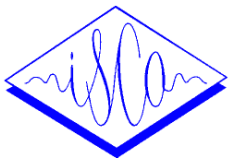

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

# A NEW JAPANESE TEXT-TO-SPEECH SYNTHESIZER BASED ON COC SYNTHESIS METHOD

*Kazuo Hakoda, Shin-ya Nakajima, Tomohisa Hirokawa and Hideyuki Mizuno*

NTT Human Interface Laboratories
Yokosuka-shi, Kanagawa 238-03, JAPAN

***ABSTRUCT***

*This paper describes a new Japanese text-to-speech synthesizer that produces far more natural and intelligible speech than existing synthesizers by using the new Context Oriented Clustering(COC) method. The COC method automatically generates speech unit variations from natural speech database. Preference tests show that the intelligibility of COC synthesized speech is better than that of the conventional dyad based method. A new LSP synthesizer which produces a wide frequency band of output speech is developed. The synthesizer is implemented with a general purpose Digital Signal Processor(DSP). Optimum design parameters, such as LSP order, parameter quantization bits are decided on the basis of spectral distortion and preference tests results. This synthesizer is constructed on a single PC board to permit easy installation in personal computers.*

## 1. INTRODUCTION

In 1987, we developed a compact Japanese text-to-speech synthesizer which converts arbitrary Japanese text consisting of Kanji and Kana characters to continuous speech[1]. The synthesizer has been used in several practical applications: such as a revision support system for newspapers and a building equipment control system[2]. The synthesized speech quality was not considered sufficiently natural and further improvement was needed.

Recently we proposed a new speech synthesis method which produces more natural and intelligible speech than conventional methods. Our 1987 prototype synthesizer used dyad units(CV,VC) as the speech units. However, some units produced unnatural speech because they failed to cover the many variations that occur in connected speech. Our new method automatically generates speech unit variations from a natural speech database by using Context Oriented Clustering(COC) method[3]. This paper describes the COC algorithm briefly, and examines the speech quality.

With recent improvements in hardware technology, speech output over a wide frequency can be produced from a LSP synthesizer by using a general purpose high-speed Digital Signal Processor(DSP). This paper describes the effect of frequency band expansion on speech quality.

By using the new COC method and a wide frequency LSP synthesizer, we constructed a text-to-speech synthesizer on a single PC board which can be easily added to a personal computer. This paper describes its features, and hardware configuration.

## 2. COC SYNTHESIS METHOD

### 2.1 Method

Conventional speech synthesis methods use phonemes, diphones, demisyllables, or syllables as speech units. However, these units are affected by phoneme context such as preceding and succeeding phonemes, because of the co-articulation phenomena. It is difficult to decide the most efficient and appropriate speech unit. Speech units used in previous text-to-speech synthesizers were decided experimentally or on an ad hoc basis, and relied on the researcher's phonological knowledge. These units, therefore, often failed to cover the many variations that occur in connected speech, and so produced unnatural speech.

To overcome these problems, allophones which are necessary for the representation of co-articulation phenomena, must be used as speech units. The key point is that they are decided automatically by using a statistical clustering technique. We have termed this technique Context Oriented Clustering(COC). The COC method searches for the phoneme concatenation strings that most affect the phoneme features, based on variations of phonetic clusters[3].

### 2.2 Speech units generation

A speech database with acoustic phonetic labels is used. Feature extraction analysis is performed on the speech data to obtain feature vector sequences. Using the feature vector sequences with phonetic labels, synthesis units are generated automatically through the COC procedure. The generation process is shown in Fig.1. In the COC procedure, each initial cluster is a set of segments

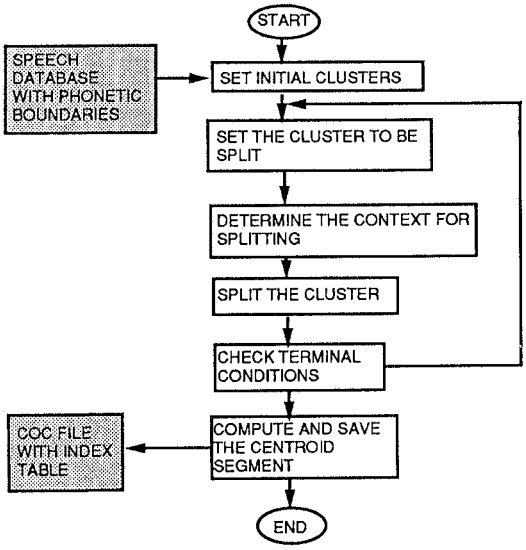


Fig.1 Speech Unit Generation based on COC Method

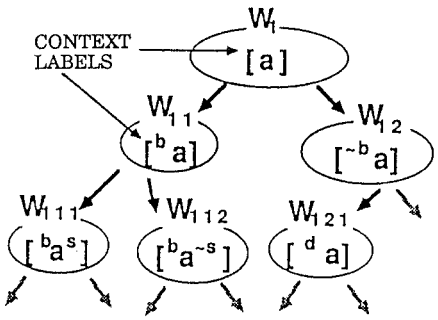


Fig.2 Cluster partitioning process in COC

with the same phonetic labels. The cluster partitioning process is applied to clusters hierarchically referring to phonetic contexts. After a number of iterations, when proper terminal conditions are satisfied, the centroid segment of each cluster is saved as a speech unit, and the corresponding phonetic context is saved as the index to determine in what context the unit should be used.

Figure 2 is an example of this process. Initial cluster $W_1$, which is a set of all segments with phonetic label /a/, is split into two clusters: $W_{11}$:/a/ preceded by /b/, and $W_{12}$:/a/ preceded by any phoneme but /b/. These two clusters can be further split into four clusters.

### 2.3 Comparison with conventional method

In order to confirm the efficiency of the COC method, the difference in speech quality between the COC method and the conventional diphone method was examined[4]. ABX preference tests were carried out. Reference X was PCM speech sample uttered by an announcer. Test sample A and B were randomly

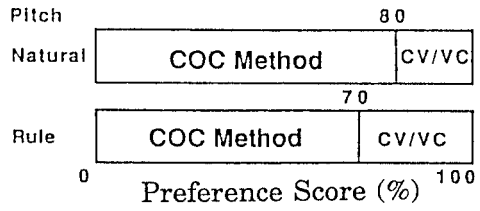


Fig.3 Results of the preference tests

selected from the COC speech and the diphone-based synthesized speech produced by concatenating CV,VC speech units. The subjects were asked to aurally judge which sample, A and B, was closer to the reference X. Two kinds of pitch pattern were used in synthesized speech: one was generated automatically by rule, the other was extracted from natural speech. Test results are shown in Fig.3. The preference score of the COC synthesized speech is 80% for the natural pitch pattern, and 70% for the rule-based pitch pattern. Therefore, the speech quality of the COC synthesized speech was confirmed to be better than that of the conventional diphone method.

## 3. TEXT-TO-SPEECH SYNTHESIS PROCESS

### 3.1 Linguistic process

The text-to-speech synthesis process consists of two stages; linguistic processsing and speech

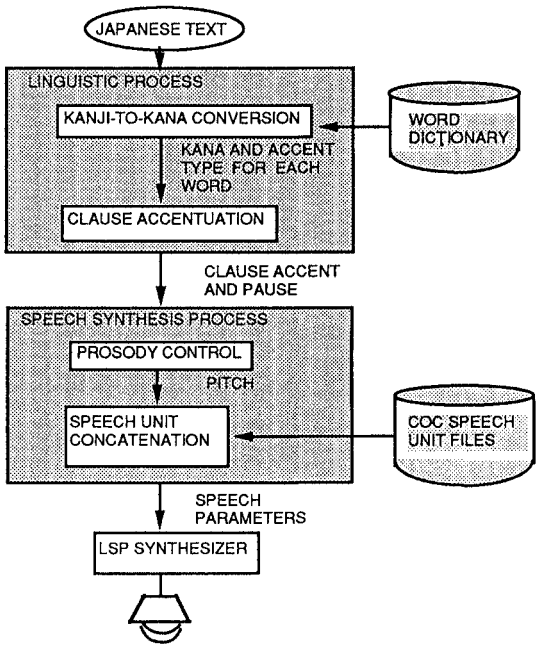


Fig.4 An outline of the Japanese text-to-speech conversion process

synthesis processing. Japanese text is composed of Kanji and Kana characters. Each Kanji character has multipule pronunciations. In the first stage of linguistic processing, words are extracted from Kanji strings using a word dictionary, and each word is given an accent type and a pronunciation with Kana notation. Next, the word accents are merged into a clause accent according to rules. Pauses of proper length are inserted at appropriate positions[1].

The dictionary used in the prototype synthesizer contained about 45,000 words with accents and Kana readings. In the new synthesizer, about 10,000 words were added to improve the pronunciation and accent of Kanji strings. Furthermore, about 500 English words were added to allow alphabet strings to be pronounced.

### 3.2 Speech synthesis process

In the second stage (speech synthesis processing), a fundamental frequency contour is produced according to rules using the clause accent[5].

In the speech unit concatenation process, given a phoneme string, those synthesis units having context labels which have the longest matching labels against a given context are selected. In our synthesizer, the length of context is limited to under 3 phonemes for preceeding and succeeding labels respectively. In this process, neither interpolation nor smoothing of speech parameters were adopted since each unit contains the essential coarticulatory phenomena and also durational characteristics. About 1,500 speech units generated automatically by the COC method were stored in the COC speech unit files with LSP and amplitude parameters[6]. Finally, these parameters are given to the LSP synthesizer to produce speech sound.

Table.1 Basic specifications

| Items | Specifications |
|---|---|
| Input Text | Kanji and Kana strings(JIS C6226 Code).<br>Kana strings with accents (JIS C6220 Code). |
| Kanji to Kana conversion and accentuation | Simplified text analysis for Kanji strings.<br>Word and clause accentuation rules. |
| Speech synthesis | *Synthesis by rule based on the COC method.<br>*Wide frequency band LSP synthesis. |
| Speech output control (control by command) | Start,stop, and restart of speech output.<br>Speech speed, loudness.<br>Male/Female speech output.<br>*Speech and signal output from LSP parameters which are stored in file memory and sent from host terminal. |
| Interface | PC-bus interface |
| Size | A single PC board |

* New functions for the new synthesizer

## 4. SYSTEM CONFIGURATION

### 4.1 Features

The text-to-speech synthesizer is controlled by a personal computer. The specifications of this synthesizer are shown in Table 1. This synthesizer has a new function for output control: speech and signal output from LSP parameters which are stored in file memory or sent from host terminals. Several prototype synthesizers have been used in telephone information services such as message passing systems[2]. However, in many cases, pre-recorded fixed messages can be used instead of text-based synthesized speech. The total quality of output speech for message services can be improved by combining a pre-recorded voice synthesized using LSP parameters and the COC synthesized speech.

### 4.2 Hardware configuration

The text-to-speech synthesizer is designed to be constructed on a single PC board equipped with a PC bus interface. This allows simple installation in a personal computer. The synthesizer is shown in Fig.5, and its hardware configuration is illustrated on Fig.6.

The synthesizer consists of a control unit, dictionary files, COC speech parameter files, and LSP speech synthesizer. The control unit includes a 16 bit microprocessor, program ROM, and work RAM. In order to improve the space factor, CPU

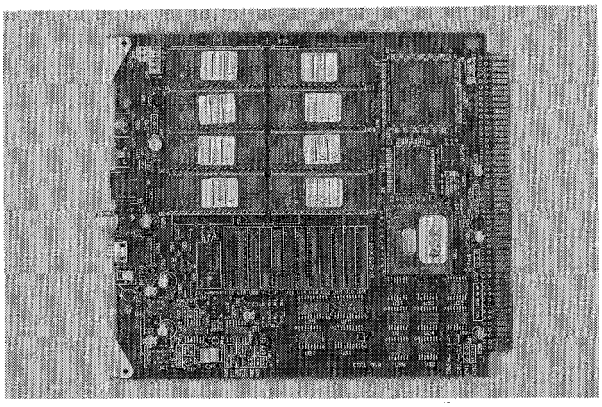

Fig.5 The new text-to-speech synthesizer

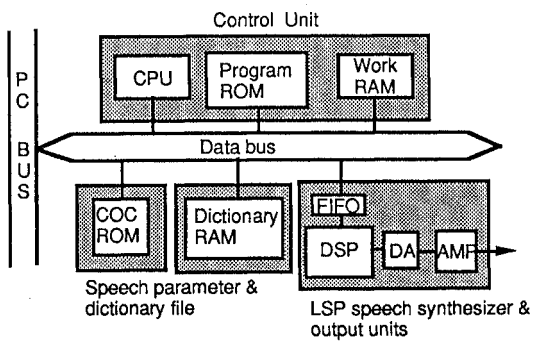


Fig.6 Hardware configuration

support circuits such as address controller are integrated into a single gate array IC, and dictionary and speech parameter files are stored in 1M high density memories. The LSP speech synthesizer is implemented on a general purpose digital signal processor(DSP). LSP parameters are sent asynchronously from the CPU to the DSP through FIFO memory.

### 4.3 LSP speech synthesizer

The new speech synthesizer was designed to improve the speech quality by expanding the frequency range from the telephone band(3.4KHz) to 5.1KHz[7]. Preference tests were carried out to examine the efficiency of this expansion. The results confirmed that the effect of expanding the frequency range to 5.1KHz is more effective to improve the intelligibility of synthesized speech than increasing the LSP parameter bit rates from 5 bits/parameter to 7 bits/parameter[6].

The basic specification of the LSP synthesizer is shown in Table 2. The LSP synthesizer was implemented on a 24-bit fixed point Digital Signal Processor. This DSP had twice the processing speed of the old-type DSP used in the prototype synthesizer. Owing to this, the LSP order for synthesized speech was increased to the 14th order for a 12 KHz sampling frequency. Optimum design parameters, such as LSP order and parameter quantization bits were decided on the basis of spectral distortion and preference test results[7].

## 5. CONCLUSION

A new synthesis method has been used to create a single PC board text-to-speech synthesizer. The new speech synthesizer has the following features.
(1) It produces natural and intelligible speech by applying a new speech synthesis method based on Context Oriented Clustering(COC).
(2) Its speech output has a wider frequency band through the use of a general purpose Digital Signal Processor(DSP).
(3) The PC board has a PC bus interface which permits easy installation in personal computers.

This synthesizer can be incorporated into a compact audio response unit (ARU) that includes a network control unit controlled by a personal computer. Various telephone information services, such as reservation and order entry, can be realized with this system.

Table.2 Basic specifications of LSP synthesizer

| Items | Specifications |
|---|---|
| Sampling Frequency | 12 KHz |
| Frequency range | 5.1 KHz |
| Frame period | Variable |
| LSP order | 14 |
| Source | Pulse, noise, and residual waves |
| Parameter quantization | LSP 5 bits<br>Pitch 8 bits<br>Amplitude 7 bits |
| DSP | 24-bits fixed point |


## ACKNOWLEDGEMENT

The authors wish to thank Dr. Sadaoki Furui, director of Speech and Acoustics Laboratory, Dr. Ryohei Nakatsu, former group leader, and Dr. Hirokazu Sato, group leader, for their encouragement during this work.

We also thank Mr. Yasuo Endo, NTT Data Communications Systems Corporation, who gave us a good suggestion for hardware construction.


## REFERENCES


[1] K. Hakoda, K. Nagakura, T. Hirahara and K. Kabeya, " Japanese Text-to-Speech Synthesizer", Journal of the American Voice I/O Society, Volume 6, pp.1-16 (1989-6)

[2] T. Hirokawa, "Application of Japanese Text-to-Speech Synthesizer", Speech Tech'89, pp.30-32 (1989)

[3] S. Nakajima and H. Hamada, "Automatic Generation of Synthesis Units Based on Context Oriented Clustering", Proc. IEEE Int. Conf. ASSP, S14.2 (April 1988)

[4] S. Nakajima and K. Hakoda, "Evaluation of Synthesized Speech based on Context Oriented Clustering", IEICE Technical Report, SP 88-128, pp.49-56 (1989-01) ( in Japanese)

[5] K. Hakoda and S. Sato, "Prosodic Rules in Connected Speech Synthesis", Systems, Computers, Controls, Scripta Electronica Japonica 3, Vol. 11, pp.28-37 (1980)

[6] N.Sugamura and F.Itakura, "Speech Data Compression by LSP Speech Analysis and Synthesis Technique", IEICE Trans, Vol. J64-A, No 8, pp.599-606 (1981) (in Japanese)

[7] K. Hakoda, T.Hirokawa and S.Nakajima, "LSP Synthesizer for Speech with Expanded Frequency Band", ASJ 90 Spring Meeting, 3-4-9, pp.285-286 (1990) ( in Japanese)