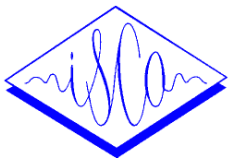

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

A JAPANESE TEXT-TO-SPEECH SYSTEM FOR ELECTRONIC MAIL

*Hiroyoshi Saito *, Motoshi Kurihara *, Ken-ichiro Kobayashi**,*
*Yoshiyuki Hara *, Naritoshi Saito**

*Information & Communication Systems Laboratory, Toshiba Corporation
70, Yanagi-cho, Saiwai-ku, Kawasaki, 210 Japan
**Toshiba Audio Video Engineering Co., Ltd.
70, Yanagi-cho, Saiwai-ku, Kawasaki, 210 Japan

ABSTRACT

This paper describes the methods for a Japanese text-to-speech system the authors have developed and its application to a voice supply mechanism for electronic mail. This system analyzes an arbitrary text morphologically with a dictionary which has 100,000 entry words, their grammatical attributes, and their phonetic and accent information. This analysis separates the text into individual terms, and the phonetic and accent data for the individual terms are obtained from the dictionary. In the proposed system, rules play an important role in improving the speech in order to produce a high quality speech. These rules decide whether the phonetic symbols for each term are separated or unified, and whether a part of them are modified or not : to modify a vowel sound to a long sound, to make a breath sound, and to make a nasal sound, whether the accent position of each term is shifted or not, and how long the duration between individual terms is.

The authors tried out this system in the field of electronic mail. If the mail receiver wishes to confirm a mail text when out of his or her office, the person can call the mail box by telephone and obtain a synthesized speech of the mail produced by the system.

1. INTRODUCTION

Speaking out the contents of texts is desirable in many cases, instead of looking at the texts on a display. Examples are such cases as verifying texts entered into a word processor, listening to the context of an electronic mail through a telephone, and realizing an interactive human-machine interface through the speech of a natural language.

To realize the practical use of such applications, a high quality meaning both clearness and naturalness is required for a text-to-speech system. First, an input text must be analyzed into words exactly so that the correct information of phonemes and accents can be obtained from a dictionary. However a high quality speech can not be produced only from such retrieved information. Many rules should be developed in order to modify and improve the information. Thus a speech synthesis unit must output clear and natural speech sounds by rules, operated by coded symbols which the text analyzing process gives.

The authors have been making effort to develop a text analysis method to supplement the phonetic and prosodic method. A text analysis technique and speech synthesis rules are being developed so that we can obtain a high quality speech for arbitrary texts. These text analysis and speech synthesis methods are described in the paper. And a mail-voice system, which is an application of the above speech synthesis technique, is also introduced.

2. JAPANESE TEXT-TO-SPEECH SYSTEM

This system consists of three processing parts, which are the analysis part, phonetic and prosodic symbolizing part, and speech synthesis part. First, the analysis part analyzes the input Japanese texts morphologically with a word dictionary and grammar into individual words. A Kanji dictionary is used for analyzing unknown words of Kanji characters. These dictionaries give each retrieved word its reading and accent information. Second, the phonetic and prosodic symbolizing part determines the phonetic and prosodic information of each word by rules. The speech synthesis part finally generates the acoustic sound for each phonetic and prosodic symbol by a speech synthesis module. Figure 1 shows the principal structure of this system.

3. ANALYSIS INTO PHONETIC AND PROSODIC SYMBOLS

3.1 Analysis into Words

Japanese sentence analysis in this system is mainly based on the morphological analysis. Morphological analysis chooses an optimal structure of words which compose the input sentence through looking up the word dictionary which has 100,000 terms, examining propriety of connection between neighboring words with grammar, and comparing each possible composition of words. Every entry in the dictionary that corresponds to each portion of the input sentence is retrieved at dictionary look up. At this time, among all retrieved word entries, those except ones which can be connected to both the succeeding and the preceding words are rejected, by referring to a grammatical connection table.

Then, after rejecting such ungrammatical compositions of words,

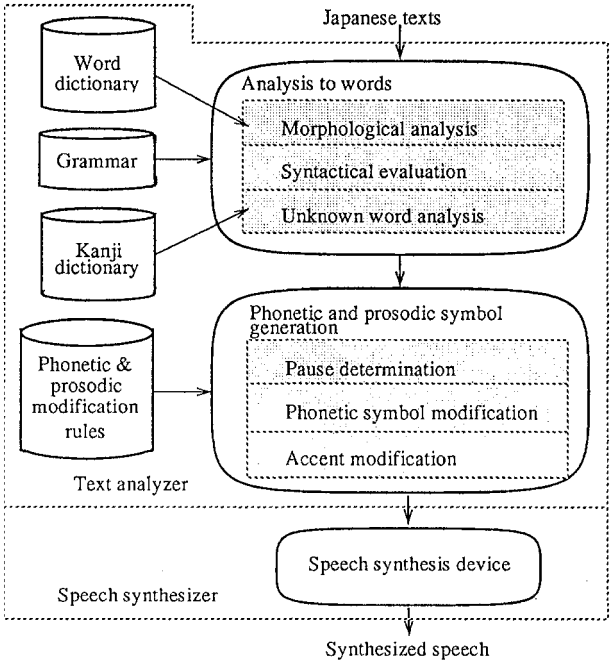

Fig. 1 Structure of Text-to-Speech System

Input text
図書館へ 行って本を読みます
(I go to a library and read books.)
Result of morphological analysis

図書館	[N]	としょかん	AccType = 2
へ	[P]	へ	
行っ	[V5ky]	いっ	AccType = 0
行っ	[V5wy]	おこなっ	AccType = 0
て	[P]	て	
本	[N]	ほん	AccType = 1
を	[P]	を	
読み	[V5my]	よみ	AccType = 1
ます	[Js]	ます	

Result of syntactical evaluation
としょかん/へ/いっ/て/ほん/を/よみ/ます/

Fig. 2 Example of Analysis to Words

a string of words is selected. In some cases, there are plural candidates of possible compositions of selected words remaining. The authors prepared priority rules based on the syntactical condition and heristics for the likelihood of composing a sentence. All the priority values evaluated from these rules for possible word compositions are compared with one another, and the optimal composition is selected. Figure 2 shows an example of the analyzing process.

3.2 Determination of Phonetic Symbol

Each Japanese Kana character represents a syllable in general. And each term in a Kanji character has a corresponding Kana character representation for reading. The basic information on phonetic symbols is given by referring to the Kana representation in the dictionary for each retrieved word. In order to realize a high quality speech, it is necessary to insert pauses in a sentence, to modify the phonetic symbols by some case, and to add or shift accents to their proper positions. In the following, important treatments the authors have designed will be shown considering typical phenomena in standard Japanese pronunciation.

3.2.1 Pause Insertion

Pause symbols are inserted between each "bunsetsu" according to the result of morphological analysis. "Bunsetsu" is the smallest phrase in Japanese. However, people usually speak continuously for some sequentially related bunsetsu clauses. So, there are some rules which have been designed to determine whether neighboring words should be unified or not, and to settle the length of a pause.

Each designed rule is mainly represented as a relation in a set of three words, especially for a set of a postpositional word and other words. To insert a pause before a post noun in a sequence as "noun + 'ha(は)' + noun", or to connect all phonetic symbols in a sequence as "noun + 'wo(を)' + verb" are examples.

3.2.2 Prolonged Vowel

One of the phonetic modifications is to prolong a vowel in some cases. A vowel following a syllable that includes the same vowel makes the previous syllable change to a prolonged vowel regularly. In the case of a vowel "i" following an "e" sound or an "u" following an "o" sound, they also become long sounds of "e" or "o" respectively. The rule for such a transition is simple but should not be applied outside the word boundary.

3.2.3 Phonological Transition

Some combination of words changes their reading. The first syllable of a quantifier which includes the consonant "k" or "h" changes to the consonant "g" or "p"("b") in some cases caused by the pronunciation of a previous numeral. When the first phoneme of a quantifier following some numeral is "k" or "p", the last sound of the numeral like "1", "6" or "8" changes to assimilated sounds. These phenomena occur regularly for some quantifiers, and rules for them can be described in sets of two terms. They do not apply to foreign terms such as a Katakana entry.

The following are examples of such phoneme changes.

ici(一) + hai(杯) --> <ip pai>
ni(二) + hai(杯) --> <ni hai>
san(三) + hai(杯) --> <san bai>

As for nasalization, the sound of phoneme whose consonant part is "g" is determined by its position in a word. Its sound doesn't change when it is in the first position, but their sounds change to a nasalized "g" in other positions or for the postpositional word "ga".

The followings show examples using "Ng" as a symbol of nasalized "g".

Gakkou(学校) ga(が) mieru(見える) ---> <gakkou Nga mieru>
Ongaku(音楽) wo(を) kiku(聞く) ---> <onNgaku wo kiku>

The other case of a changing sound is devoicing such as follows.

tasikameru(確かめる) ---> <tas(i)kameru>
arimasu(有ります) ---> <arimas(u)>

There are two patterns in a Japanese devocalization. One is the case that a consonant among "k", "s", "t", "h" or "p" exists just after any following syllable.

["ki" "ku" "si" "su" "ci(ち)" "cu(つ)" "hi" "hu" "pi" "pu" "shu"]

The other case is when a syllable among the following exists just before a pause.

["ki" "ku" "si" "su" "ci" "cu" "pi" "pu" "shu"]

It is necessary that a proper pause position for each term has been determined before this devoicing rule is applied. In the case when an object syllable has an accent, this rule is not applied.

3.3 Accent Modification

The initial accent information on each word is obtained from the dictionary as prosodic information. The Japanese accent does not really mean an accent in the Western language (which expresses the strength of pronunciation) but means the position of the last mora of the high-pitched portion in a term or phrase. It is more of an intonation than an accent. A type of accent is used to express the accent information on standard Japanese speeches for each term.

The accent types in the dictionary are described for most words when the words exist alone. But the accents of some words like some postpositional words are not shown in the dictionary because they shift or disappear in somes. Some modification rules should be applied to such initial accent information in order to obtain natural speech sounds.

A speech synthesis module can generate a pitch controlled speech sound for a specified accent type.

3.3.1 Varying for Inflection

The accent form varies regularly for each inflection of verbs, adjectives, or verbal adjectives. A retrieved initial accent information, which is given for the basic form of an inflectional word, is shifted to its proper position for the inflected form by the specified rule for inflection. The following is an example of such accent variation.

hasiru(走る) --- <type 2> , initial
hasira(走ら) nai(ない) --- <type 3>
hasiri(走り) masu(ます) --- <type 0>
hasiru(走る) toki(とき) --- <type 2>
hasire(走れ) ba(ば) --- <type 2>

3.3.2 Varying for Word Unification

In general, the accent position often shifts when one of the words among independent words and dependent words is connected to the other to make a compound word. These variations have the characteristic of accent unification or accent rise. Examples of accent changes are as follows:

yo^mu(読む)<type 1> ("^" shows an accent)
yomi(読み)na^gara(ながら) : <type 3>
naku(泣く)<type 0>
naki(泣き)nagara(ながら) : <type 0>
bu(部)<no type> / ka(家)<no type> / go(後)<no type>
kacudo^u(活動)bu(部) : <type 3>
kacudou(活動)ka(家) : <type 0>
kacudou(活動)go^(後) : <type 5>
kika^i(機械)<type 2> / hon yaku(翻訳)<type 0>
kikai(機械)ho^n yaku(翻訳) : <type 4>

Basically, the rule for such variation control can be considered under conditions in a set of two words, but sometimes conditions for three or four items are needed. An existence of a word which possesses the initial accent should also be considered for such control.

3.4 Speech Generation for an Unknown Word

The system must estimate the reading of an unknown word in Kanji characters after morphological analysis. We use a Kanji dictionary in which reading information with high frequencies of use is shown for each Kanji character. And each reading is shown with its frequency and a sign which indicates whether it is classical Chinese-style reading or Japanese-style reading in this dictionary.

In general, there is a tendency that the kind of reading, such as Chinese-style or Japanese-style successively uses the same kind of reading for a Kanji sequence. The initial phonetic symbols for an unknown Kanji word is determined by referring to this dictionary, and when a sequential Kanji string composes the unknown word, each phonetic symbol is adjusted to have the same kind of reading.

Then, these phonetic symbols are modified the same as other ordinal terms, and their accent information is added by referring to the table in which the accent type is defined as a function of the total number of syllables, and the kind of character of the word. This function comes from the general feature of Japanese speaking, as shown in Table 1.

Table 1 Examples of Relations between Japanese Word and Accent Position

Word	Accent position
Katakana(2 mora)	1st mora
Katakana(>2 mora)	3rd mora from end
Kanji(1 mora)+Kanji	1st mora

3.5 Control Procedure

After analyzing a text and retrieving the basic information of phonetic and accent data from the dictionary, the above modifications of the phonetic and accent symbols progress in the order as follows:

(1) Determination of the reading for an unknown word
(2) Word unification and pause insertion
(3) Changing phonetic symbols to prolonged vowels, nasalized sounds and some particular pronunciation
(4) Adding devoicing symbols to particular syllables
(5) Accent shifting for inflections
(6) Accent modifications for unified words

4. SPEECH SYNTHESIS [4][5]

The authors' speech synthesis device uses a cepstral parameter as a feature parameter of speech. A cepstral parameter is defined as the inverse Fourier transform of the short-time logarithmic amplitude spectrum obtained from the vocal data of a CV unit through discrete Fourier transform (DFT). This cepstral parameter has many characteristics. It can separately represent a spectral envelope and a fine structure. Here, the cepstrum of low frequency elements representing the spectral envelope is used. This speech synthesizer has an M-sequence generator producing a unvoiced sound, and an impulse generator producing a voiced sound as a sound source, and a speech is synthesized through an LMA (log magnitude approximation) filter which has the cepstral parameter as a direct factor. [3]

4.1 The Outline of Speech Synthesis

Information processing for speech synthesis is shown in Fig. 3. A phoneme string with a prosody control symbol given from the text analyzer is entered into a generator of phonetic parameters and a generator of prosodic parameters. The phonetic controller generates phonetic parameter strings, and the prosodic controller generates prosodic parameter strings.

The generator of phonetic parameters extracts some control parameters corresponding to the entered phoneme string from a file of CV-syllable parameters and determines the segmental duration of each phoneme using the segmental duration parameter. Using this segmental duration, each of the CV-syllable parameters (cepstral parameter), which correspond to the entered phoneme string, is interpolated and concatenated by the interpolation-concatenation rule.

The cepstral parameter C0 to CM, and voiced-unvoiced information (V/U) resulting from this process are given to the synthesizer. After the segmental duration is determined for each phonetic parameter, the generator of prosodic parameters produces pitch patterns using the segmental duration and prosody parameters.

A pitch pattern IP can be made by superposing the elements for intonation and accent to the pitch pattern on the logarithmic axis. The element of intonation means a base declination pattern which generally appears in human speech. This pitch pattern IP is entered to the synthesizer. The synthesizer has an impulse train generator and an M-sequence generator as a sources of sounds.

The impulse train generator is used as the source for voiced sounds, and the M-sequence generator is used as the source for unvoiced sounds. Which generator will be active is determined by the voiced-unvoiced information (V/U).

The fundamental frequency pitch (IP) for a voiced sound is used as the data entered to the impulse train generator. The amplitude is calculated from the following equation, using the data resulting from the impulse train generator.

$$Am = \exp(C0)\ \mathrm{sqrt}(IP)$$

This data is entered to the LMA filter. This LMA filter synthesizes the speech using the cepstral parameter entered to it.

4.2 Structure of the Speech Synthesis Device

The hardware structure of the speech synthesis device is shown in Fig. 4, and its main characteristics are shown in Table 2. The device is constructed on only one board whose size is 100 × 340 mm, as large as an extension board of a personal computer (J-3100) sold by Toshiba. The generator of synthesis parameters (as shown in Fig. 3(a)) consists of MP68000 as a controller, and the synthesizer (shown in Fig. 3(b)) consists of DSP (TMS320C25) as a controller.

5. AN APPLICATION TO MAIL-VOICE

An electronic mail system can be considered as a probable application of text-to-speech synthesis. In the system, we can send a mail to the other person with a personal computer. If the text-to-

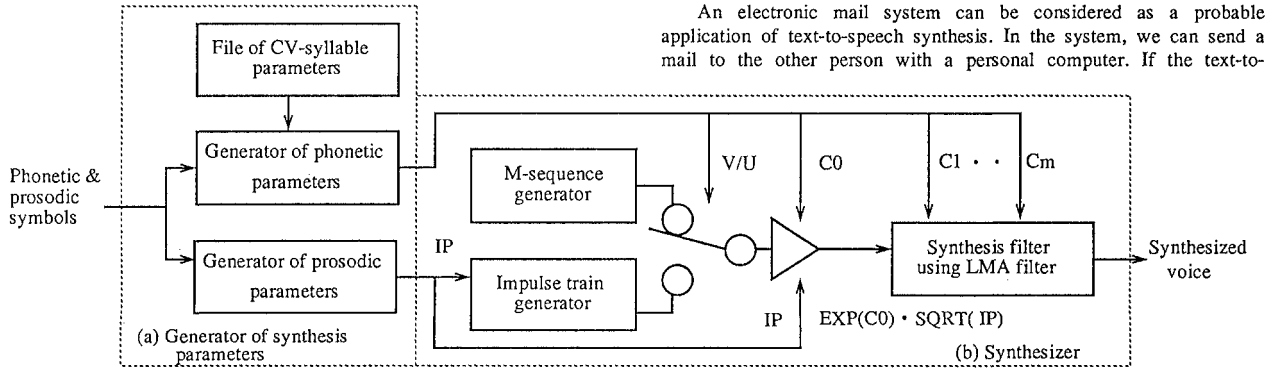

Fig. 3 Processing Flow of Speech Synthesis

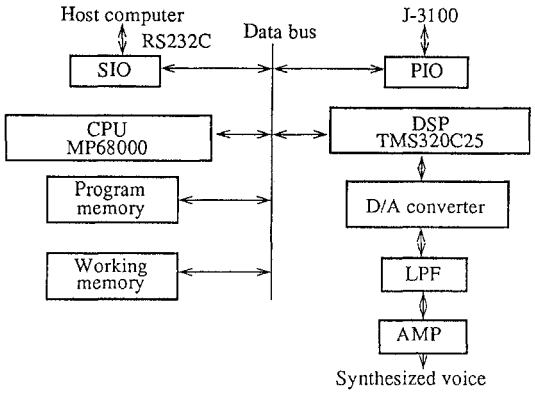

FIg. 4 Hardware Structure of Speech Synthesis Device

Table 2 Main Characteristics of Speech Synthesis Device

Synthetic Parameter	Cepstral parameter
Synthesis Unit	5 vowels, a syllabic nasal /N/ and 131 open monosyllables, Male / Female
I/O Port	RS232C, I/O port for J-3100
Others	Speaking rate and pitch are variable by software

speech synthesis is built into the electronic mail system which generally has electronic mails as character data and holds them in a mail box of a central host computer, we can not only read the contents of some mail on the personal computer's display but also listen to the contents through speech synthesis.

5.1 System Structure

There are two methods on how to build a mail-voice system using text-to-speech synthesis.

(1) The personal computer has a text-to-speech synthesis device and an NCU (telephone network control unit).
(2) The central host computer has a text-to-speech synthesis device and an NCU.

Method (1) has an advantage that we can use the text-to-speech device not only for electronic mail but for other purposes. But it also has a disadvantage that we have to buy an expensive machine. The authors adopted method (2), because it is desirable that we can easily use the mail-voice system even if we don't purchase a text-to-speech synthesis device and an NCU. So, it is expected that this method will promote the mail-voice system to spread. Figure 5 shows the system structure.

5.2 Information Processing in Mail-Voice

The NCU is connected to the central host computer and the text-to-speech synthesis device. It plays the role of interface to transmit information from a telephone to the computer and the device. For example, if the NCU receives a call signal, a BT signal and a PB signal from a telephone, it converts the signal data to code data and sends them to the computer. If the NCU receives some commands from the computer, it does such things as automatical calling and turning off the call.

Furthermore, the NCU has a function to send vocal signals generated by the text-to-speech synthesis device to the telephone. The text-to-speech synthesis device synthesizes a speech from phonetic and prosodic data generated by the computer from Japanese sentences including Kanji and Kana characters. The speech, i.e. vocal signals, is sent to the NCU, and we can listen to the content of Japanese sentences through a telephone.

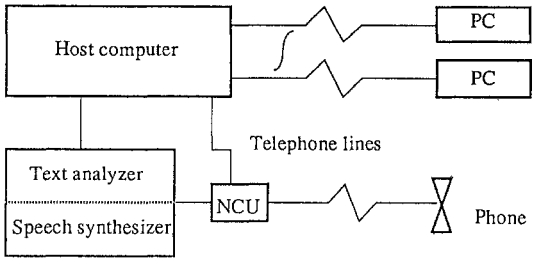

Fig. 5 Mail-Voice System

Using a personal computer connected to the central host computer through a telephone line, we can send some mails to an other person and retrieve some mails sent from an other person and store some mails in the computer. The central host computer manages and controls all of the network service. In such a system, we can listen to the content of some mails and coded information stored in the computer through a telephone. This system has many characteristics as follows,

(1) The text-to-speech synthesis device enables us to listen to the content of some mails sent from an other person and the content of coded data stored in the host computer through a telephone.
(2) If we do not have a personal computer, we can listen to the content of many kinds of information data stored in the host computer.
(3) Any speech can be generated by the text-to-speech synthesis device, so we can easily change vocal data for any guidance without recording a human voice speech.
(4) A speech for guidance can be synthesized by female CV-syllable parameters and a speech for an electronic mail can be synthesized by male CV-syllable parameters so that each part can be clearly distinguished.

CONCLUSION

A Japanese text-to-speech synthesis system for an arbitrary text has been developed. Particularly, the development of a dictionary with a large vocabulary and rules for various cases has realized high quality speech. However, there still remains the subject to estimate the correct reading of Kanji words that have plural readings. To solve this problem, more detailed semantic analysis or discourse understanding must be adopted.

Moreover, how to recognize the speaker's feelings and to apply them to speech synthesis are important subjects. For instance, rules to increase the pitch of the last syllable for an interrogative sentence, and to increase the amplitude of the sound for a focused word would be desirable. Studies on such subjects and other applications will be discussed in the future.

REFERENCES

[1] "Meikai Nihongo Akusento Jiten (Japanese Accent Dictionary)", Sanseido, 1981.
[2] "Nihongo Hatsuon Akusento Jiten (Japanese Pronunciation and Accent Dictionary)", Japan Broadcasting Publishing Company Ltd., 1985.
[3]Imai, "Log Magnitude Approximation (LMA) Filter", The Transactions of The IEICE, vol.J62-A No.12, 1980.
[4]Hara and N.Saito,"Implementation of a Rule Based Speech Synthesis Board for Laptop Computers",1990 Spring National Convention,The IEICE ,1990.
[5]N.Saito and Hara,"A Speech Synthesis-by-Rule System for Telephone Information Services",1990 Spring National Convention, The IEICE ,1990.