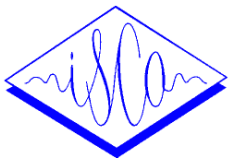

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Japanese Text-to-Speech Equipment: Current Applications and Trends

Hirokazu Sato

NTT Human Interface Laboratories
Yokosuka-shi, Kanagawa-ken, 238-03 Japan

ABSTRACT

This paper introduces the Japanese text-to-speech equipment developed by Nippon Telegraph and Telephone Corporation, and describes several current applications of text-to-speech systems such as ANSER. Besides the original role of generating speech from unrestricted texts, other application possibilities are discussed. Moreover, this paper discusses the importance of achieving low-priced equipment, improving the synthetic speech quality, and realizing the controllability of voice quality and speech style. The solution of these problems will expand the future application fields of text-to-speech systems.

1. INTRODUCTION

Over the last decade, the technology of speech synthesis by rule has been increasingly developed to text-to-speech systems. As the systems transform input orthographic text to speech without requiring complicated phonological representation, they have increased the application areas of speech synthesis. At present, commercial text-to-speech equipment and systems are available for several languages.

This paper introduces Japanese text-to-speech equipment that Nippon Telegraph and Telephone (NTT) has developed, and their current applications and potentials. Text-to-speech systems are most useful when very large quantities of arbitrary texts or messages must be synthesized. The development of low-priced synthesizers has lead to their use in several non-traditional applications. This paper discusses other possible application areas for text-to-speech systems. It must, however, be recognized that the market for text-to-speech synthesizers has not increased as much as experts in this field had hoped. Current problems that restrict the application of text-to-speech system are also discussed.

2. TEXT-TO-SPEECH EQUIPMENT

NTT has advanced Japanese speech synthesis by rule and text-to-speech techniques, and has developed several systems and equipment sets over the past twenty years. The first laboratory system for speech synthesis developed by NTT combined VCV (Vowel-Consonant-Vowel) composite synthesis units with formant synthesis [1]. VCV-based speech synthesis was combined with a linear prediction technique (PARCOR) [2], and was applied to ANSER (Automatic Answer Network System for Electrical Request) for the banking industry in 1981 [3]. The speech synthesizer in ANSER is compositely structured with a rule-based speech generation component, which allows an unlimited vocabulary, and a speech parameter storage component for the output of fixed messages. A new version of ANSER's synthesizer was developed that used CVC (Consonant-Vowel-Consonant) -based speech synthesis [4] to improve the synthetic speech quality.

ANSER does not, in the strict sense, employ a text-to-speech system, because input characters are composed of only kana strings that approximately correspond to phonetic symbols. After the laboratory system for Japanese text-to-speech, to transform orthographically written text including Chinese characters (Kanji) into speech, was examined [5], the first commercial Japanese text-to-speech equipment was developed in 1987 [6]. As this equipment includes a common interface for personal computer control and network control, compact and personal audio response units can be easily constructed. This equipment was named "Petit ANSER"[7]. In this text-to-speech equipment, dyad type speech elements, CV and VC, are used as concatenation units. CV is the basic Japanese syllable, and VC is a transition from the vowel to the next consonant. Petit ANSER has been used to construct a revision support system for newspaper editing, several information passing systems, and so on. These applications will be described in section 3.

The problem with Petit ANSER was that it was very expensive. A low cost, single PC board version based on the same techniques as Petit ANSER was developed by NTT Data Inc. in 1989, and is currently available. By using the new version, a full-featured voice output system can be created inexpensively.

To realize a text-to-speech synthesizer that produces more natural sounding speech, we are developing a new text-to-speech technique. The speech synthesis method employed is called the COC method because synthesis units with phoneme context information are automatically generated with the Context Oriented Clustering technique[8]. Coarticulatory characteristics and allophonic variations in speech can be accurately reproduced by concatenating these synthesis units. The current number of the units is about 1500, but it is possible to enhance speech quality by increasing the number of units. The details of the COC text-to-speech synthesizer are given in these proceedings [9].

3. EXAMPLES OF APPLICATIONS

3.1 ANSER system

ANSER is a combined system of speech synthesis and recognition, that provides banking information services over telephone networks. Rule-based speech synthesis was used to output proper nouns (personal and company names) in a money transfer notification service. When ANSER was first developed in 1981, the system had only audio response capability. Later, speaker-independent speech recognition was introduced, which made dial telephone access possible. Moreover, ANSER has been enhanced to a system that offers facsimile and modem access capabilities. The ANSER system configuration is shown in Fig. 1.

ANSER generally provides two kinds of banking services. One is the notification of money transfers. The computer automatically calls users and notifies them of the transactions using synthetic speech. The other is an inquiry service. In response to a user's call, the computer provides information about account balance or other bank services.

According to a recent report on ANSER [3], the system has spread to more than 70 Japanese cities, and 95% of the financial institutions in Japan offer ANSER services. The average traffic is about 16 million calls a month. Recently, insurance companies have begun using ANSER systems to provide information about fund transfers. The number of customers, including financial and insurance companies, is steadily increasing.

ANSER is one of the most successful speech synthesis systems. The reasons for its success are:
(1) Transaction notification services were already common in Japan but human operators were used. ANSER was an effective means of both expanding services and achieving work efficiencies.
(2) ANSER was designed so that several banks could share one system. Since sharing lowered the expense for each bank, many banks in the provinces utilized ANSER systems.
(3) In Japan, touch tone telephones are not wide spread. About 70% of all telephones are still the rotary type. This may be, interestingly, one reason for ANSER's success, because the majority of users with the rotary dial telephones can access ANSER services via its speech recognition techniques.

3.2 Revision support

The Japanese text-to-speech equipment, Petit ANSER, has been successfully used to realize a revision support system for newspaper proofs. Such systems have been used in a newspaper publishing company since 1987. This system makes it easy for revisers to correct transcription errors in articles.

Japanese newspaper articles are input into the computer from character input terminals as Kanji and kana characters. The text-to-speech synthesizer "narrates" the article including input errors, and revisers check the article by listening to the synthetic voice while watching the text on a display terminal. The display of the article was arranged so that unresistered words or homonyms are displayed in special colors to attract the reviser's attention.

Previously, this work of newspapers required two revisers. One reviser narrated the article loudly, while the other checked it. It is obvious, therefore, that the introduction of this system reduced man power cost. The more important advantage of the system is that it has prevented the mental strain of continuous personal interaction. Experience has shown that users quickly adapt to the synthetic speech and can comfortably listen to its output at its maximum speed of 13 syllables per second. It is interesting that most users prefer the male synthetic voice, although the system can output both male and female voices. This is because low pitched male voice is easier to listen to for a long time.

3.3 Information providing system

Text-to-speech systems are being used in several information providing services such as those offering baseball match results through telephone networks. This section introduces an example of an office information providing system.

A Japanese trading company has installed a support system for money transfer that effectively uses text-to-speech equipment. The system receives a remittance receival notice from abroad via the network between a bank computer and the company computer. The message of the remittance arrival notice, including country name and amount of money, is converted to speech with the text-to-speech synthesizer and sent to the proper section by the autocalling function. The section that receives the message confirms that the remittance is for the section. The configuration of this information providing system is shown in Fig. 2.

3.4 Another application

A building equipment control system has been constructed with the text-to-speech equipment. This system responds to client requests for building equipment control such as temperature or lighting control. A client calls the system and enters the data using a touch tone telephone, while being prompted by a synthetic voice. The building equipment is controlled automatically and then the system reports to the client that the request has been completed with.

4. TEXT-TO-SPEECH APPLICATION AREAS

4.1 Applications using original function

It is well known that text-to-speech systems have been successfully applied to services for which speech output of unlimited words or text-to-speech transformation is essential. This is the original function of text-to-speech systems. The above-mentioned ANSER services and the revision support system, or the confirmation by voice of texts input into word-processors are examples of such applications. However, we would like to show that other applications for text-to-speech systems can be found in practical systems.

4.2 Use expansion with lower price

Progress in semiconductor and VLSI technologies has resulted in a dramatic lowering of device cost. Text-to-speech equipment has naturally benefitted from this trend. For instance, the Petit ANSER speech synthesizer was originally priced at about $8000, but the new text-to-speech PC board is priced at just $1400. A much lower-priced version of this board is now being developed, and this tendency will only accelerate in the future.

With the above in mind, the text-to-speech equipment has begun to be applied to voice response

systems that require fixed messages and limited vocabulary. Traditional audio response units are configured to send messages simultaneously over multiple lines because multiplex processing lowers the cost per line. One of the latest order entry systems using audio response has been constructed with individual text-to-speech boards for each output line. This no-multiplex hardware is advantageous in its high reliability and easy maintenance. The low cost of the equipment has made this simple structure possible.

Moreover, the PC interface makes it easy to construct an economical small-scale audio response system for one output line. Thus, if very high quality of output speech is not necessity, text-to-speech equipment will become utilized by a greater variety of services, even in applications that need just fixed message outputs, such as in the building equipment control system.

4.3 Use as speech file producing equipment

A text-to-speech system can eliminate the need to prepare speech files. The fixed message audio response services, such as information providing services or information inquiry services, demand a lot of speech recording and speech file preparation. If the output words are frequently changed or new words are often added, the speech file preparation task will significantly raise the cost of the audio response system. Moreover, it is not easy to ensure that the same narrator is used.

An audio response system constructed with text-to-speech equipment, therefore, is advantageous because it eliminates speech file preparation overhead. If a text-to-speech system is used as speech file producing equipment, many auxiliary input symbols to control intonation or speech style will be needed to produce a great variety of speech.

5. APPLICATION EXPANSION

The application areas of the text-to-speech system have increased gradually. Unfortunately, applications remain sporadic and are not widely diffused in the public domain. The big bottle neck is the price. According to investigations made on the purchasers of Petit ANSER, almost all of them were corporations and research institutes because the equipment was so expensive [8]. Low-priced equipment that can be freely purchased by private individuals will be the key for expanding the application area.

Besides the price problem, here are several barriers against increasing the number of text-to-speech applications. They are described in the following.

(1)The unsatisfactory synthetic speech quality.

The quality of synthetic speech is the chief requisite for applications. It has been improved to the point where it is clear and intelligible. Synthetic speech is practical in many respects, but it is still quite different from human speech, particularly in its naturalness. Synthetic speech is inferior to human voice in :

① Natural voice quality
② Articulatory movements
(dull or over articulated mouth movements)
③ Prosodic characteristics
(accent, intonation, rhythm and so on)
④ Mixture of unnatural sounds
(related to waveform synthesizer, synthesis unit concatenation or amplitude control)

If synthetic speech quality could be made equal to that of human speech, text-to-speech will naturally be applied to most audio response services.

Accurate and stable text analysis is also very important for generating high quality speech, because reading errors, for example, in Chinese character pronunciation, incur an unacceptable degradation in synthetic speech quality. Input character decomposition errors and accent mis-assignments sometimes degrade prosodic naturalness. In addition, sophisticated text analysis can give useful linguistic information for generating more natural prosodic characteristics. In text analysis, the most difficult problem to be solved is how to "pronounce" unknown words (words unregistered in the system).

(2)Uncontrollability of voice quality and speech styles

Current text-to-speech systems output speech in an ordinary narration style, and the intonation is relatively monotonous. They cannot generate conversational speech, emotional intonation or other specially styled utterances. Moreover, the voice quality is also restricted to the male and/or female voices prepared in the system. It is necessary that voice quality or speech style be controllable according to the service or user preferences. Synthesis of conversational speech is inevitable for services using human-machine communication by voice. The monotonous intonation and the unchangeable voice quality will cause users to be dissatisfied. When the controllability for generating various speech becomes possible, text-to-speech systems will have advanced beyond the compilation technique of pre-recorded speech.

(3) Restricted audio response services

Most current audio response services do not require many output messages, because the services are restricted to specific domains, or because the systems guide the dialogue process. For such restricted services, text-to-speech systems are not needed, and audio response units of the fixed message type can be most effective.

If the conversation systems linking humans to machines become more intelligent, and if machines flexibly respond to human demands, text-to-speech technology will become essential, because its output speech is almost unrestricted. Progress in the technology of speech synthesis from concept will help realization of intelligent audio response systems.

6. SUMMARY

This paper has described text-to-speech systems developed in NTT and their applications. As examples of such applications, the ANSER system, a revision support system, an information providing system and a building equipment control system were introduced.

Although the application areas of text-to-speech systems are still restricted, the author is confident that text-to-speech systems have a great deal of potentials. Currently, several voice response services have begun to use text-to-speech systems even for fixed message services. The full potential of text-to-speech will only be realized when three requirements are fullfilled: lower equipment cost, improved speech quality, and controllable speech style.

ACKNOWLEDGEMENT

The author wishs to thank Dr. S. Furui for his helpful guidance, and also thanks T. Hirokawa and K. Hakoda for their comments and discussions.

References

[1] S. Saito and S. Hashimoto, "Speech Synthesis System Based on Interphoneme Transition Units", Proceedings of 6th ICA, B-5-12, p.B-195 (1968)

[2] H. Sato, "Speech Synthesis Based on PARCOR-VCV Concatenation Units (in Japanese)", Trans. Commit. Speech Res., Acoust. Soc. Jpn., S74-22 (1974)

[3] R. Nakatsu, "ANSER: An Application of Speech Technology to the Japanese Banking Industry", IEEE Computer, Vol.23, No. 8, p.43 (1990)

[4] H. Sato, "Speech Synthesis Using CVC-Concatenation Units and Excitation Waveform Elements (in Japanese)", Trans. Commit. Speech Res., Acoust. Soc. Jpn., S83-69 (1984)

[5] H. Sato, "Japanese Text-to-Speech Conversion System", Review of the ECL, Vol.32, p.179 (1984)

[6] K. Hakoda, K. Nagakura, T. Hirahara and K. Kabeya, "Japanese Text-to-Speech Synthesizer", Journal of the American Voice I/O Society, Vol.6, p.1 (1989)

[7] T. Hirokawa, "Application of Japanese Text-to-Speech Synthesizer", Speech Tech'89, p.30 (1989)

[8] S. Nakajima and H. Hamada, "Automatic Generation of Synthesis Units Based on Context Oriented Clustering", Proc. IEEE Int. Conf. ASSP, S-14.2 (1988)

[9]K. Hakoda, S. Nakajima, T. Hirokawa and H. Mizuno, "A New Japanese Text-to-Speech Synthesizer Based on COC Synthesis Method", these proceedings

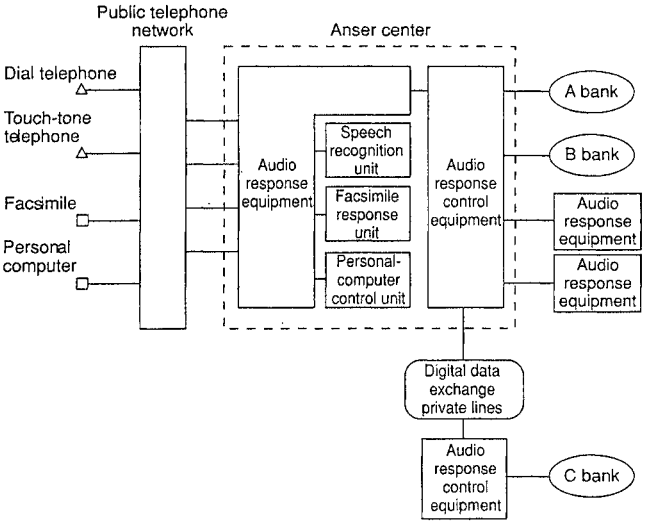

Figure 1. ANSER system configuration

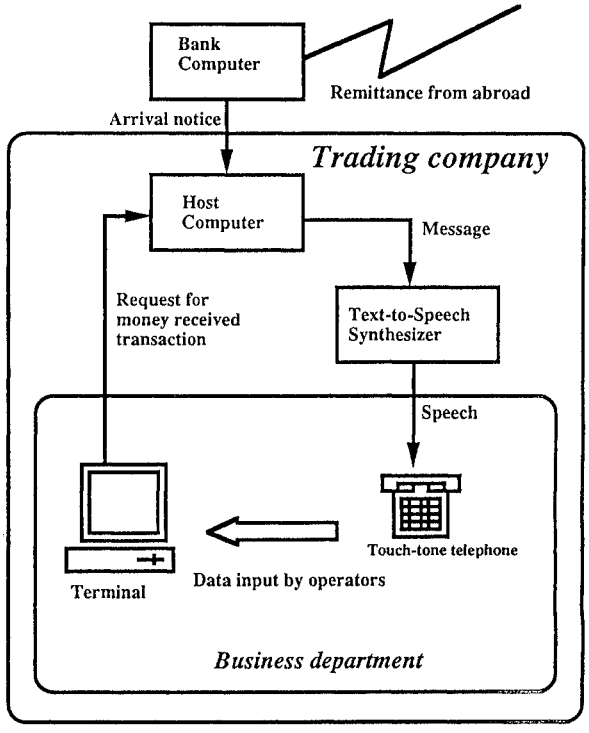

Figure 2. Information providing system configuration