free. Thus, if we want to solicit information from a computer and get the desired result, natural speech input is the most appropriate method, especially for novice users.

However, the current reality is far from this dream. While it is true that there are some devices that provide speech interfaces, for example, speech-enabled car navigation systems or personal game terminals, their usability and performance leave a lot to be desired. This is because conventional speech interfaces impose unusual speech styles on the user. They are far from intuitive, so many users become confused because they have no idea how to use the interface. With more experience, the user learns that conventional speech-enabled systems cannot accept everyday spontaneous speech: the systems demand unusual utterance styles and are not flexible enough to handle the speech fillers and hesitations common in everyday speech. This is a long way from the goal of eliminating the need for special training.

In this article, we introduce our new speech-recognition technology, VoiceRex2003, which can accept spontaneous utterances. VoiceRex2003 is extremely powerful and flexible. It handles real-world speech that includes fillers and hesitations and can control dialog initiatives to better support the user. We have applied VoiceRex2003 to a multi-modal cyber attendant system CASYS2003, which will lead to many service applications, not only telephony-based systems, but also IP-based multi-modal speech services.

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: hirasawa.j@lab.ntt.co.jp

2. VoiceRex2003: spontaneous speech interface

2.1 Problems and challenges

Good speech interfaces should not require us to use rigid and unfamiliar speech styles. A spoken dialog system must accept everyday spontaneous utterances. What then is the difficulty in processing spontaneous utterances? First of all, spontaneous utterances include speech fillers and hesitations. Since it seems impossible for most speakers to create perfect speech in real time, we are often forced to use fillers such as

"um" and hesitations. Next, every spoken language provides great flexibility in terms of the expressions used. An idea can be expressed in many different ways. Finally, listeners often counter the speaker's question with another. Moreover, the listener's responses often overlap the speaker's utterances. These phenomena are not handled well by existing systems. Our goals include handling the following problems:

(i) various linguistic expressions, including speech fillers, hesitations, and alternate expressions
(ii) variation in dialog progression, including deviation from the system's prompt
(iii) various utterance timings, including the user's utterance overlapping the system's prompt.

NTT Cyber Space Laboratories has been developing a series of speech recognition engines under the name "VoiceRex". This article describes the latest version, "VoiceRex2003" and shows how it overcomes the constraints that cause conventional speech interfaces to fail.

2.2 Stochastic language modeling

Speech recognition requires the use of linguistic constraints to determine whether word strings form utterances. We lump these constraints together under the term "grammar" or "language model". In conventional methods, the developers of a recognition system must manually define the "grammar", i.e., the word-connection rules, for the system, which makes new services expensive to develop. This is especially true for spontaneous utterances. It is almost impossible to draw up an adequate set of manual grammar rules that can handle all possible linguistic expressions.

Our solution is to use stochastic language modeling. This approach is called class N-gram modeling. A large sample size of utterances, called a corpus, yields adequate stochastic connectivity probabilities for language modeling, allowing us to cover a broad range of spontaneous expressions and thus achieve truly effective speech recognition.

2.3 Preparing language models

Since we use class N-gram language modeling, we dispense with the effort of manually forming grammars when creating new services. However, class N-gram language modeling requires an utterance corpus containing many samples, which is usually considered to be difficult to create. To solve this problem, VoiceRex2003 provides a convenient language model generation tool. This tool has a graphical user interface (GUI) and currently starts with 100,000 sample utterances as a spontaneous expression corpus. This built-in corpus establishes a good initial level of language modeling.

Although other speech recognition engines could also be provided with stochastic language modeling technology, it is not clear whether they could provide a less cumbersome environment. Since VoiceRex2003 provides a generation tool with a built-in corpus for stochastic language modeling, it can handle various linguistic expressions, including speech fillers, hesitations, and alternate expressions, more effectively.

2.4 Speech understanding

Most conventional speech recognition systems are based on isolated words or short fixed phrasal forms rather than on spontaneous utterances. In spontaneous speech recognition, however, recognition results could contain words that hinder speech understanding. Therefore, such words must be removed and only important keywords must be extracted from the recognition result. This process is called "speech understanding". **Figure 1** describes the information extraction process from input speech to speech understanding result representation. The results are represented as domain slot entries.

2.5 Flexible dialog control

Conventional speech dialog interfaces restrict possible utterances within each phase of the dialog. For example, the system prompt "Please say the station that you want to go to" tries to restrict the user's response to just a station name. The system prompt "Do you want to go to Yokohama station? Please, answer yes or no" aims to narrow the user's response to just yes or no. Thus, if the dialog system misrecognizes the user's response, it takes the user a long time to correct the error because the dialog imposes restrictions on each phase of the correction process. The left-hand side of **Fig. 2** shows a sample dialog of a conventional system.

VoiceRex2003, on the other hand, can accept the user's input utterance even if it does not conform to the system's prompt. For example, if the system misrecognizes the user's response and creates the prompt "Do you want to go to Yokosuka station?", the user can respond with "No, I didn't say Yokosuka. I said Yokohama", where the system's error is directly corrected, as in the right-hand side of Fig. 2. Conventional technologies, e.g., voiceXML dialog scenario interpreter, cannot accept direct correction, because

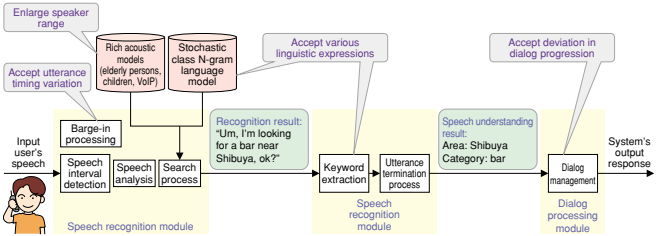


Fig. 1. VoiceRex2003 handles spontaneous utterances.

System:	Please say the station that you want to go to.	System:	Which station do you want to go to?
User:	Yokohama.	User:	Yokohama.
System:	Do you want to go to Yokosuka station? Please, answer yes or no.	System:	Do you want to go to Yokosuka station?
User:	No.	User:	No, I didn't say Yokosuka. I said Yokohama.
System:	Then, please say the correct station name.		
User:	Yokohama.		
System:	Do you want to go to Yokohama station? Please, answer yes or no.	System:	Sorry, do you want to go to Yokohama station?
User:	Yes.	User:	Yes, that's right.
System:	All right, I'll begin to find the way to Yokohama station…	System:	All right, I'll begin to find the way to Yokohama station…
	(a) Rigid dialog control		(b) Flexible dialog control

Fig. 2. Dialog samples of spontaneous utterances.

such technologies cannot change that predetermined slot value. Dialog management in VoiceRex2003 is controlled by a different method. The types of changes in a slot value control the progression of a dialog. This mechanism makes the system feel as if it is working with the user.

2.6 Handling a wide range of speakers and barging-in

Conventional systems have several constraints on the user's utterances. First, they often fail to recognize atypical user utterances, such as those made by elderly people and children. VoiceRex2003 can cope with these voices because it has rich acoustic models and can use them appropriately. Moreover, VoiceRex2003 suits the VoIP network. Second, conventional systems restrict the timing of the user's utterance so that they can detect when the user's speech begins. A typical system prompt is "Please speak after the tone". In everyday conversations, you can start talking at any time except when that is impolite. VoiceRex2003 can accept the user's utterance at any time, even during its own utterance because the speech recognition module is running all the time. This action is known as "barging-in".

2.7 Application to telephony

The technologies described so far enable the system to process spontaneous speech as part of the services implemented on CTI (computer telephony integration) platforms. One example is the commercially available IVR (interactive voice response) platform "Advice" produced by NTT-IT Corporation.

3. CASYS2003: cyber attendant system with speech interface

Most speech services have been created for telephony-based services. Given the rapid acceptance of the IP network, however, speech recognition services should be designed to support IP-based applications too. Such applications use personal computers or personal digital assistants, which provide a visual display in addition to speech. IP-based applications allow users to input their commands to the system via a keyboard and mouse not just by speech. It is obvious that such systems can output information to the users via figures or tables on the displays, or can output the system's response via humanoid animated agents that speak. This is usually called a "multi-modal" service. We have developed a multi-modal cyber attendant system, CASYS2003 (**Fig. 3**).

3.1 CASYS2003: core technologies

CASYS2003 supports common web browsers such as Internet Explorer. It allows humanoid agents to implement multi-modal services. This requires CASYS2003 to process several different modal input/output signals in an integrated manner. To do this, it uses the Active X controls provided by Internet Explorer. In CASYS2003, the dialog scenario interpreter plays a central role. When implementing a service, the interpreter first downloads a document describing a dialog scenario written in CASYS-ML (CASYS dialog scenario mark-up language) and starts the dialog scenario interpretation processor and the speech recognition engine for spontaneous utterances, VoiceRex2003. It can also simultaneously accept inputs via mouse click and keyboard entry besides speech, so it offers multiple input modalities.

For output, the CASYS scenario interpreter can control the speech output, text-to-speech engine, and pre-recorded audio files. It can also provide graphical depictions using HTML. The animated character agent is based on MS (Microsoft) Agent technology. This achieves multiple output modalities.

An XML document written in CASYS-ML can describe dialog scenarios and control the CASYS scenario interpreter, so services developers are not required to write system-level programs to control speech devices or image processing.

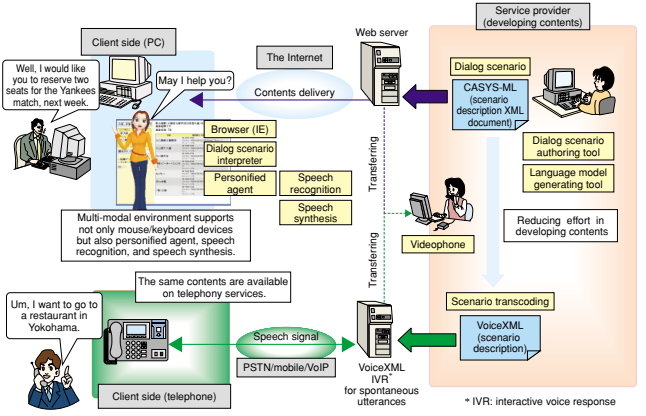


Fig. 3. Cyber Attendant System CASYS2003 for multi-modal dialogs.

3.2 Environment that supports service development

As shown in Fig. 3, designing and coding a dialog scenario document is a complex task, so CASYS2003 offers a support environment for service development. During dialog scenario design, the CASYS2003 developmental environment provides GUI-based dialog scenario authoring tools (dialog scenario authoring tool and language model generating tool in Fig. 3). This greatly reduces the effort needed to develop a scenario.

CASYS2003 also has a unique authoring tool environment for transcoding a dialog scenario into a telephony service. It can automatically transform a dialog scenario written in CASYS-ML into a document in VoiceXML format (scenario transcoding in Fig. 3). When a contents provider wants to offer the same service in a multi-modal environment and in a telephony environment, the transcoder in CASYS2003 reduces the costs incurred.

CASYS2003 has one more feature. Since it runs within the most common web browser, it is easy to apply CASYS2003 to familiar web-technologies. For example, the CASYS scenario interpreter can connect the user to a human operator waiting in a contact center, because it is easy to apply CASYS to an existing videophone system (videophone in Fig. 3). This lets service providers develop services in which CASYS2003 initially accepts a simple order from the user and routes it to an appropriate human operator depending on the user's request.

3.3 Examples of services

CASYS2003 can be applied to various information services. The most obvious application field covers service reservation service tasks. A CASYS2003 multi-modal system can provide user-friendly ticket reservation, dental appointments, and other kinds of reservation services that are easy for novices to use.

Information retrieval services are another good application area for CASYS2003. A train route planning service, shop finding service, and other kinds of information retrieval services would free the user from annoying keyboard input, by using spontaneous speech to make service input very smooth. **Figure 4** shows an example of a shop finding service. Service users receive a list of appropriate shops and information after they fill out the desired shopping area and category fields.

Since spontaneous speech is also suitable for con-

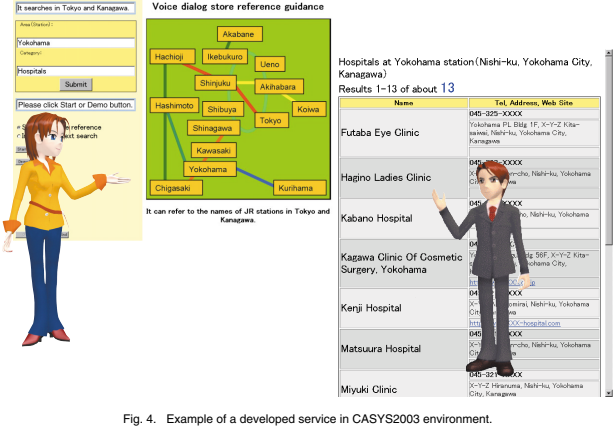


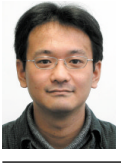
Fig. 4. Example of a developed service in CASYS2003 environment.

sultation, some kinds of preference-based services tasks could be achieved if sufficiently powerful intellectual knowledge processing can be achieved. Finally, it is reported that most of the queries that contact centers receive are very simple ones. A cyber attendant system that can handle such queries would allow the human operators to devote themselves to more difficult tasks.

4. Conclusion

This article introduced our new speech recognition technology, VoiceRex2003, which can accept spontaneous utterances from the user. We also described the cyber attendant system, CASYS2003, which allows us to communicate with cyber attendants in multi-modal environments. VoiceRex2003 and CASYS2003 will contribute to natural and effortless communication with computers via spoken dialog. We hope that they will be applied in real-world services soon. To encourage this trend, we are planning to continue development in two directions: using distant-microphone technology and utilizing personal dialog histories.

Jun-ichi Hirasawa
Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.
He received the B.S. degree in behavioral science from Hokkaido University, Sapporo, Hokkaido in 1993 and the M.S. degree in information science from Nara Institute of Science and Technology, Nara in 1995. Since joining NTT in 1995, he has been working on spoken dialog systems. He is a member of the Information Processing Society of Japan and the Acoustical Society of Japan (ASJ). He is currently on the editorial board of the International Journal of Speech Technology.



Tetsuo Amakasu
Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.
He graduated from the Graduate School of Information Sciences of Tohoku University, Sendai, Miyagi in 1999. Since joining NTT in 1999, he has been working on R&D of spoken dialog systems. He is a member of ASJ and the Japanese Society for Artificial Intelligence.

Shunichiro Yamamoto
Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.
He received the B.E. degree in information systems engineering from Osaka University, Osaka in 1997 and the M.E. degree in information science from Nara Institute of Science and Technology, Nara in 1999. In 1999, he joined NTT Cyber Space Laboratories, Yokosuka, Japan. He is a member of ASJ.

Yoshikazu Yamaguchi
Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.
He received the B.E. and M.E. degrees in electrical engineering from Osaka Prefecture University, Osaka in 1993 and 1995, respectively. Since joining NTT Human Interface Laboratories, Yokosuka, Japan, in 1995, he has been working on R&D of speech recognition technologies. He is a member of ASJ.

Akihiro Imamura
Senior Research Engineer, Supervisor, Media Processing Project, NTT Cyber Space Laboratories.
He received the B.E. and M.E. degrees in electrical engineering from Nagaoka University of Technology, Niigata in 1982 and 1984, respectively. In 1984, he joined Nippon Telegraph and Telephone Public Corporation (now NTT). He is a member of ASJ, IEEE, and the Institute of Electronics, Information and Communication Engineers of Japan.