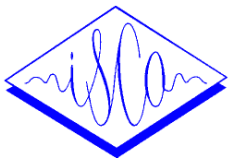
ISCA Archive
http://www.isca-speech.org/archive



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

TRIAL PRODUCTION OF A MODULE FOR SPEECH SYNTHESIS BY RULE

Mikio Yamaguchi

Information and Electronics Laboratories,
Sumitomo Electric Industries, Ltd.
1-1-3, Shimaya, Konohana-ku, Osaka, 554 Japan

ABSTRACT

A module-type speech synthesizer was developed for rule-synthesis of Japanese, in order to find practical applications of rule-synthesis devices. The module can be built into any machine and can be used almost anywhere.

Input to the module consists of *katakana*, accent marks, punctuation marks, and modification marks. The modification marks represent accent sandhi between words on both sides. The size of the module is 66x46x15 mm and it weighs 59 grams. All necessary hardware and software components are included in the module. The electrical interface is compatible with common CPU buses. Syllable articulation is 63% comprehensible. It uses at most 160 mA from two 5V power supplies.

1. INTRODUCTION

Spoken language is an effective tool for human communication, and therefore, it is thought that public demand for speech synthesis by rule will be great when used as a man-machine interface. Several types of rule-based speech synthesizers are already available on the market [1]. The potential market size is thought to be large, but wide-spread applications and significant market segments are not yet clear [2]. Two problems should be solved to encourage a wide variety of uses: (1) The quality of synthesized speech is still relatively poor, and (2) Practical applications of mass-produced synthesizers have not been shown. The author decided to focus on the latter problem first, under the assumption that if commercially-practical applications are known (or expected by many), then efforts of proper range and degree can be allocated to the former problem.

The author developed and implemented a speech-synthesis-by-rule system which has been previously reported [3][4][5][6]. However, the unit was not applicable for a wide variety of practical uses, because (1) the place of use was restricted to a desk top, and (2) the operation was restricted to use with a personal computer. From the viewpoint of application, the location of the speech output device should not be restricted. Also, a new type of speech synthesis device would be helpful when considering further applications. Furthermore, an actual device is necessary to involve engineers from other fields in order to develop new applications. Therefore, the author decided to produce and promote a module-shaped device (Photo. 1) for trial use.

This paper presents a detailed description of this module which has already been implemented. The module is capable of installation into any equipment as an output device for a built-in microprocessor. It is hoped that this practical implementation of a speech synthesis device will promote discussion about the limits of application of such devices and encourage the technological development of speech synthesis by rule for actual use.

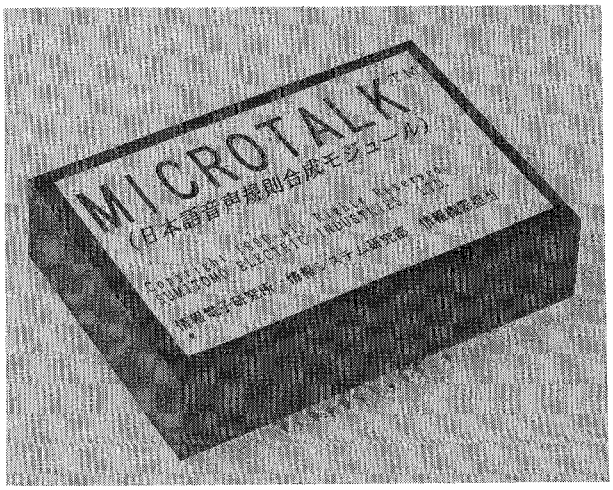


Photo. 1 The module for speech synthesis by rule (approximately life-sized).

2. SYNTHESIS PROCEDURE (Fig. 1)

2-1. THE PHONOLOGICAL PROCESS [3][6]

The phonological process accepts accent marks('), punctuation marks(。、・), modification marks(_) and *katakana* (Fig. 2). Modification marks show word boundaries of accent sandhi which should reflect syntactic structure. This process generates prosodic symbols (Table 1) and CV symbols.

(1) Generation of Pause and Phrase Symbols

Three different types of punctuation marks, which correspond to different levels of syntactic structure, are used to generate pauses. Punctuation marks are interpreted as follows:

"。" → "P0 S1 P1" (sentence boundary)
"、" → "P0 S2 P1"
"・" → "S3 P2"

P3 is used at word boundaries which do not have a modification mark and when more than 13 mora exist between neighboring phrase symbols.

(2) Generation of Accent Symbols

Accent marks show both the accent type and the accent boundary. A1 and A2 are used to show a rise in F_0, and A0 is used to show a drop in F_0. A1 is assigned for the accent type which has a rapid downward transition of the F_0 contour. A2 is assigned for the accent type which does not have a rapid downward transition of the F_0 contour. Accent sandhi rules are then applied to prosodic words which are on both sides of the modification mark (Fig. 3).

(3) Allophone Processing

Any /su/ which precedes a pause is devoiced. /i, u/ are devoiced in the following cases except in the interval from an accent rising point to an accent falling point ("high level"):

C_1=/s/ and C_2=/p, t, k/ or
C_1=/p, t, k, h/ and C_2=/s, p, t, k/
where /C_1VC_2/ (V=/i, u/).

A syllabic nasal is interpreted as [m] before [m, b, p], [ŋ] before [k, g], [N] before a pause, and [n] for all other cases.

2-2. THE ACOUSTIC PROCESS [4][6]

(1) Determination of Timing

In Japanese, each syllable is pronounced with semi-constant rhythm, and durational differences between consonants are partially compensated for with anterior and posterior vowels [7]. Each CV's stored pattern contains timing information. This information includes a time-adjustment value which represents

Table 1 Prosodic symbols and their values.

Pause		Phrase		Accent	
Symbol	Value	Symbol	Value	Symbol	Value
S1	700	P0	-0.30	A0	-A1, -A2
S2	300	P1	0.43	A1	0.40
S3	80	P2	0.26	A2	0.26
	(msec)	P3	0.12		

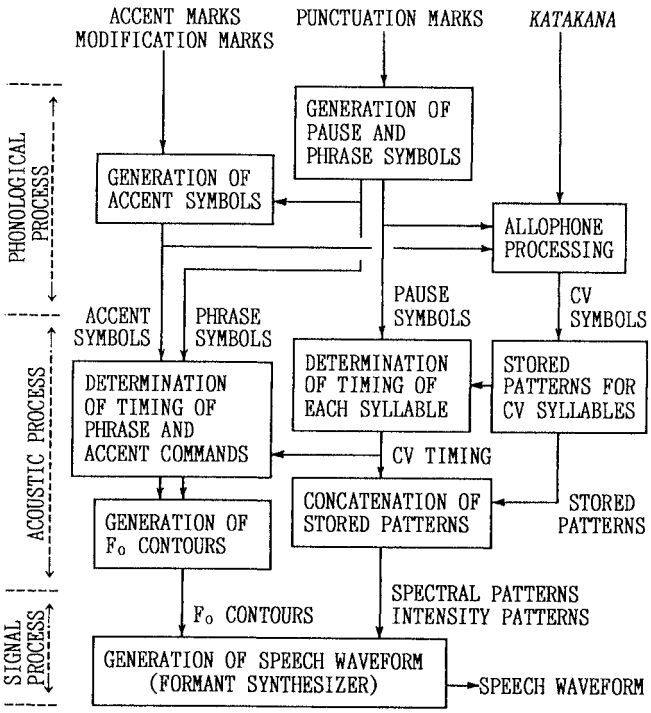


Fig. 1 Block diagram of speech synthesis by rule in the module.

テンキガ 'イキョー_デ 'ス.
ニシニホ'ンオ_オーッテイル イト'ーセーコーキ'アツワ、シダ イニ ヒガ シエ_
イト'ーシ、イツカワ、キアツノタニ'ガ ツーカスル_ミコミ_デ 'ス.
コノタメ、ア'サカラ ア'メノ_フ'ル_トコロガ _オ'ーク、ニッチユーワ
カ'クチトモ トキト'キ ア'メニ_ナ'ル_デ ショ'ー.
マタ・オンダ ンセ 'ンセンノ_ツーカニ_トモナ'ッテ、ヨ'ルワ イチ'ジ
テ'ンキガ _カイフクスル_デ ショ'ー.
ウ'ミデ ワ、トコロト'コロ コ'イ_キリノ_タメ ミトーシカ' _ワ'ルク、
タショー ナミ'ガ _ア'ル_デ ショ'ー.
オオサカ'フ、キョ'ー・ニッチューワ ヒガ シノ_カセ'、クモ'リ トキト'キ_ア'メ.
ヨ'ルワ ミナミノ_カセ'ガ _ツ'ヨク、クモ'リ トキト'キ_ハ'レ.
トコロニ_ヨッテ'ワ、ニワカ'アメガ _フ'ル_デ ショ'ー.
ウ'ミデ ワ、タショー ナミ'ガ _ア'ル_デ ショ'ー.
アス'ワ、ゼ ンバ ンニ、クモ'リ イチ'ジ _ハレ'ノ_ミコミ_デ 'ス.

Fig. 2 An example of input sentences for the module.

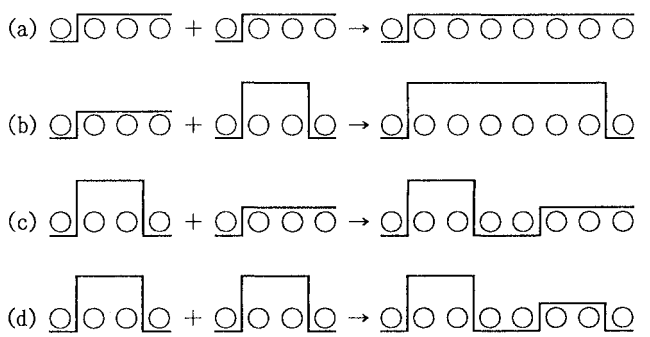


Fig. 3 Accent sandhi rules in the case of four-mora words.
(a) A2+A2 (b) A2+A1 (c) A1+A2 (d) A1+A1

that CV's deviation from the average CV time length and also includes a mora-synchronous timepoint. The timing of each CV is determined by adding the CV's time-adjustment value and pause length to the mora time interval.

The timing of the F_0 command (the onset or end of an accent command or the onset of a phrase command) is determined relative to the onset of the vowel in each CV. The offset values are -210 msec for P1, -80 msec for P0, P2 and P3, and -70 msec for A0, A1 and A2 [8][9].

(2) Generating F_0 Contours

The values from Table 1 are assigned to F_0 commands. F_0 contours are generated with the Fujisaki model as the sum of the phrase components and the accent components, which are the impulse response and step response of a 2nd-order critical damping system in the logarithmic scale, respectively [10].

(3) Generating Segmental Features

Formant parameters and intensity parameters are generated by concatenating stored patterns for each CV as segmental features. Stored patterns consist of 3 portions: a connection part, a fixed part, and a flexible part. Interpolation between CVs is done in the connection part (Fig. 4). S-curve interpolation is generally done for formant frequencies, but F2-F3 crossing interpolation is done for the combination of front vowel and back vowel. Linear interpolation is used for other parameters (formant bandwidths and intensities). The fixed part is not influenced by the CVs on either side. In the flexible part, the end value of the fixed part is extended to the beginning of the connection part of the next CV.

Intensity values of the voicing source and the fricative source are decreased by 5 dB in the 500 msec at the end of each sentence as a suprasegmental feature.

2-3. THE SIGNAL PROCESS

A Klatt-type formant synthesizer is used [11]. The variable parameters of the (anti-) resonaters are F_1, B_1, F_2, B_2, F_3, B_3, F_4 and FZ (frequency of the nasal zero). The parameter values are updated pitch-synchronously.

3. IMPLEMENTATION

All components for the above procedures are included in the module's package (Fig. 5). It operates with two 5V power supplies for the digital and analog circuits. The significant functional units are the following:

CPU(16bit, 10MHz), ROM(128KB), RAM(32KB):

The phonological process and acoustic process are performed by the CPU using the ROM and RAM. The program size is 19 KB. The data size in ROM is 63.2 KB for CV stored patterns and other data.

RAM is for both CPU operation and parameter transfer from the CPU to the DSP.

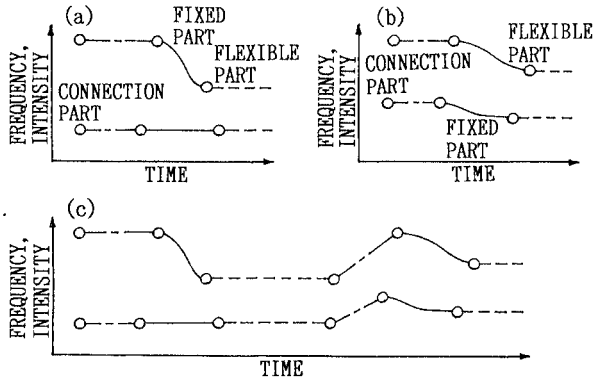


Fig. 4 Data structure of CV syllable patterns and the method for their concatenation.
(a), (b) CV syllable patterns
(c) After concatenation of (a) and (b)

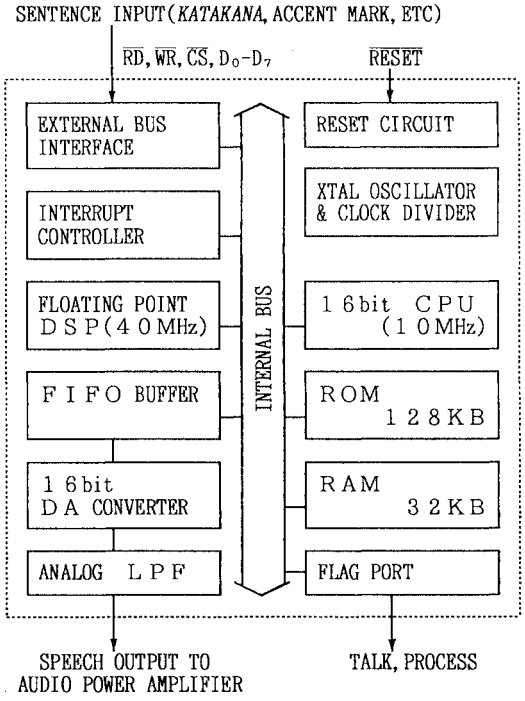


Fig. 5 Internal hardware block diagram of the module.

DSP(Floating Point type, 40MHz):

The signal process is performed by the DSP. The DSP also oversamples in order to decrease the scale of the analog LPF. The DSP usually calculates using its own local on-chip RAM and ROM, but acquires the internal bus from the CPU when it reads parameters from RAM and sends the waveform to the DAC's FIFO buffer.

XTAL OSCILLATOR AND CLOCK DIVIDER:

These generate and supply clock

signals to the CPU, DSP and DAC.

FIFO BUFFER:

The DSP generates 4 samples of waveform data during one operation cycle of the synthesizer by oversampling. The FIFO buffer regulates the sampling interval for the DAC.

ANALOG LPF:

The anti-aliasing LPF consists of a 1st-order passive filter and a 2nd-order active filter with one operational amplifier.

RESET CIRCUIT:

This generates an internal reset signal from an external reset signal and at power-up time.

EXTERNAL BUS INTERFACE:

This supplies $\overline{RD}$, $\overline{WR}$, $\overline{CS}$, D0-D7 and ACK (request next $\overline{WR}$ operation) signals similar to a common peripheral LSI of a microprocessor system.

FLAG PORT:

This shows synthesis activity with the PROCESS flag and TALK flag.

PACKAGE:

A black plastic box is used to help other engineers to think of speech synthesis by rule as just a "black box" for speech output (Photo. 1). It has 24 dual-in-line type pins for electrical connection.

4. SUMMARY

This general-purpose module was produced with current hardware technology. Syllable articulation was 63% comprehensible. Other specifications are shown in Table 2. This module is capable of installation into any equipment as an output device for a microprocessor, and can be used in a variety of situations.

It is hoped that this module will promote discussion about possible applications and encourage technological developments of speech synthesis by rule for actual use. This module is under survey for a new field of application in the author's company.

Table 2 Brief specification of the module.

ITEM	CONTENTS
TYPE OF SYNTHESIS	FORMANT CV
NUMBER OF CV SYLLABLES	119
F_0 CONTOUR MODEL	FUJISAKI MODEL
SIMULATED FREQ. RANGE	~5 kHz
FRAME PERIOD	10 msec
SPEECH RATE	7 mora/sec
VOICE TYPE	MALE
SIZE	66x46x15 mm
WEIGHT	59 grams
POWER SUPPLY VOLTAGE	5V
POWER SUPPLY CURRENT	100-160 mA (IN SYNTHESIS) 60 mA (STAND-BY)

ACKNOWLEDGEMENTS

The author would like to thank his managers and colleagues for valuable discussions about speech synthesis by rule from the viewpoint of industry.

REFERENCES

[1] M. Kato, "Speech Synthesizer by Rule, Re-entering the Market with Improved Usability and Economy," Nikkei Electronics, Nikkei Business Publications, Inc., 1988.11.28 No.461, pp.171-176, 1988.

[2] H. Sato, "The Present and the Future of Japanese Text-to-Speech Technologies in View of Their Application," Reports of Spring Meeting, Acoust. Soc. Japan, 1-4-2, 1990.

[3] K. Hirose, H. Fujisaki, M. Yamaguchi and M. Yokoo, "Synthesis of Fundamental Frequency Contours of Japanese Sentences Based on Syntactic Structure," Trans. of the Committee on Speech Research, Acoust. Soc. Japan, S83-70, 1984.

[4] K. Hirose, H. Fujisaki, M. Yamaguchi, H. Minemura and M. Yokoo, "Synthesis of Segmental Features in the System for Connected Speech Synthesis," Reports of Spring Meeting, Acoust. Soc. Japan, 2-2-10, 1984.

[5] M. Yamaguchi, H. Fujisaki, K. Hirose and H. Kawai, "Trial Production of Speech Synthesizer by Rule Using Floating Point DSP," Reports of Autumn Meeting, Acoust. Soc. Japan, 2-2-3, 1988.

[6] K. Hirose, H. Fujisaki, H. Kawai and M. Yamaguchi, "Speech Synthesis of Sentences Based on a Model of Frequency Contour Generation," Trans. of Institute of Electronics, Information and Communication Engineering, Vol. J72-A, No.1, pp.32-40, 1989.

[7] N. Higuchi, "A Study on the Segmental Duration in Connected Speech of Japanese," Doctral Dissertation, Faculty of Engineering, University of Tokyo, 1982.

[8] K. Hirose, N. Noboru and H. Fujisaki, "Analysis, Synthesis, and Perception of Fundamental Frequency Contours in Spoken Sentences — An Investigation on Phrase Components —," Trans. of the Comittee on Speech Research, Acoust. Soc. Japan, S81-36, 1981.

[9] H. Fujisaki and K. Hirose, "Fundamental Frequency Control in Spoken Sentences," Trans. of the Comittee on Speech Research, Acoust. Soc. Japan, S80-73, 1980.

[10] K. Hirose and H. Fujisaki, "Analysis and Synthesis of Voice Fundamental Frequency Contours of Spoken Sentences," Proc. 1982 IEEE International Conf. Acoustics, Speech, and Signal Processing, Paris, pp.950-953, 1982.

[11] D. Klatt, "Software for a Cascade/Parallel Formant Synthesizer," J. Acoust. Soc. Am., Vol.67, No.3, pp.971-995, 1980.