STUDIA PHONOLOGICA V (1970)

On Line Speech Synthesis through I/O Interface Unit

Toshiyuki SAKAI, Shūji DŌSHITA,
Makoto NAGAO and Kenji ŌTANI

SUMMARY

This paper describes the on-line use of the computer, especially the application to speech synthesis. On line use of the computer with other kinds of equipments is a new trend in computer application, by which computer extends its ability to all the possible information processing fields. To the central processing unit (CPU) of the computer can be attached several kinds of equipments such as magnetic tape, disk, drum, console etc. as standard equipments. The interface between CPU and these equipments are usually designed specifically to meet the requirement of each case. The same interface can not be connected to several different equipments. Therefore, when a new equipment is to be connected to the computer, we must design a new interface for it.

Considering these situations, it seems an essential problem to design a universal type interface which is independent from the specific structures of equipments to be connected and through which we can connect a variety of equipments to the computer.

The PART I of this paper presents the universal I/O interface unit (peripheral adaptor) designed so as to connect any kinds of data processing equipments to NEAC-2200 series computers. In the PART II of this paper the speech synthesizer which is connected in on-line mode to the computer through this peripheral adaptor is described. The synthesizer uses the zero-crossing wave as the wave element with which speech sound is constructed.

PART I

INTRODUCTION

The developments of computer technology such as speed, memory capacity, mass file and new I/O devices have made more and more easy the application of computer to a variety of non-arithmetic, information processing systems. In these applications it is essential that the computer, making a large complex system, works on line with other information processing systems and human beings. In

Toshiyuki Sakai (坂下利之) Professor, Department of Information Science, Kyoto University
Shuji Doshita (堂下修司) Assistant professor, Department of Physical Electronics, Tokyo Institute of Technology
Makoto Nagao (長尾真) Assistant professor, Department of Electrical Eng. Kyoto University
Kenji Otani (大谷謙治) Graduate course, Department of Electrical Eng. Kyoto University

man-machine cooperation system, logical interruption both in programming and execution stages is achieved by on line console, which can provide the user with the efficient data exchange between man and the computer. The manner of cooperation of the computer with other kinds of information processing systems is far complex, depending on the purpose of the system, and requiring the different interface characteristics for each case. It is, however, desirable that the computer has some standard interface applicable to any of these requirements.

The I/O interface unit (peripheral adaptor) presented in this part was designed to connect to the computer a variety of systems whcih arise in experiments of information processing and other laboratory researches. Through this adaptor the user can match the computer with his own devices. The adaptor works like the control device of the conventional computer peripheral devices, but the concrete function is decided by the structure of each peripheral equipment to be attached and by the program.

The Computer System

The peripheral adaptor was designed to work with NEAC 2200 series computer of Nippon Electric Co. Ltd.. The configuration of the present system is illustrated in Fig. 1.1. The central processor unit (CPU) is NEAC 2200 series model 200, a medium size computer, 16 K character core memory, three read-write channels and eight I/O trunks are attached to it. NEAC 2200 is a character machine,

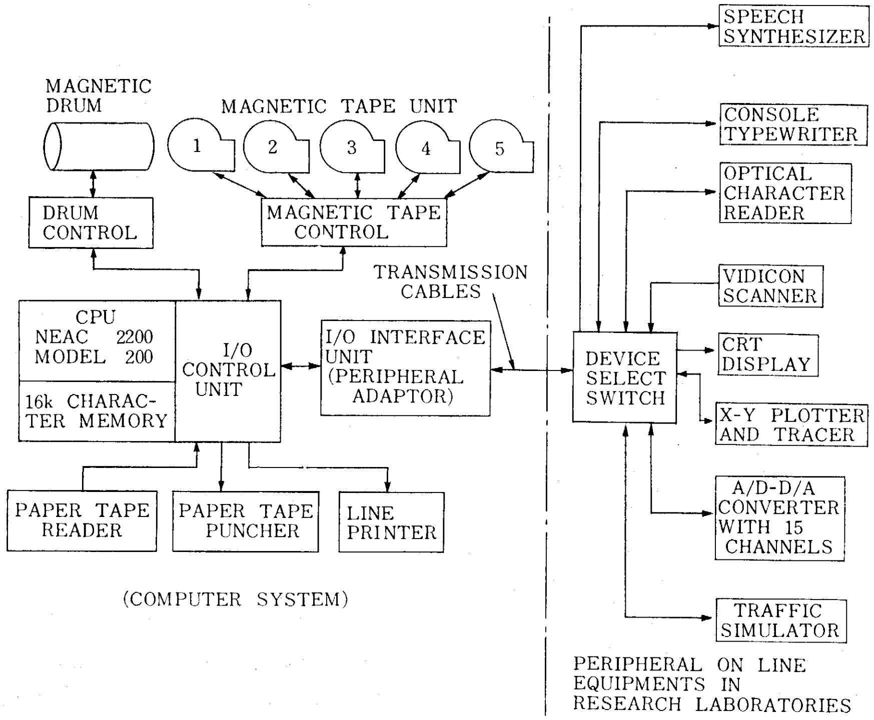


Fig. 1.1. Configulation of computer system with I/O interface unit and miscellaneous equipments,

each character having six information bits and two punctuation bits (word mark and item mark) like IBM 1400 series. The cycle time of core memory of model 200 is 2 μsec/character. The central processor unit is connected through read-write channels and I/O trunks to the peripheral devices. The 2 μsec access timing of core memory is distributed to three channels in turn, thus allowing the three concurrent I/O operations with the maximum data transfer rate of 6 μsec for each channel. The peripheral adaptor can be connected as one of the peripheral control circuit through a channel and a trunk like the other conventional peripheral control circuit.

Two I/O instructions are prepared: "Peripheral Control and Branch (PCB)" instruction and "Peripheral Data Transfer (PDT)" instruction. Both have the format of PCB A C1 C2 C3......, and PDT A′ D1 D2 D3, respectively, in which A and A′ are addresses, C1 and D1 select the channel and C2 and D2 select the trunk.

In PCB, C3 decides the function of control or tests the state of the peripheral equipment designated by C1 and C2. If the test condition is satisfied the control of the program is transferred to address A. PDT performs the transfer of data between the core location subsequent to A' and the peripheral equipment designated by D1 and D2 under the control of parameter D3. The function of C3 and D3 can be designed properly for each device. The peripheral adaptor is installed within the computer main drawer and 53 transmission cables connect it to the equipments in each laboratory room via device selection switch. The cables can be extended up to 100 meters without special modulation-demodulation system.

The system is not constructed to operate in time-sharing mode, but dedicated to one job at one time because of the limitation of the size of CPU and memory capacity. It runs under simple operating system and the user can utilize the full resources for his job. At present the system is operated in open shop for the laboratory experiments of wide range of information processing, in which closer interactions among computer, operator and other equipments have successfully been achieved.

The Peripheral Adaptor

The main rolls of the peripheral adaptor are to give the interface for the exchange of the control signals and data. Central processor unit can interpret the control signal through PCB's C3 parameter or PDT's D3 parameter. Several signals that control the transfer of data are also prepared.

The interface signals of this adaptor to a peripheral equipment are as follows: (1) four indication signals which send the states of the computer program to the equipment, and are sensed by equipments when needed, (2) six set or control signals to the equipment from the computer, (3) eight indication signals from equipment to computer which are sensed by computer when needed, (4) eight interruption signals which interrupt the normal sequence of the program.

In this formulation both equipment and computer are connected in equal relationship through the adaptor. That is, the peripheral equipments can give the interruption and indication signals, and can accept the control and indication signals. On the same way the computer gives and accepts the information by the interruption signal and by the interpretation of the signals through C3 of PCB instruction.

When CPU is in normal mode, it accepts the interruption signal from peripheral equipment and is set to interruption mode. Entering this mode, program sequence automatically jumps to interruption routine, the origin of which has been stored at the interruption register by program. In the routine, PCB tests the interruption source and jumps to A address when source of interruption is detected. When many sources are activated simultaneously, they are managed in turn. Each interruption signal is enabled or disabled by setting or resetting the corresponding allow flip-flop by PCB previously.

Data transfer is started by PDT instruction. Parameter D3 can also be used to control the equipment. During the operation of PDT the channel is kept in busy state which may be tested by PCB. Data are transmitted in six bits parallel. The simultaneous input and output are not allowed. The timing of data transfer is decided by the peripheral equipment at the arbitrary interval longer than 6 μsec, thus permitting maximum transfer rate of 166 K characters per sec. The peripheral equipment can select the input, output operations. The successive data are stored into or fetched from storage either in ascending or descending order or on the fixed location. The peripheral equipment can also control the contents of the read-write current location counter and start location counter, which increases the flexibility of data arrangements in core storage. On account of these operations, a PDT instrnction can continuously transfer the data of the length, that exceeds the size of the core memory, between the peripheral equipment and the core storage of the computer.

The termination of the PDT is instructed either by "End of Order" signal from peripheral equipment or by the detection of record mark in the core storage. It is also possible to neglect the function of record mark by special signal from equipments. Such special functions permit the full use of the ability of computer and increase the flexibility of operation of equipment.

Equipments are placed in different rooms. The signals are transmitted through 53 cables in parallel. The maximum distance is 100 meters with propagation delay of 0.4 μsec. For the farther extension of the cables or in the serial transmission of data some limitations must be imposed on logical structure and data transfer rate of the system. At the present system the equipments are selected by device selection switch in manual operation, although the automatic selection by program is possible.

An experimenter of research laboratory can design his own on-line equipments and connect it through the adaptor to the computer. Such information processing devices as shown in Fig. 1.1 have been designed. They are the devices for pattern input/output, graphic data input /output and multichannel analog data input/ output, character reader, speech analyzer/synthesizer, traffic simulator console typewriter etc...

PART II

INTRODUCTION

In this part one of the speech synthesis methods is described as the application of the previous chapter.

It is very interesting problem to consider the use of the synthesized speech sound for the means of the computer cutput. For this purpose it is necessary to be able to synthesize the speech sound in real time and it is demanded that one computer can control multi synthesizers at the same time and also each synthesizer must be simple. One of the most suitable methods with these necessary conditions is the compilation method. The method described in this part is based on this compilation method.

The unit of compilation is not a short sentence or a word, but the shorter time unit named "wave element" which corresponds to a part of speech sound such as a pitch period of a vowel. These wave elements are memorized in a simplified form, that is, the zero-crossing wave whose zero-crossing intervals are digitalized by 20 kHz sampler. By the intervalgram analysis of speech sound of a male announcer and the computer simulation of a so-called terminal analog model, about 550 wave elements are selected. The amount of information of one wave element is about 100~150 bits, and total amount of information adds up to about 80 k bits.

In Fig. 2.1, block diagram of this speech synthesis system is illustrated. The

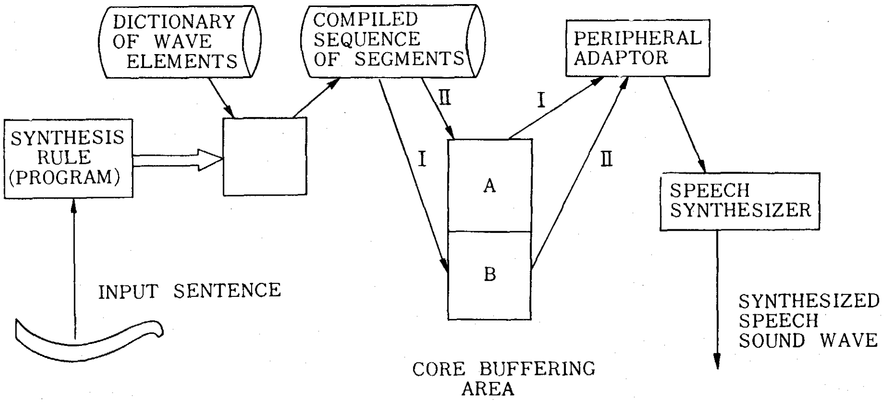


Fig. 2.1. Block diagram of speech synthesis system

computer program interprets an input roman style Japanese sentence, and the sequence of wave elements and their additional information which corresponds to the sentence are decided and read out from the dictionary of wave elements, and the corresponding zero-crossing wave is reproduced by the speech synthesizer which is combined with the computer on-line through the adaptor. The time for synthesis of each sentence is about 1.2～1.3 times real time. Intelligibility tests of the synthesized words scored about 85%.

Analysis for Synthesis

Mono-syllabic sounds were analyzed by the computer beforehand. Recorded mono-syllabic speech sounds were sampled at a rate of 20 K samples/sec and digitalized to 11 bit/sample. 20 kHz rate was dictated by the 10 kHz frequency of range. The digitalized speech utterances are stored in a magnetic tape and are analyzed by the computer. Analysis programs are written in Fortran language. Mono-syllabic sounds are subjected to infinite peak clipping and intervals of zero-crossing points are quantized into 26, 19 and 14 steps. In terms of frequency scale these approximately correspond to 50 Hz, 100 Hz and 200 Hz quantum respectively. This experiments guaranteed that the intelligibility was not lost so much even in 19 steps quantization. The sequence of quantized sampling numbers between successive zero-crossing points are expanded into Fourier series pitch-synchronously. Intensity of each component is computed every 100 Hz and the results are printed out by the line printer. In this results, it is indicated that the structure of the first and second formant is conserved considerably even in zero-crossing wave.

From now on, we call this sequence as sequence of "symbolic intervals".

Expression of Phonemes and Transitional Part

Every phoneme can be categorized into vowel, semi-vowel or consonant. Vowels are considered to be damped oscillation excited by vocal cords. In normal articulation, the vocal tract is maintained in a relatively stable configuration during most of the sound. The uttered speech sound is a damped sinusoid repeated at a excitation pulse rate. The output sound pressure of nasal or liquidlike consonants and also of buzzbar parts of voiced consonants are analogous to that of vowels at the point of its periodicity.

It is true that this periodicity of natural speech sound is not perfect, but in this synthesis system every voiced phoneme is synthesized by repeating the rectangular wave, that is, wave element which is expressed with sequence of symbolic intervals quantized into 26 steps in one pitch period. Burst parts of stop consonants are also memorized in the dictionary as a sequence of 20～100 symbolic intervals which were selected from the typical burst parts of [pa], [ta] and [ka]. For the fricative parts of phonemes /s/ and /z/ and so-on, 100 symbolic intervals were prepared. By shuffling the given symbolic intervals several times, fricative-like rectangular speech sounds are synthesized whose zero-crossing intervals coincide with time intervals decided by each given symbolic intervals. Besides, the amplitude of synthesized rectangular wave is variable at every wave element. In the other words, wave element is the minimum time unit of synthesized speech in which the amplitude is constant. In these ways every phoneme can be expressed with the several sets of the wave elements, their repetition times or shuffling times and their amplitudes. We define these set as "segment".

Phoneme /x/ is represented as follows;

$/x/=a_1P_1^{r1},\ a_2P_2^{r2},\ a_3P_3^{r3},$

Where a_i is the amplitude of wave element P_i and r_i means the repetition times of P_i. In Fig. 2.2 several examples of the expression are listed. As the amplitude of P_{sp} is zero, P_{sp} indicates the interval of pause about 20 msec. P_s, P_s and so on whose repetition times (r) are distinguished by the asteriscs symbol (*) from others are fricative-like sounds. These wave elements are arrayed r_i times, by shuffling the contents of wave elements (P) every time.

/a/	$10P_a^8$	/s/	$2P_s^{8*}$
/i/	$6P_i^8$	/ʃ/	$2P_ʃ^{9*}$
/u/	$6P_u^8$	/z/	$3P_{Buzz}^3\ 3P_z^3\ 3P_s^{2*}$
/e/	$8P_e^8$	/dʒ/	$3P_{Buzz}^3\ 3P_z^1\ 3P^{3*}$
/o/	$8P_o^8$	/hi/	$3P_{hi}^{7*}$
/m/	$8P_m^6$	/ch/	$OP_{sp}^6\ 4P_{ch}^{6*}$
/n/	$8P_n^9$	/ts/	$OP_{sp}^6\ 4P_{ts}^{6*}$
/p/	$OP_{sp}^6 2P_p^1$	/h/	$3P_h^1$
/t/	$OP_{sp}^6\ 2P_t^1$	/r/	$4P_r^5$
/k/	$OP_{sp}^6\ 3P_k^1$		
/b/	$2P_{Buzz}^6\ 2P_p^1$		
/d/	$2P_{Buzz}^6\ 2P_t^1$		
/g/	$2P_{Buzz}^6\ 3P_k^1$		

Fig. 2.2 Examples of expression of the phonemes by segments

Transitional parts between phonemes are also constructed with the combination of segments whose repetition times (r) are one and amplitude (a) are decided by the linear interpolation technique according to the amplitude of two adjacent phonemes. Wave elements of transitional parts are selected from the dictionary of wave elements so that the loci of first and second formants may be connected continuously between phonemes.

For this purpose about 500 wave elements were obtained by the computer simulation of a so-called terminal analog model. Lower four poles are simulated with the series of four single-tuned circuits. Center frequencies of lower two single-tuned circuits which correspond to first and second formants are varied every 50 Hz in the frequency range of 300～950 Hz and 700～2500 Hz respectively. The output of this circuit is subjected to infinite peak clipping and one pitch period in steady state is adopted as the wave element.

The transitions of formants are decided by the combination of vowel-consonant-vowel, that is, V-C-V set. This rule of interpolation was based on the measurements with sonagram of V-C-V sets.

Synthesis Method

In Fig. 2.3 the flow chart of synthesis program is illustrated. The computer system is NEAC-2200 model 200, and synthesis rule (program) is written by the assembly language.

Input Japanese sentences punched on a paper tape are interpreted and the speech is synthesized one by one. At first roman style Japanese is translated into phonemic expression. Then the sustained parts of this expression are substituted by the segments, as is in Fig. 2.2. Transitional parts between the phonemes are also interpolated by the sequence of segments.

For instance, input Japanese sentence,......da...... in Fig. 2.4 (a) is transformed to the sequence of segments shown in Fig. 2.4 (b), where (400, 1450) means a wave

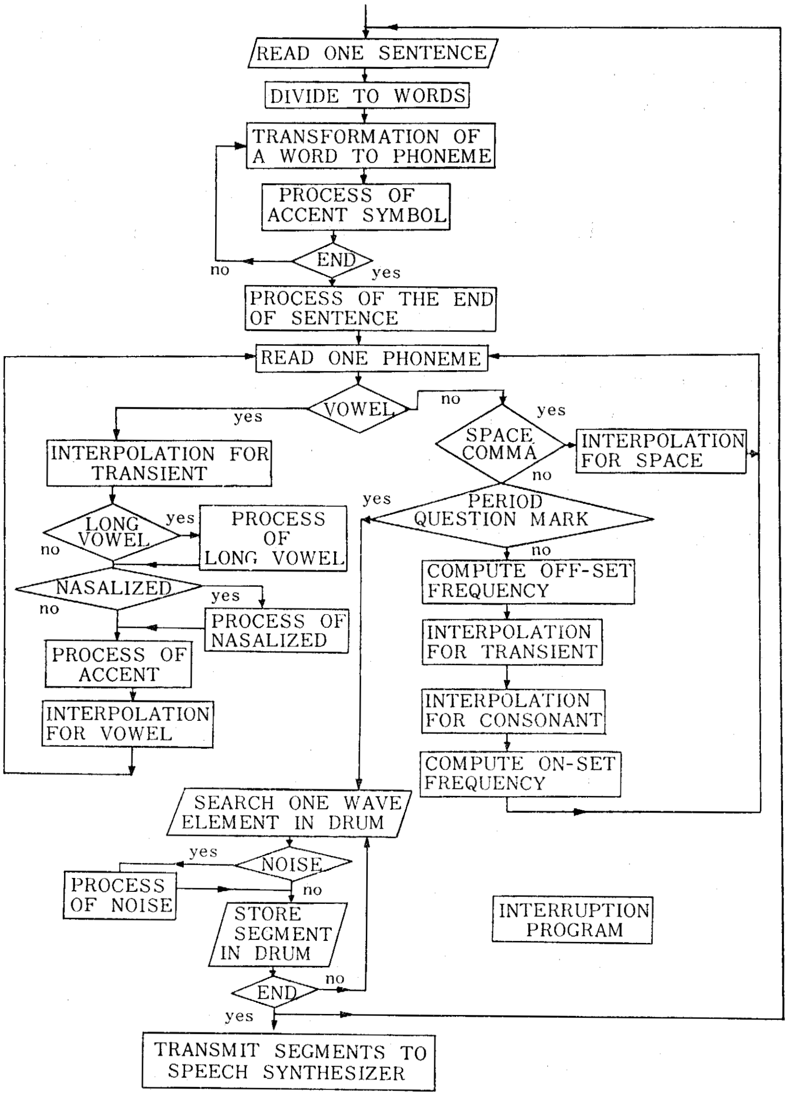


Fig. 2.3. Flow chart of speech synthesis program

(a) ······da······

(b) 2(BUZZ)^6, 2(T)^1, 4(400, 1450)^1, 6(600, 1400)^1, 8(750, 1350)^1, 10(A)^{10}

(c) 2(61, 41, ···1)^6, 2(6, 5···10)^1 4(······)^1 4(······)^1, 6(······)^1 8(······)^1 10(8,8···9,7)^{10}

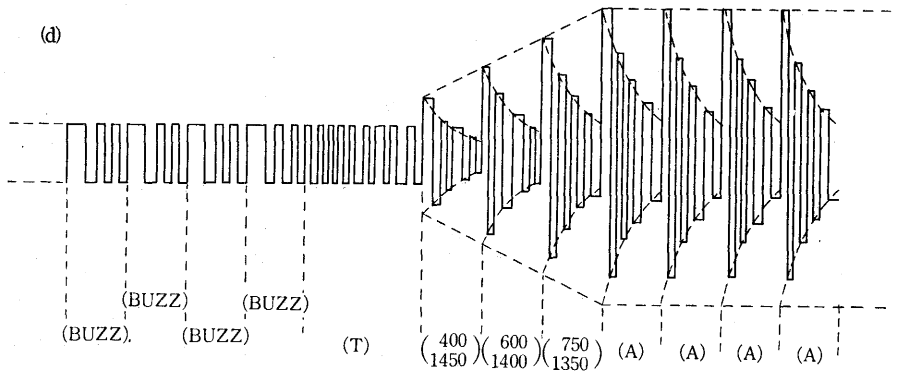
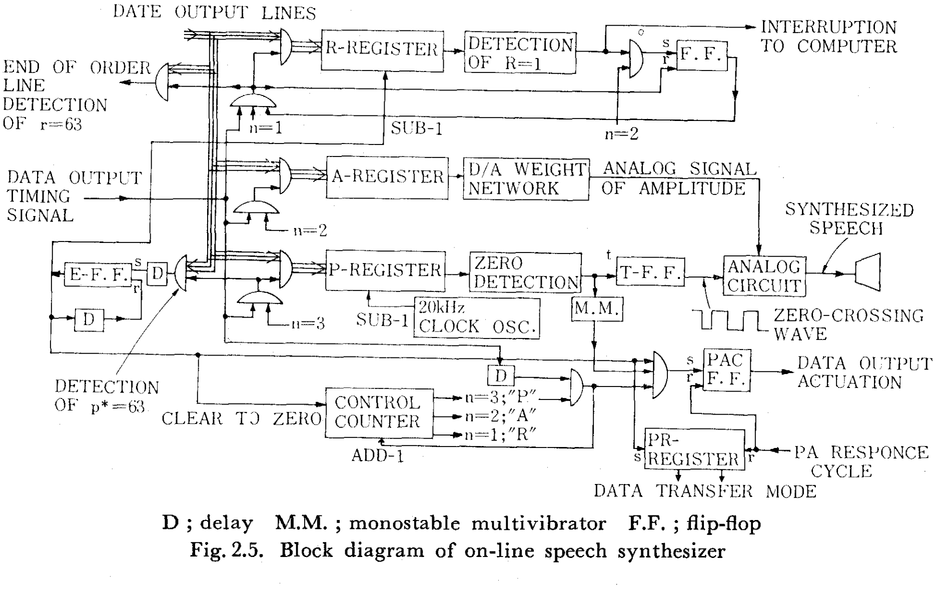


Fig. 2.4. Example of transformation of input sentence and its synthesized speech sound wave

element whose first and second formants are 400 Hz and 1450 Hz respectively. Next as is in Fig. 2.4 (c) every wave element is substituted with the stored sequence of symbolic intervals and stored in drum in order. As soon as one sentence is transformed to the sequence of segments and stored in drum memory completely, a bundle of segments is pushed out in core memory and transmitted to speech synthesis device continuously.

In Fig. 2.4 (d) synthesized speech sound wave ......da...... is illustrated. The core memory area is divided into two subareas, "A" and "B", which are the memory units of one Read/Write instruction. Subareas "A" and "B" are used as the buffer area for input of segments from drum and also for output of segments to speech synthesis device alternatively. When the data in one subarea is going to be trans-



D ; delay M.M. ; monostable multivibrator F.F. ; flip-flop

Fig. 2.5. Block diagram of on-line speech synthesizer

mitted to the speech synthesis device, next bundle of segments in drum memory are pushed down to another subarea in interruption mode.

Speech Synthesizer

Speech synthesizer is connected to the computer through the peripheral adaptor mentioned in PART I. This device operates asynchronously under control of the computer. Seventeen signal and control lines are used in this connection. In Fig. 2.5 the block diagram of the synthesizer is shown, which is constructed with about 80 integrated circuits. This device is composed of 5 sections roughly, that is;

1) R-register which consists of 5 bit set-reset sub-1 counter,

2) A-register which consists of 6 flip-flops,

3) P-register which consists of 6 bit set-reset sub-1 counter,

4) control counter and PR-register which control the gate signals and data transfer mode and

5) analog circuit which processes analog speech sound wave.

The control register has four states, and can control the gate signals of R. register, A-register and P-register. The states "R", "A" and "P" correspond to the timing of receiving the data of repetition times, amplitude and the sequence of symbolic intervals respectively which construct the segment. The state of the control register becomes "E" when the end of the segment is detected.

R-register to which the repetition times of the wave element is set consists of 5 bit set-reset sub-1 counter and 1 bit memory which is used as information whether the damping is added to the wave element or not. The contents of this register are subtracted by one when the last data of segment (P*) has been transmitted. As soon as the content becomes one, the interruption signal is transmitted to the computer, and the control of the computer is transferred to the interruption program. If the data whose 6 bit are all one (r*) is accepted in state "R", "End of Order" line is obliged to one which ends "PDT" instruction and the data channel possessed by this synthesizer becomes free.

A-register in which the amplitude information is stored is constructed with 6 set-reset flip-flops. The contents of A-register are converted into analog quantity by D/A weight network, and modify the amplitude of synthesized zero-crossing wave.

P-register consists of 6 bit sub-1 counter. Sequence of symbolic intervals are set in order in this register and are subtracted by one every 50 μsec. As soon as its content becomes zero, next data is fetched from the core memory in the computer and the state of T-FF is reversed. In the result the output of this flip-flop is zero-crossing wave. If the datum whose 6 bit lines are all one is accepted in state "P", which means that it is the end data of one wave element, and moreover if the contents of R-register are not zero, the same wave element is fetched again. In these way the identical speech sound wave is produced repeatedly according to the given number of repetition times in R-register.

RESULTS AND DISCUSSION

In Fig. 2.6 sonagram of synthesized speech "onsei bunseki" is shown.

In Table I Confusion matrix for consonants of the synthesized monosyllabic sounds is illustrated. The number indicates the times the items in the row were responded. The confusion contained in column "ϕ" means that consonant+vowel (C-V) was heard as vowel only. In the intelligibility test, synthesized monosyllabic sounds were arrayed in random two sequences and presented to three male listeners who had not been trained in particular to hearing of synthesized speech sound. The confusion between /m/ and /n/ was hardly observed. Fur-

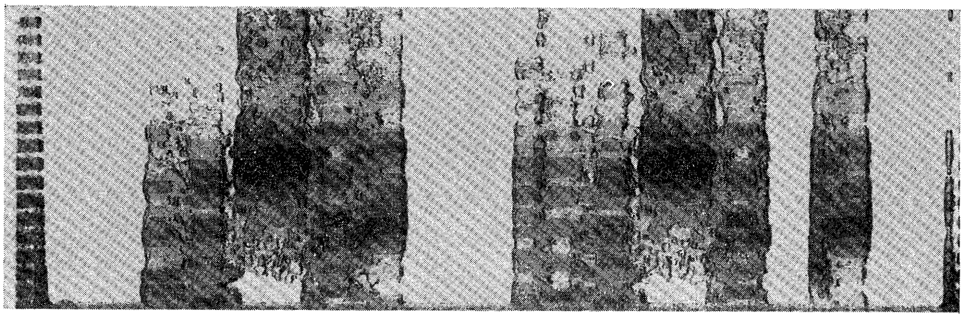


Fig. 2.6 Sonagram of synthesized speech sound "onsei bunseki"

consonant received

consonant sent	p	t	k	b	d	g	h	s	z	n	m	r	c	y	w	ϕ
p	20	7	2			1										
t	1	14	3													
k	2	3	25													
s	4	1		14	4	2				2	1	2				
b		1		3	11	3										
d					2	27						1				
g							25	1	2							2
h								30								
z									30							
n	2			2		2				12	9	3				
m										11	19					
r				2	2	9				1		16				
c													12			
y														18		
w	2			2											2	

Table I Confusion matrix for consonants

thermore, considerably many monosyllabic sounds including /p/ and /b/ were heard incorrectly. On the other hand, most of the monosyllables including /k/, /g/, /h/, /s/ and /z/ were judged correctly. Total score of intelligibility test of synthesized monosyllables is about 75%.

The features of this system are as follows,

1) It takes considerably short time to interpret the input sentences and synthesize the speech from them.

2) Compared with other compilation methods, total amount of necessary information to be stored is very small. This is caused by using a short element in the form of zero-crossing wave as the unit of compilation. We can synthesize without so much memory capacity and processing.

3) The synthesis program is very simple.

4) It is easy to control the synthesizer, the hardware of which is very simple.

5) These advantages enable this system to be connected with other information processing systems.

On the other hand, the defects of this system are;

(1)' There exist some inferiorities in naturality of synthesized speech sound, and in fluency of the sentence.

(2)' It is difficult to control the pitch frequency and amplitude continuously. The reason of (1)' is mainly caused by using only the information of zero-crossing points as wave elements and by going without analyzing the context of input sentence.

It is very important problem to shorten the required time for synthesizing the speech sound. We must pay attention to the construction of the dictionary of wave elements. And synthesis program can be improved more efficiently, in respect of decision of the interpolated transitional parts.

Intelligibility or quality of synthesized speech sound may be improved better by choosing the wave elements more elaborately and increasing the number of wave elements. For instance, it is very effective means to prepare several wave elements for one consonant according to the following vowel.

At any rate the synthesized speech sound has considerable high intelligibility. Application of this system are considered for the computer output of other information processing systems, such as, automatic translation system from English to Japanese or information retrieval system and so on.

(Sept. 30, 1968, received)