Technologies for Voice Portal Platform

● Yasushi Yamazaki ● Hitoshi Iwamida ● Kazuhiro Watanabe

(Manuscript received November 28, 2003)

The voice user interface is an important tool for realizing natural, easy-to-use human-computer interfaces that can bridge the digital-divide. This paper gives an overview of voice interface applications and information access systems that operate over a telephone network. It also introduces some voice portal solutions that use speech recognition/synthesis and dialogue control and outlines the speech processing technologies used in voice portal solutions. The developers of a voice portal system must understand the characteristics of speech and ensure that the speech recognition part can cope with interjections, background noise, and the distortions caused by telephone lines. Speech synthesis produces natural speech and other speech styles by using a corpus-based method and automatically constructing a waveform database. The information services of municipal offices, traffic information services, stock price services, and so on are widely expected to incorporate voice portal systems in the near future.

1. Introduction

In the near future, human-computer voice interfaces will become important tools for solving the accessibility limitations of conventional human-computer interfaces such as the keyboard, mouse, and GUI.

Progress in research and development of speech recognition/synthesis technologies and improvements of microprocessors have brought many commercial products in fields such as personal computers, mobile phones, car navigators, and CTI (Computer Telephony Integration).[1)] Voice dialing and e-mail reading using mobile phones have been realized, and voice commands and voice navigation can be used in car navigators.

This paper gives an overview of speech recognition/synthesis applications, particularly information access systems that operate over a telephone network. It also describes voice portal solutions that apply speech technologies and dialogue control and the speech processing technologies used in voice portal solutions.

2. Voice-operated information access

Most information systems are still operated using the dial tones of a standard telephone. However, much work is now underway to develop voice-operated information access systems. Fujitsu has developed a voice and fax response system called VoiceScript[2)] that uses speech synthesis technology. This system is being applied in various solutions as a major information access system.

Application systems with speech recognition were first introduced in the middle 1990s, and with the big boom in mobile phones, there has been a rapid growth in voice portal service trials and implementations. The global market for voice portal services is estimated to become $12.3 billion in 2005.

Many interactive scenarios used in voice por-

tals have been created by custom design. A voice interactive control standard called VoiceXML (Voice Extensible Markup Language)[3] has recently become available. In addition, SALT (Speech Application Language Tags) for multi-modal operation has been released. Voice interactive systems that conform to the standard specifications will be hot items in the market in future years.

3. Voice portal solutions

We have developed a trial system for a voice portal solution using our Japanese speech recognition/synthesis and dialogue control technologies. This system copes with interjections and background noise by using an enhanced word spotting technique. On the speech synthesis side, it produces highly natural speech and various speech styles by using a corpus-based method and automatically constructing a waveform database. Although this paper describes a voice portal system for Japanese, the basic ideas used in these technologies are language independent.

3.1 System configuration

Figure 1 shows the configuration of the trial voice portal system. A voice interactive scenario is prepared using an existing Web system accessed from a personal computer or other device. The information required by the user is obtained by operating the speech recognition/synthesis engine using instructions from the dialogue control engine. Because an existing telephone network is used, users can easily access the information over the telephone at anytime and from anywhere.

3.2 Features

This system was designed to not only replace the existing system with voice but also to consider the characteristics of voice. The system has the following features.

1) Speech recognition engine

The system uses an enhanced word spotting technique that can cope with interjections and can counteract background noise and the distortions caused by telephone lines. As a result, high-performance telephone speech recognition was realized.

2) Speech synthesis engine

The speech synthesis engine uses corpus-based speech synthesis, which generates highly natural synthesized speech using a large volume of speech data (called the corpus), and automatic speech corpus generation, which generates speech corpora automatically from various samples of human speech.

The speech synthesis engine has already been used in many applications in Japan.

3) Dialogue control engine

We developed an interpreter that conforms to the VoiceXML2.0 standard for dialogue control. This interpreter enables us to realize interactive scenarios appropriate for various user conversation modes, for example, system-initiated conversation and user-initiated conversation.

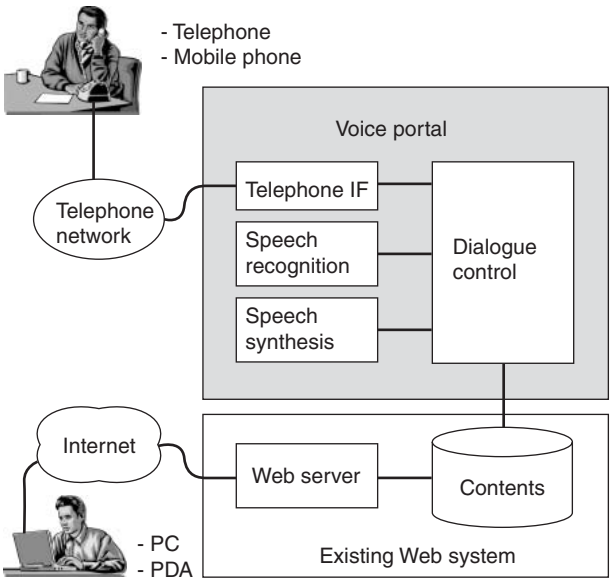

Figure 1
Trial voice portal system.

3.3 Example applications

We expect this system will have a wide range of applications, for example, the information services of municipal offices, stock price services,

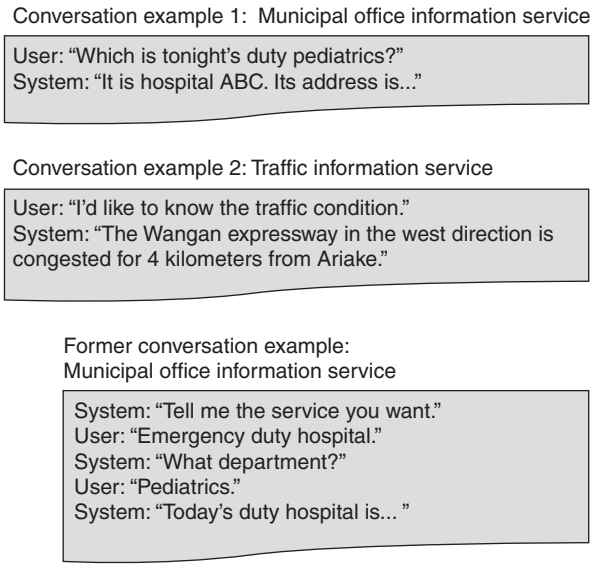

Figure 2
Example dialog of voice portal applications.

traffic information, and other real-time information services. In particular, it will help municipal offices realize natural, easy-to-use human-computer interfaces that can bridge the digital-divide. **Figure 2** shows some example conversations that have taken place in the voice portal service. Until now, most conversations have been in a simple question-and-answer mode. However, as can be seen in the examples, this system allows users to speak more naturally and include multiple information items in the same speech string.

4. Speech recognition

This section describes the high-performance telephone speech recognition technology of this system.

4.1 System configuration

Figure 3 shows the general configuration of the speech recognition system. After acoustic processing such as speech detection and noise suppression of the input speech, the features of the speech are extracted. Based on the extracted features, matching is performed using an acoustic model and a linguistic model to obtain the recognition result. The acoustic model is a model of the phonemes (e.g., the vowels and consonants) of the input speech. The Hidden Markov Model (HMM)[4] acoustic model is generally used. The linguistic model uses a word dictionary describing the vocabulary to be recognized. It also uses a network grammar that describes the connections between words using a finite-state automaton or a statistical grammar such as an N-gram grammar that describes word connection probabilities.

4.2 Telephone quality speech recognition

For speech recognition in a voice portal, it is essential to counter distortions caused by telephone lines. It is also essential to consider how different first-time users will use the system. Fujitsu therefore developed an acoustic model using a large volume of speech data of telephone-line quality. This model has been extensively evaluated and improved in various environments to intensify its robustness. In particular, we developed various processing techniques for speech detection and noise suppression. These include techniques for reducing the influence of distortion by speech coding and measures to counter background noise. Thanks to these developments, our system performs highly accurate recognition of telephone quality speech, even speech from a mobile phone.

4.3 Word spotting

In casual, spontaneous speech, it is difficult to get correct recognition results. For instance, spontaneous utterances include many interjections such as "well" and "uh-uh" and the mode of expression varies in many ways (e.g., "tell me" or "please tell me"). As a result, recognition errors arise and conversations cannot progress as intended. To achieve a higher recognition performance, Fujitsu developed a word spotting technique with grammatical restrictions that ignore interjections and other distractions.

Word spotting extracts only predetermined words in the word dictionary from input speech. **Figure 4** shows an example of word spotting with

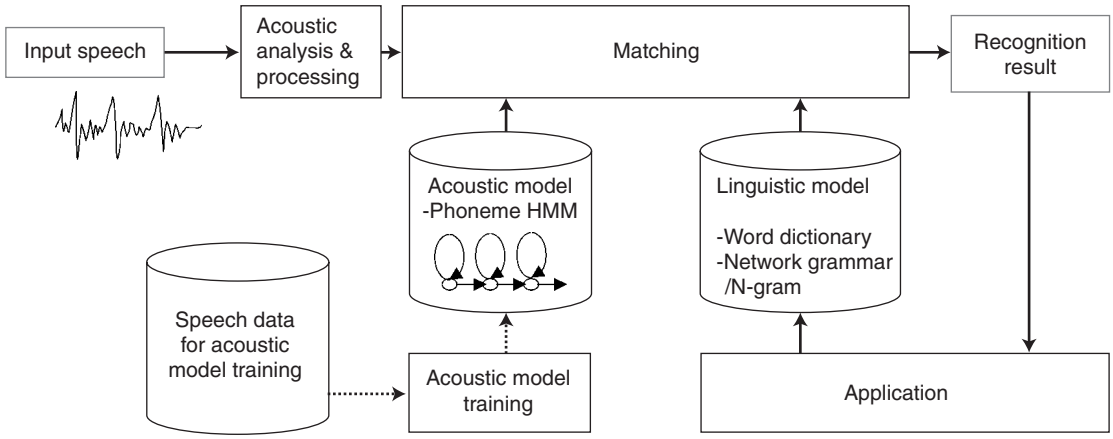

Figure 3
Speech recognition system.

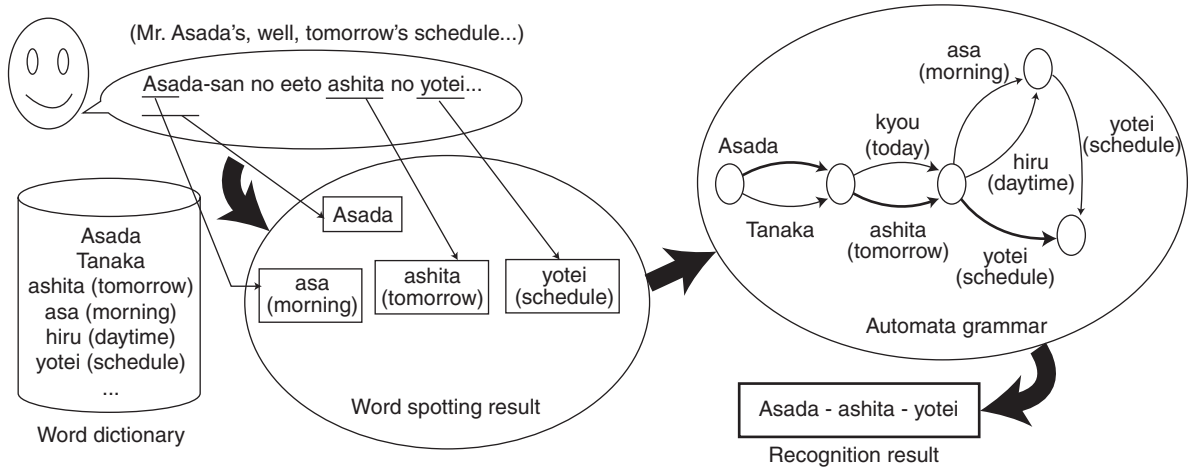

Figure 4
Example of word spotting with grammar.

grammatical restrictions. **Table 1** shows the meanings of the Romanized Japanese words used in the following sentences. When a user says, "Asada-san no eeto ashita no yotei...,"[*1] the four words, "asa,"[*2] "Asada," "ashita,"[*3] and "yotei"[*4] in the word dictionary are extracted from the input speech. When only word spotting is used, sufficient recognition accuracy cannot be ensured because similar sounding words might be erroneously extracted. In this example, "asa" is extracted from the input speech because "asa" is similar to "Asada." Therefore, another grammatical restriction is imposed. That is, only applicable word sequences are searched according to the predefined automata grammar. In this example, "asa" at the top of sentence is deleted by the grammatical restriction and "Asada – ashita – yotei" are selected as the result.

5. Speech synthesis

In this section, we describe the speech synthesis technology used in this system to realize natural voice responses and provide rich voice variations.

5.1 Essential points of technical development

Speech recognition/synthesis technology is being actively introduced to information and ticket reservation services in order to increase business

without increasing the number of operators. To meet the different requirements of different services, it is necessary to produce natural-sounding voice responses and rich voice variations. Our system produces these using corpus-based speech synthesis, which generates highly natural synthesized speech, and automatic speech corpus generation, which generates various speech corpora automatically from various samples of human speech. These two methods are described below.

5.2 Corpus-based speech synthesis

Our system synthesizes speech using a large volume of speech data (called the corpus) and a method usually referred to as the corpus-based speech synthesis method. The system consists of a linguistic processing module, prosody generating module, and waveform generating module (**Figure 5**). The synthesis method is used in the waveform generating module.

The linguistic processing and prosody generating modules convert the input text into a phoneme sequence and prosody (i.e., the duration of each phoneme, the pitch contour, and the amplitude pattern). The first process of the waveform generating module selects waveform units from the speech corpus. The longer waveform units that will result in less discontinuity for waveform concatenation are selected for the phoneme sequence and prosody. In the example of Figure 5, the following are selected for the text input "Yamanashi ken no JR Chuo sen..."[*5]: the phoneme sequence "y-a-m-a-n-a-sh-i" from sentence "Yamanashi no koukou kara...,"[*6] the phoneme sequence "k-e-N-n-o" from sentence "bekken no kaigi ga...,"[*7] etc. In this example, the "k-e-N" phoneme sequence in the text input means "prefecture" and the selected phoneme sequence "k-e-N" means "case."

Although these are different words, the difference in meaning is not a problem if these

Table 1
Meanings of the romanized Japanese words.

ID	Romanized Japanese	Meaning
*1	Asada-san no eeto ashita no yotei...	Mr.Asada's, well, tomorrow's schedule...
*2	asa	morning
*3	ashita	tomorrow
*4	yotei	schedule
*5	Yamanashi ken no JR Chuo sen...	JR Chuo line of Yamanashi prefecture...
*6	Yamanashi no koukou kara...	From a high school in Yamanashi...
*7	bekken no kaigi ga...	There is another meeting...

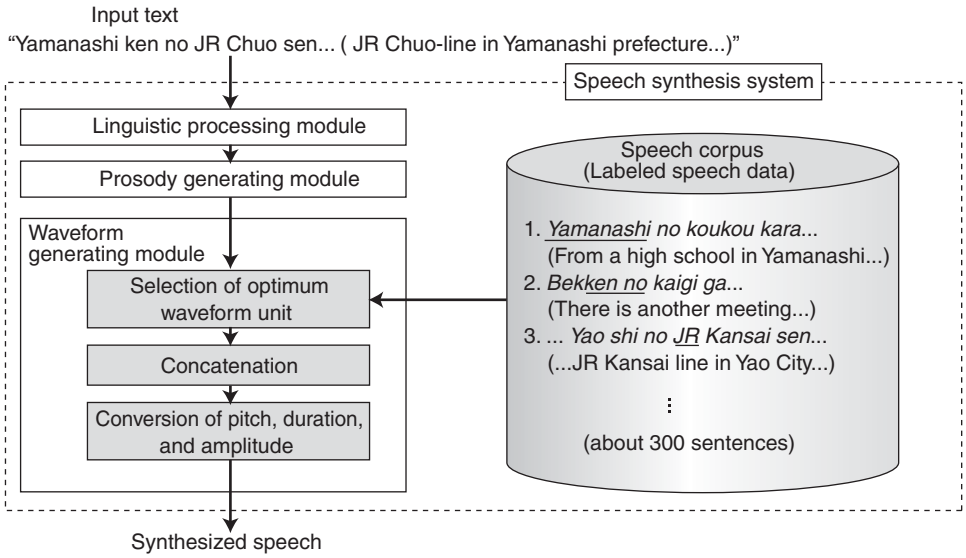

Figure 5
Outline of corpus-based speech synthesis.

phoneme sequences are the same. The speech corpus includes data of about 300 sentences and 180 syllables (a,i,u...) of Japanese speech. In the second process of this module, the selected waveform units are concatenated and their pitch, duration, and amplitude are transformed according to the input text.

Compared to the former method[5] of concatenating fixed and short phoneme-waveform units, this method allows concatenation of longer and variable length phoneme-sequence-waveform units. Therefore, the number of concatenation points, which cause deterioration of synthesized speech quality, can be remarkably reduced. In addition, the use of a longer waveform that matches the input phoneme sequence results in good preservation of the original voice quality and improvement of the synthesized speech quality.

Figure 6 shows the averages of evaluations of synthesized speech quality based on a scale of 1 to 5. The evaluation score for this method is 3.8, which is a remarkable improvement in speech quality in comparison with the 2.3 score of the former method.

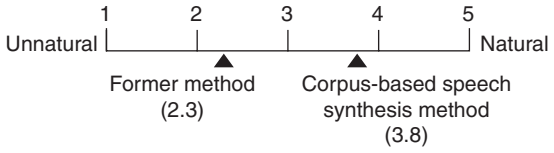

Figure 6
Evaluations of speech synthesis systems.

We demonstrated that our speech synthesis system can express emotions such as joy, anger, and sorrow by preparing an emotional speech corpus and prosody generating model for each emotion.[6]

We have also built a speech synthesis library that is compatible with Microsoft Speech API 5.0. This library, called FineSpeech, is in use in many IVR (Interactive Voice Response) systems and voice portals.

5.3 Automatic speech corpus generation

Because speech synthesis is being used in more and more applications, demand for a wider range of voices has increased. To prepare a speech corpus for a new voice type, phoneme labeling and pitch marking must be done for the new speech data to be used (**Figure 7**). Phoneme labeling determines the time segments corresponding to the phoneme sequence in the voice waveform, and pitch marking marks the pitch locations. Pitch marks are necessary to control prosody by pitch modification in the speech synthesis process. Conventionally, these tasks are performed through visual inspections by experts that can take as long

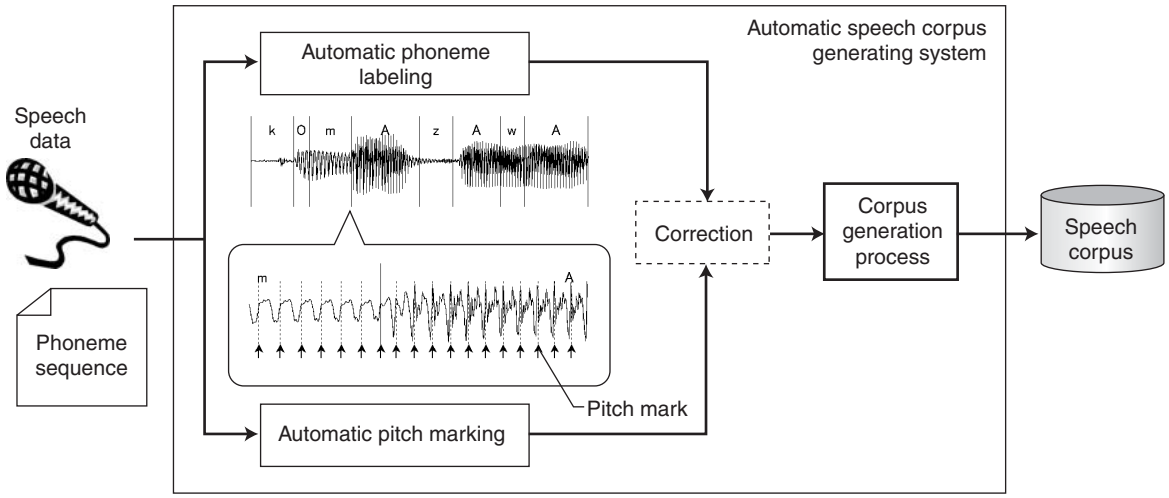

Figure 7
Automatic speech corpus generating system.

```
<form id=airplane_service>
    <grammar src="Information.grxml"type="application/grammar+xml"/>
    <block>Welcome to Fujitsu airline schedule information service.</block>
    <field name="Place of departure">
        <prompt>From which airport do you depart?</prompt>
        <grammar src="Airport.grxml" type="application/grammar+xml"/>
    </field>
    <field name="Arrival airport">
        <prompt>Which airport do you arrive?</prompt>
        <grammar src="Airport.grxml" type="application/grammar+xml"/>
    </field>
    <field name="Date of departure">
        <prompt>When do you depart?</prompt>
        <grammar src="Date.grxml" type="application/grammar+xml"/>
    </field>
    <block>
        <prompt>Thank you for your access. </prompt>
        <submit http="http://www.vxml.com/servlet/airplane_service">
    </block>
</form>
```

Figure 8
Example of VoiceXML script.

as several months. To reduce the work, we developed an automatic speech corpus generation technique. Using this technique, the process can now be done automatically and new voice requirements can be rapidly met.

1) Automatic phoneme labeling

We developed a technique for automatically marking the segment boundaries of each phoneme in speech data using a phoneme recognition technique with the Hidden Markov Model. We also developed a post-processing unit for correcting errors such as assigning phoneme labels to silent segments and extremely short segments.

2) Automatic pitch marking

The locations of pitch marks significantly influence the quality of synthesized speech. We developed a technique to automatically extract the basic frequency of each frame from speech data and detect the pitch mark locations that achieve the best quality of synthesized speech.

Although these techniques are less accurate at points of unstable utterances than the manual method, they have made it possible to prepare a speech corpus for synthesizing speech after automatic generation processing in just a few hours. To improve the quality of synthesized speech, manual correction after automatic generation is required. We have confirmed that our automatic speech corpus generation technique can greatly reduce the labor time in comparison with a fully manual preparation of the speech corpus.

6. Dialogue control

Many of the interactive scenarios used in voice portals have been created by custom design. Standardization activities for an XML-based script language called VoiceXML (Voice eXtensible Markup Language) were started in 1999. VoiceXML is a task (objective) oriented language that acquires the information required to meet an objective through conversational interactivity. With this language, the system allows a user to request or provide multiple information items in the same speech string. A feature of VoiceXML is its high affinity with the existing Web system. **Figure 8** shows an example of a voice interactive scenario written with VoiceXML.

Fujitsu has developed a dialogue control engine that conforms to VoiceXML2.0. This engine is composed of two processing blocks: an interpreter and an interpreter context. The interpreter interprets voice-interactive scenarios written in VoiceXML, and the interpreter context controls the platform and the interface with the voice rec-

ognition/synthesis engines and other components.

When evaluating a voice interactive system, a refined voice interactive scenario is essential. Replacing the dialog of an existing GUI with voice interactivity is not sufficient, because developers must consider the characteristics of speech. For example, the proper expression and length of a prompt to the user and the sentence coverage of the recognition grammar should be thoroughly evaluated for each conversation with respect to the time it takes to meet the objective, the rate of attainment, the objective evaluation of the subject, and other details. Fujitsu has developed a prototype system for practical application in areas such as facility reservation and schedule guides and has accumulated practical know-how for voice interactive scenarios by repeating in-house trials and making improvements to refine the voice interactivity.

7. Conclusion

This paper gave an overview of voice interface applications and information access systems that operate over a telephone network. It introduced a voice portal solution for providing various information services that use speech recognition/synthesis and dialogue control technologies. It also described various advanced speech processing technologies that Fujitsu has developed for realizing human/computer voice communication. With the aim of popularizing voice portal solutions in various fields, we will continue to improve these speech technologies and also evaluate and improve our voice portal systems in practical fields.

References

1) S. Kimura: Progress in Speech Synthesis and Speech Recognition Technologies. (in Japanese), *FUJITSU*, **49**, 1, p.41-46 (1998).
2) H. Tsujiuchi: Development Tool for Telephony Information Service Using Speech Synthesis. (in Japanese), *FUJITSU*, **49**, 1, p.57- 60 (1998).
3) The World Wide Web Consortium. *http://www.w3.org/*
4) S. Nakagawa: Speech Recognition by Probability Model. (in Japanese), first edition, Tokyo, The Institute of Electronics, Information and Communication Engineers, 1988.
5) N. Katae et al.: High-Quality Speech Synthesis Technologies. (in Japanese), *FUJITSU*, **49**, 1, p.47-51 (1998).
6) N. Katae et al.: Effects of Voice Quality and Rhythm on Feeling Speech Synthesis. (in Japanese), Autumn Meeting of Acoustical Society of Japan, 2-1-7, 2000, p.187-188.

Yasushi Yamazaki received the B.S. and M.S. degrees in Computer Information Sciences from Tokyo University of Agriculture & Technologies, Tokyo, Japan in 1987 and 1989, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1989, where he has been engaged in research and development of speech processing systems.

Hitoshi Iwamida received the B.E. degree in Electronic Engineering from Hokkaido University, Sapporo, Japan in 1983. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1983, where he has been engaged in research and development of speech recognition systems.

Kazuhiro Watanabe received the B.E. and M.E. degrees in Electronics and Communication Engineering from Waseda University, Tokyo, Japan in 1982 and 1984, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1984, where he has been engaged in research and development of speech synthesis systems and ultrasonic systems.