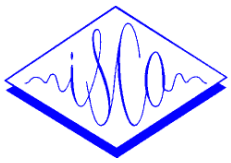

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Japanese Text-to-Speech Conversion System

Hiroki Kamanaka, Takashi Yazu, Keiichi Chihara and Makoto Morito

Human Interface Laboratory, OKI Electric Ind. Co., Ltd.
550-5 Higashi-asakawa-cho, Hachioji-shi, Tokyo, 193 Japan

Abstract

This paper describes a newly developed Japanese text-to-speech conversion system. The phonetic and prosodic information for speech synthesis is extracted from the input text by the Japanese text analyzer so that Kanji and Kana texts can be input directly to the system. The synthetic speech is generated by concatenating speech synthesis units which contain symmetrical waveforms. As synthesis units, VCV syllables are used as well as CV syllables and symmetrical waveforms have been obtained by PSE analysis. Therefore, the synthetic speech sounds smooth and natural. The system has been built on a compact board which can be housed in a PC slot.

1. Introduction

So far, we have already developed a speech synthesis method using symmetrical waveform [1]. In this method, each one-pitch symmetrical waveform is shifted for the pitch period and added to the preceding waveform to synthesize continuous speech. Symmetrical waveforms are generated as follows: first, a spectral envelope is extracted per frame from the original speech by the frame synchronous improved cepstral method, and then a symmetrical waveform is obtained from the spectral envelope by zero-phasing and inverse Fourier transformation. This method has several features such as high clearness of synthetic speech and good ability to control the pitch period. Using this method, we have also built a rule-based speech synthesis system in which only CV syllables are used as speech synthesis units [2].

This system has a few problems. Two of them are about the quality of the synthesized speech. It is true that the synthesized speech sounds clear, but it does not sound smooth. And it does not sound natural but artificial. The former is because the representation of co-articulation from vowel to consonant is not good enough, and the latter is because the extraction of spectral envelopes is not accurate. The other problems are concerning the form of the input text. Though the system is designed for an unlimited Japanese vocabulary, it can accept only Kana character strings and not Kanji character strings and it cannot synthesize sentences but only words (person's names).

Recently, we have made some improvements to the system in order to solve these problems and have built a Japanese text-to-speech conversion system on a speech processing board. The improvements are as follows: First, we have introduced not only CV syllables but also VCV syllables as speech synthesis units to make the synthetic speech smooth. Second, we have adopted the power spectrum envelope (PSE) method as a new method of envelope extraction instead of the improved cepstral method to make the synthetic speech more natural. Third, we have added a Japanese text analyzer to the system to pronounce Kanji characters, accentuate words and intonate sentences.

In this paper, the spectral envelope extraction methods, the text-to-speech conversion process, the hardware features and the result of a system evaluation are described.

2. Spectral Envelope Extraction

Speech segment data for speech synthesis (i.e., symmetrical waveforms in our system) are obtained by analyzing the original speech. Therefore, the analysis method is important to improve the quality of the synthesized speech. In this section, we compare four kinds of analysis methods :

(a) frame synchronous improved cepstrum analysis.
(b) pitch-pair synchronous improved cepstrum analysis.
(c) frame synchronous PSE analysis.
(d) pitch-pair synchronous PSE analysis.

In the frame synchronous analysis, the analysis window for short term power spectrum is shifted every fixed frame period. The window length is fixed. On the other hand, in the pitch-pair synchronous analysis, the window is shifted synchronizing its center with the center of two adjacent pitch peaks. The window length is decided by the pitch period (i.e., it is 3.2 times as long as the pitch period).

The PSE analysis is a method of spectral envelope extraction as well as the improved cepstrum analysis [3],[4]. This method is based on sampling the logarithmic power spectrum at the positions of whole number multiples of the pitch frequency. If the sampled data are not at the peak of the spectrum, they are replaced by the maximum value which is obtained out of ten neighboring points on the spectrum.

Figure 1 shows some examples of the spectral envelope analyzed by the four methods (a)-(d). In the frame synchronous analysis, the fine structure of the FFT spectrum becomes disturbed above 4 KHz. Therefore, the envelope above 4 KHz is not reliable, especially in the

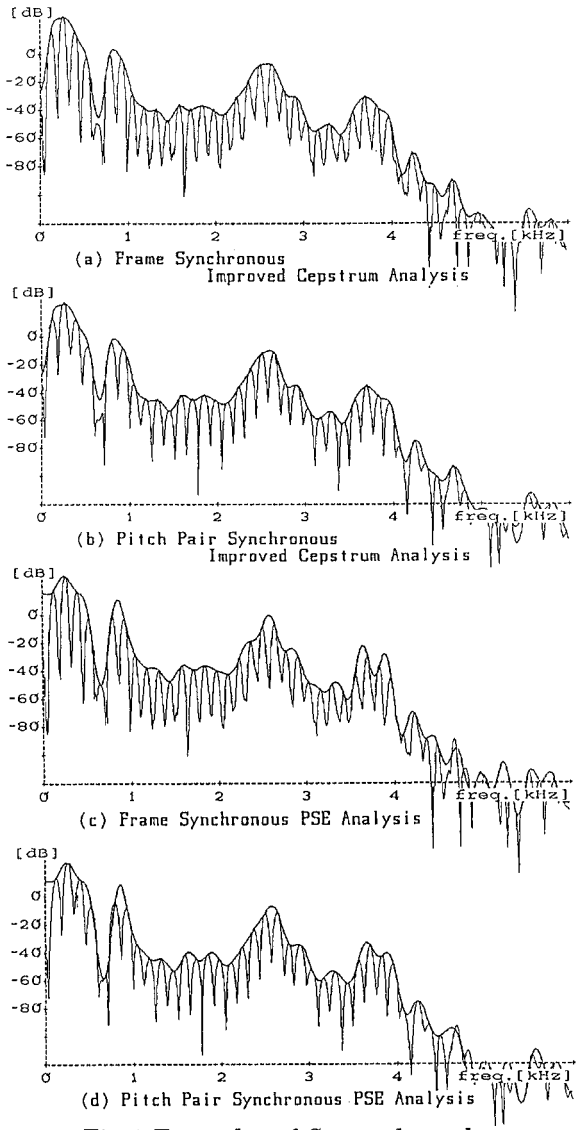

Fig.1 Examples of Spectral envelope

PSE analysis which uses only the peak data of the spectrum. On the other hand, the FFT spectrum is relatively stable in the pitch-pair synchronous analysis. The improved cepstrum analysis and the PSE analysis are not very different so far as the envelope traces the spectral peaks correctly. But they differ in the characteristics of the envelope between the spectral peaks. The improved cepstrum analysis connects the spectral peaks nearly straight on the envelope. On the other hand, the PSE analysis emphasizes the formants (i.e., the peaks of the envelope) between the spectral peaks.

3. Text-to-Speech Conversion System

Figure 2 outlines the process of the Japanese text-to-speech conversion system. The system consists of three parts : the Japanese text analyzer, the synthesis data generator and the speech synthesizer. According to Figure 2, each part is explained in order as follows :

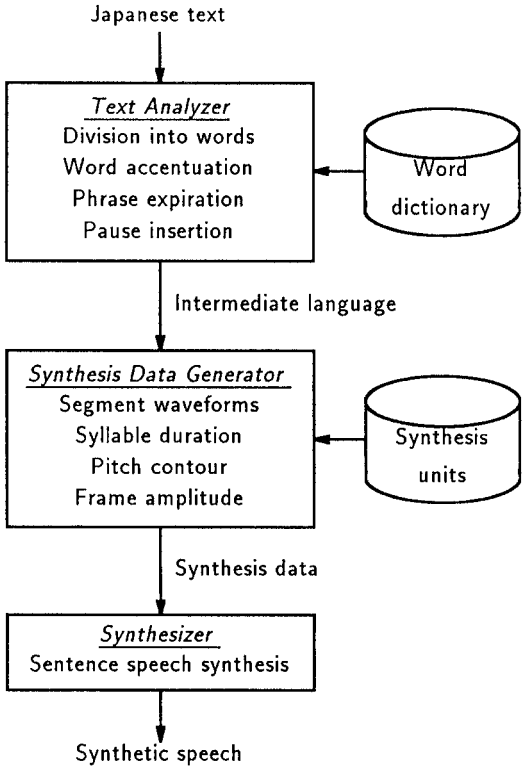

Fig.2 Japanese Text-to-Speech Conversion System

3.1 Japanese Text Analyzer

Since the pronunciation and accentuation of a Japanese text cannot be decided automatically from its spelling only, text analysis is necessary for text-to-speech synthesis. This part analyzes the input text which consists of Kanji and Kana characters. And it outputs the intermediate language which describes phonetic and prosodic information of the input text. Figure 3 is an example of the output. The intermediate language is sent to the next part : the synthesis data generator.

Input Japanese Text :
任意の文章を音声合成します。
(It converts arbitrary text to speech.)
Output Intermediate Language :
P1ni]*ino/bu}*syooo/P3o*seego}oseesimasuP0.

Fig.3 An Example of Intermediate Language

3.1.1 Division into Words

Japanese words are not spaced out in a written sentence, which is different from a language such as English. First, the input text is divided into words by referring to the word dictionary for the longest matching word under constraint of word connection rule [5]. The dictionary has about 60,000 words and the phoneme sequence, accent type and grammatical type of each word have been registered inside.

3.1.2 Accentuation

In Japanese speech, word accents are decided according to the accent type of the words. However, syntactical concatenation of words makes word accents move or appear or disappear [6]. In this system, such phenomena have been described by accentuation rules. These rules are applied to prefixes, suffixes and compound words.

3.1.3 Expiration and Pause Insertion

Expiration and pause insertion are based upon the syntactical and semantical structure of the sentence. In this system, however, they are decided by simple rules. The expiration strength of a phrase is determined according to the syllable number of the phrase. Pauses are put at punctuation marks such as a period, comma and parenthesis, and its length is determined according to the mark type. Besides, pauses of proper length are inserted between phrases.

3.1.4 Processing of Unregistered Words

It is probable that the input text includes some words which have not been registered in the dictionary. In this system, unregistered words are extracted from the text by detecting the change of character types such as Kanji and Kana. And they are given a proper reading and accent type according to the character type.

3.2 Synthesis Data Generator

Referring to the intermediate language, this part generates synthesis data such as segmental waveforms, duration, pitch pattern and amplitude pattern. And it sends these data to the synthesizer.

3.2.1 Speech Synthesis Units

The former speech synthesis system uses only CV syllables as speech synthesis units [2], and its synthetic speech does not sound smooth because CV units do not have the transition part from vowel to consonant. Therefore, this new system uses not only CV units but also VCV units which include the VC transition part.

3.2.2 Speech Segment Waveforms

First, the intermediate language is converted to the code sequence of synthesis units such as CV and VCV. Then the speech segment waveforms of each unit are continuously obtained from the data ROM according to the address sequence on the synthesis unit table. Because of using address sequences, the waveforms of each unit need not be successively recorded in the ROM, and moreover, they can be shared among different units. The synthesis unit table is storing not only the address sequence but also the classification of voiced/unvoiced sounds for the consonant part, the pause length preceding a plosive and the information necessary to control the pitch, amplitude and duration in a syllable.

3.2.3 Duration Control

In Japanese speech, the rhythmical timing points locate near the vowel onset of each syllable and they are basically isochronic [7]. And the change of duration in a CV syllable is greater in the vowel part than in the consonant part. In this system, therefore, not the consonant duration but the vowel duration is arranged to make the interval of timing points constant. The interval length is determined according to the syllable number of the phrase. The timing point location in a vowel and the fixed intrinsic length of a constant have been recorded in the synthesis unit table.

3.2.4 Pitch Control

If the pitch contour is represented by a logarithm pattern of the fundamental frequency along the time axis, it can be approximated by the sum of phrase components and accent components. Making use of this nature, the Fujisaki model has been proposed as a parametric model of the pitch contour [8]. This model consists of two functions : the impulse response function for phrase components and the step response function for accent components. When a phrase command or an accent command is input to the model, each component is generated by the function mentioned above, and the sum of all components (including formerly generated ones) is output as the pitch pattern.

In this system, the pitch pattern is generated by the Fujisaki model according to the phrase symbols and the accent symbols in the intermediate language. The phrase symbols and the accent symbols give the respective commands to the model. The magnitude of the commands is decided by the type of the symbols (phrase symbol – four types, accent symbol – two types). The timing point of the commands is set around the beginning or end of a vowel based upon some simple rules.

3.2.5 Amplitude Control

The amplitude control in this system has two steps. First, the amplitude level of each syllable in a phrase is determined by the exponential attenuation function according to the syllable number of the phrase and the syllable position in the phrase. Then the amplitude value of each frame in a synthesis unit is decided by the intrinsic amplitude pattern of the unit and the amplitude level of the syllable where the frame belongs. The intrinsic amplitude pattern of each unit has been normalized and stored in the synthesis unit table.

3.3 Speech Synthesizer

Speech segment waveforms are decoded by the ADPCM method, which is of one memory type [9]. If a decoded waveform is a voiced sound, the symmetrical component is generated by folding the waveform. One kind of waveform is obtained per frame. Waveforms in a border frame of synthesis units are interpolated between the preceding and following frames. Each waveform is given a proper power determined by the amplitude information, and is added to the preceding waveform at the interval of the pitch length.

4. Hardware Features

The Japanese text-to-speech conversion system has been built upon a general-purpose speech processing board. Table 1 shows the system specifications. The processing of the text analyzer and the synthesis data generator is performed by a 16-bit μ-CPU (Motorola, Inc. MC68000). The speech synthesizer uses a floating point DSP. The synthesis data are sent from the CPU to the DSP through the common RAM. The board is connected with a host machine such as a personal computer by the parallel/serial interface, and it can be housed in the expansion slot of a personal computer.

Table.1 System Specifications

Synthesis method	Using symmetrical waveforms
Sampling freq.	12KHz (frame period : 8ms)
Synthesis units	CV / VCV (800Kbytes)
Word dictionary	ab. 60,000 words (800Kbytes)
Input characters	Shift-JIS code
CPU	MC68000 (10MHz)
DSP	MSM699210 (10MHz)
ROM	2Mbytes
RAM	1Mbytes
Interface	Parallel / Serial (RS-232C)
Speech I/O	12-bit AD/DA
Board size	120(W) × 220(D) mm

5. System Evaluation

As an evaluation of the new system, we examined by a listening test how much the PSE analysis and VCV units improve the synthetic speech quality. Synthetic speech for the test was generated by five kinds of methods :

[A] frame synchronous improved cepstrum analysis and only CV units (i.e., the method of the former system).

[B] frame synchronous improved cepstrum analysis and CV/VCV units.

[C] pitch-pair synchronous improved cepstrum analysis and CV/VCV units.

[D] frame synchronous PSE analysis and CV/VCV units.

[E] pitch-pair synchronous PSE analysis and CV/VCV units (i.e., the method of the new system).

The test was executed by the paired comparison method and given to 11 listeners (including 5 researchers in speech processing).

Figure 4 is the result of the preference test. There is wide difference in the preference score between with and without VCV units (i.e., between [B]-[E] and [A]). This means that the arrangement of speech synthesis units has a great effect on the synthetic speech quality. The result also shows the following. The PSE analysis is preferred to the improved cepstrum analysis, especially in combination with the pitch-pair synchronous analysis (method [E]). This meets our expectation. On the other hand, the improved cepstrum analysis is preferred by non-researchers when it is combined with the frame synchronous analysis (method [B]). We think this is because the speech by the method [B] is regarded as rather smooth than dull by non-researchers who are not familiar with the synthesized speech.

6. Conclusion

We have improved the former speech synthesis system and have developed the Japanese text-to-speech conversion system which is built onto one board. This system has the following features :

1) We can directly input a Kanji and Kana text to the system.
2) The synthetic speech sounds more smooth and natural than the former one.

We are going to evaluate this system in various ways in order to better it.

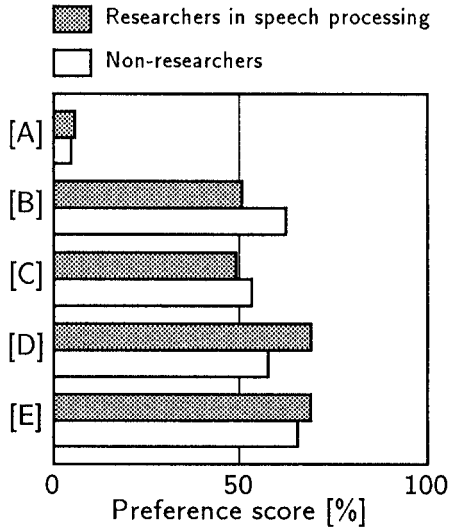

Fig.4 Result of the Preference Test

Acknowledgment

The authors wish to thank Mr. Katsuhisa Saito (director of the laboratory) and Mr. Isamu Nose (head of the division) for their advice and encouragement during the work.

References

[1] T. Yazu, K. Miki, M. Morito and K. Yamada, "Speech Synthesis Method Using Symmetrical Waveform" (In Japanese), Trans. of the Committee on Speech Research, Acoust. Soc. Jpn., S83-67 (1984)

[2] T. Yazu and K. Yamada, "The Speech Synthesis System for an Unlimited Japanese Vocabulary," Proc. IEEE-IECEJ-ASJ ICASSP, Tokyo, pp.2019-2022 (1986)

[3] T. Nakajima and T. Suzuki, "Power Spectrum Envelope (PSE) Analysis Based on Pitch Frequency Interval Sampling on Short Term Power Spectrum," Proc. 1st Symp. Advanced Man-Machine Interface Through Spoken Language, pp.155-164 (1988)

[4] S. Imai and Y. Abe, "Spectral Envelope Extraction by Improved Cepstral Method" (In Japanese), IECE Trans., Vol.J62-A, No.4, pp.217-223 (1979)

[5] S. Sagayama and K. Kogure, "Japanese Text Analysis for Speech Synthesis" (In Japanese), Trans. of the Committee on Speech Research, Acoust. Soc. Jpn., S82-78 (1983)

[6] H. Sato, "Text-to-Speech Transformation Techniques" (In Japanese), Jour. of IECE Jpn., Vol.70, No.4, pp.373-378 (1987)

[7] H. Sato, "Segmental Duration and Timing Location in Speech" (In Japanese), Trans. of the Committee on Speech Research, Acoust. Soc. Jpn., S77-31 (1977)

[8] H. Fujisaki and S. Nagashima, "A Model for Synthesis of Pitch Contours of Connected Speech," Annu. Rep. of Eng. Res. Inst., Univ. Tokyo, 28, pp.53-60 (1969)

[9] N. S. Jayant, "Adaptive Quantization with a One-Word Memory," B. S. T. J., Vol.52, No.7, 1-119 (1973)