Development of Audio Response Unit

YUKIO KIMURA*, AKIRA ICHIKAWA**, KAZUO NAKATA**,
TAKESHI HYODO*** AND TETSU ASO***

Abstract

Recent development of information processing networks requires the further expansion of the function of audio response unit (ARU) to increase response vocabulary and number of multiplex channel.

The audio response unit reported in the paper meets these requirements by the principle of speech synthesis of combined sum of acoustical elements, which are damped sinusoidal waves for voiced sounds and bandpassed noises for unvoiced sounds.

These elements are pre-recorded on core memory because the memory capacity needs less than 40 k bits only.

The control data for speech synthesis are also less than 3,200 bit/sec and a magnetic drum of 512 tracks and 2 MHz clock can store more than 2,000 words of 0.8 sec, duration with 20 m sec or less maximum access time.

Digital store of acoustical elements on core memory and relatively simple processing of speech synthesis enables multiplexing operation by time devision, and the possible number of multiplex channel is 40 under the conditions described in the paper.

In many case, response messages have several pre-determined formats, accordingly, only the assignment codes for message format and variable words in the message are transmitted from CPU to ARU and one telephone line of capacity no more than 120 Baud can transmit all necessary codes for multiplex operation of 40 channels.

1. *Preface*

Realization of a practical "Audio Response Unit", used to transfer the results of processing of information by a computer to users in remote places in the form of speech, has been strongly required by the public[1, 2]. Conventional techniques of edition of pre-recorded words have already been practically realized, however, to provide a valuable information service, an audio response system having a larger number of vocabularies is required.

This paper first appeared in Japanese in Joho-Shori (Journal of the Information Processing Society of Japan), Vol. 12, No. 7 (1971), pp. 397~405.

* Railway Technical Research Institute, J. N. R.
** Central Research Laboratory, Hitachi, Ltd.
*** Totsuka Works, Hitachi, Ltd.

We developed a multiplex audio response system by adopting the principle of speech synthesis which may fulfill the above requirement for the "Test system of seat reservation usiny telephone networks"[3] for the Japanese National Railways (JNR), and we trial-manufactured this unit. In this paper, we report an outline of this unit.

2. *Principle of Speech Synthesis*

When a person speaks, pulsation of air generated by vibration of the vocal chords is affected by the resonance of the mouth (more specifically, the vocal tract) and, radiated into space from the lips. This is the idealized mechanism of speech wave generation. To simplify the understanding, if vibration of the vocal chords is assumed as a cyclic vibration, resonant waveforms made by an excitation would be repeated. The act of resonance by the mouth changes with the change of shape of the mouth (in other words, by moving the jaws and tongue), and it can be dissolved into a sum of single resonance which corresponds to each resonant mode. From the lowest resonant frequency, these are termed 1st formant. 2nd formant, 3rd formant, and so on. Since each formant corresponds to a single resonant mode, each one of these resonance is expressed in a damped sine wave having a resonant frequency and a constant of decay determined by the loss of the resonant. Thus, this damped sine wave is termed "an acoustic element of speech". In other words, speech can be synthesized by combining and editing three to four acoustic elements[4]. In a practical case, the number of acoustic elements to be prepared constitutes a problem.

As the result of examination through various speech synthesizing experiments, it was verified that a total of 51 elements would be sufficient, as listed in Table 1.

Table 1. Number of the Acoustic Elements of Speech

	Frequency Range (Hz)	Method of Allocation	Bandwidth (Hz)	Number of Elements
1st formant	240～840	constant difference $\Delta f = 40$ Hz	50	16
2nd formant	800～2,340	constant ratio $r = 1.05$	70	23
3rd formant	2,200～3,411	constant ratio $r = 1.05$	100	10
Nasal consonants	/*m*/ and /*n*/ /*N*/			2

Length of an acoustic element is determined depending upon how far the pitch cycle of speech which can be synthesized should by extended. It was determined to select an element length in 10 msec.

When eight bits are used for an amplitude and 8 kHz sampling, a total capacity required for the memory of all acoustic elements is $8\times8\times10^3\times10\times10^{-3}\times51=32.64$ k bits, which can be sufficiently small to memorize in a core memory.

In a synthesis effected by combining and editing acoustic elements, the cycle to read out acoustic elements subsequently from the memory may be externally controlled by pitch information.

Process for speech synthesis consists of:

(1) Selection and readout of a sample value of each element which corresponds to the 1st to 4th formants from the memory by control information.

(2) Simple addition of each sample value.

(3) Amplitude control by multiplication of the result of addition and amplitude control information.

(4) Conversion to speech waveforms by D/A conversion.

(5) Setting the readout address of the sample controlled by the pitch period (resetting to the head address after the completion of one pitch period) and rewriting of control information.

3. *Control of Synthesis*

There are also voiceless sounds which require synthesizing consonants; however, these can be processed in exactly the same manner if random noise having particular frequency characteristics is considered to be an acoustic element.

Table 2. Control Data of Speech Synthesis

Parameter	Number of Elements	Levels of Control	Information Bits of Data
1st formant	18*		5
2nd formant	23		5
3rd formant	10		4
Pitch frequency		16	4
Voice intensity		32	5
Hiss frequency		6	3
Hiss intensity		32	5
		Total	31

* Including two nasal elements.

The total amount of information required for speech synthesis is 32 bits or less as shown in the Table 2.

As the result of synthesis experiment, it was revealed that the time interval to renew control information may be within 10 msec; consequently, the amount of information required in synthesizing speech for one second is 3200 bits. In comparison with the system in which eight bits of acoustic waveform are memorized digitally by 8 kHz sampling, our method reduces information to 1/20.

Control information required in synthesizing speech by this system is 3200 bits/sec., and an access of within 10 msec is required.

Structure in the number of syllables of a Japanese word features the fact that 90% or more of them are five syllables or less, and mean time length of one syllable is 160 msec. Thus, if the unit of a word to be synthesized is assumed to be 0.8 sec., necessary control information will be 2.56 k bit/word.

Generally, the total amount of control information is so large that only a magnetic drum is used for the memory medium. Present drums have been practically used with a clock of 2 MHz or more; then, 2000 words can be recorded with drum of 512 tracks. However, some words carry 0.8 seconds or longer, requiring a space for two words. As a result, it is possible to synthesize an actually effective 1500 or more words.

We adopted a system to extract control information by analyzing the human voice.

Regarding the data analyzing method, an maximum likelihood method[5] for estimation of speech spectrum is mainly used and voice amplitude, 1st to 3rd formant frequencies, and pitch frequency are extracted.

4. *Multiplex Control for Response*

To allow the audio response unit to perform multiplex response in accordance with an order from the central main computer, required is a control which processes control and operation for the synthesis in a time sharing and in a multiplexing mode.

The control of speech synthesis requires a complete real time operation, and for the purpose of this time, the process of conversion and editing as described below was required. Moreover, even if the unit time length of the synthesizing speech is constant, sufficiently natural speech can be synthesized by properly selecting pause duration at the head and tail of a word. Thus, to simplify and ensure control, we adopted a system in which control by each channel is performed synchronously.

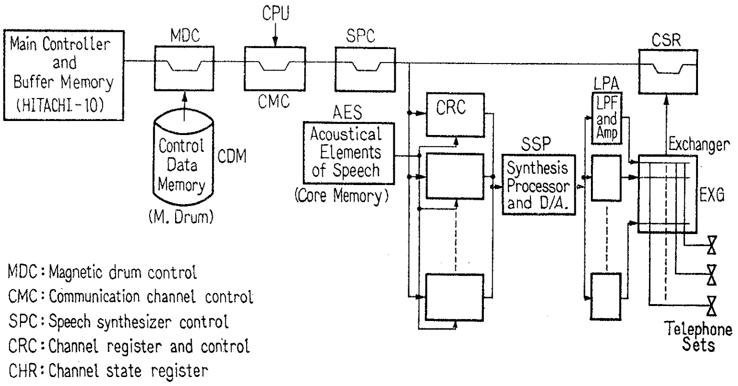
Regarding how to use the audio response unit, in most cases, the number of frames (types) of response is fixed at some number, and variable portions among them—for example, year/month/day, amount of price, name of person, name of place, name of station, etc.—are selected from vocabularies.

Thus, information received by the audio response unit from the main computer is limited to designation of type and designation of variables only and the process of converting and editing the information to understandable speech is performed on the audio response unit side.

5. *Outline of the Audio Response Unit Made for Trial Experimentation*

In accordance with the principle described in the foregoing paragraphs, we designed and produced an audio response unit on a trial basis.

Fig. 1 is a block diagram indicating unit construction.



MDC: Magnetic drum control
CMC: Communication channel control
SPC: Speech synthesizer control
CRC: Channel register and control
CHR: Channel state register

Fig. 1. Block Diagram of the Audio Response Unit

Including transmission control, all controls were effected by software of the the HITAC-10 except for the operation of synthsis. It was found possible to increase the multiplexing channel to approximately 10 channels with 8 k words or up to approximately 38 channels with 16 k words.

The portion where time increases as multiplexity increases is the control signal transmission routine and this requires approximately 100 μsec per channel. As far as processing time is concerned, it is possible to multiplex up to 80 channels. Table 3 lists the main performance of the audio response unit made for trial experimentation this time.

Table 3. Main Description of the Audio Response Unit

Item	Description
Multiplexing	12 channels
Vocabulary	1,500 words (average word length: 0.8 sec)
Response Format	Fixd formats (several types) Selection of variable parts from vocabulary
Control	Periodical multiplex control
Communication	1,200 Baud 1 channel, half duplex center-drive
Error Detection	Horizontal, vartical parity and data definition
Error Display	Alarms on data type writer
Additional Functions	Monitoring of channel states and timing data collection for statistical operation analysis

6. *Experimental Setup of Seat Reservation System Using Telephone Networks*

The audio response unit was used, as a trial, in a experimental setup of the seat reservation system using telephone networks of the Japanese National Railways[6,7]. Fig. 2 shows the construction of this experimental setup.

When a customer desires to reserve a seat, he calls the nearest seat reservation center by telephone. When the telephone channel is connected, the JNR

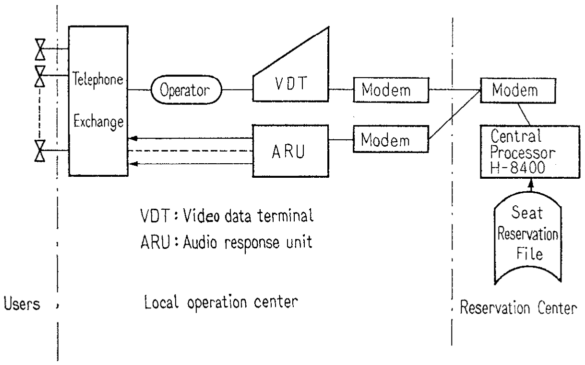


Fig. 2. Experimental Setup of Telephone Seat Reservation System

(Japanese National Railways) personnel answering the call asks whether the caller desires to make a seat reservation or an inquiry only, and then the JNR personnel inputs the information on the type of call by the typewriter of the video data terminal (VDT). The main computer, upon receipt of the information, displays a format required for a reply to be subsequently given to the customer on the CRT of the VDT. The JNR personnel queries the customer on the necessary conditions in accordance with this format, and inputs the information by operating the typewriter when the customer replies. When all conditions are completed, the main computer effects necessary processing, and indicates the resut on the VDT. If a seat reservation is not achieved at this stage, the JNR personnel interrupts the line between the customer and computer once again for further negotiation and a substitute plan. When a seat reservation is accomplished, the JNR personnel switches the telephone channel to the audio response unit, and the JNR personnel takes care of another customer's subsequent calls. The audio response unit confirms the description of the seat reservation and answers the required and new information such as transportation fee, seat number, and reservation number by an order from the main computer. The customer hears this audio reply, takes necessary notes, and thus, the seat reservation is completed.

For the effect of man-power reduction in the telephone seat reservation service by the use of a audio response unit or for increasing the number of calls to be processed which would be the equivalence of the man-power reduction (improving service ability), we are now analyzing various statistical data. According to the analysis, it has been found that time required in answering a customer is reduced one-third to one-fourth (mean value) per answering personnel. (Audio response unit connecting time/total time of service using the audio response unit=20 to 40%)

In a semiautomatic system like this which requires personnel, when it is intended to improve service, more personnel are consequently required, and it is difficult to economically establish a service system. In the future, a fully automatic system by means of a push-button dialing system may be developed. In this case, an audio response units will be required not only to confirm the result of seat reservation, but also to verify input data through push-button dialing by direct voice answer back. Consequently, the audio response unit becomes increasingly important, and more complicated response description will be required.

7. *Conclusion*

This system possesses flexibility in that this system can be used for other conventional telephone reservation services and information services. As the push-button dialing system and data communication systems are further developed, it is expected that the necessity and importance of the audio response unit will be increased.

The future problem concerning this system lies in improving the quality of synthetic speech, and we are continuously studying the preparation of control information and the improvement of synthetic speech qualities.

In closing this report, we express our deep appreciation to Mr. Hosono and his staff of the Telecommunication Division, Electrical Engineering Department, Head Office of the Japanese National Railways, who kindly offered us assistance and guidance on this JNR project, to Mr. Ohno and his staff of the Technical Research Institute of the JNR (especially to Mr. Maki).

References

[1] H. Ishida: "The State-of-the-Arts of I/O Equipment" *J. of I.P.S.J.*, Vol. 12, No. 1, pp. 37–46 (1971).

[2] J.L. Flanagan, *et al.*: "Synthetic Voices for Computers" *IEEE Spectrum*, Vol. 7, No. 10, pp. 22–45 (1970).

[3] Railway Telecommunications Association: "A Report on the Integrated Travel Information Processing System" (1970).

[4] K. Nakata *et al.*: "A Method of Speech Synthesis for Multiplexed Audio Response" *Trans. C of ECEJ*, Vol. 52–C, No. 10, pp. 579–586 (1969).

[5] F. Itakura and S. Saito: "A Statistical Method for Estimation of Speech Spectral Density and Format Frequencies" *Trans. A of ECEJ*, Vol. 53–A, No. 1, pp. 35–42 (1970).

[6] H. Hosono, H. Inoue and Y. Kimura: "Audio Response System for Travellers" 7th Domestic Symposium on Railway Cybernetics, I–124, pp. 35–39 (1970).

[7] M. Zen-nyoji and Y. Kimura: "Seat Reservation System using Telephone Networks" *J. of Japan Railway Engineers Association*, pp. 35–39 (1971).