COMPUTER SIMULATION, DEVELOPMENT AND EVALUATION OF A HIGH SPEED SPELLED SPEECH CODE

by

CHING YEE SUEN

B.Sc.(Eng.), University of Hong Kong, 1966
M.Sc.(Eng.), University of Hong Kong, 1968
M.A.Sc., University of British Columbia, 1970

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in the Department

of

Electrical Engineering

We accept this thesis as conforming to the

required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June, 1972

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Electrical Enginering

The University of British Columbia
Vancouver 8, Canada

Date July 28th 1972

ABSTRACT

A high speed spelled speech code has been developed for a reading machine for the blind. Keeping within the constraints of high reading speed and high intelligibility, a main contribution of the work has been to minimize the memory size and thus the cost of the digital spelled speech reading machine. In order to reduce the amount of memory required to generate letter sounds of this code, redundant phonemes were eliminated and a selected set of 18 basic phonemes was extracted by a segmentation program. Letter sounds were then synthesized by concatenation of these basic phonemes. Also, vowels and vowel-like sounds have quasi-periodic waveforms. These sounds were reproduced satisfactorily by repeating over and over again a pitch period extracted from the original waveforms. Another reduction of digital memory storage was accomplished by providing each individual phoneme with a minimum number of bits per sample.

The segmentation program developed runs on a PDP-9 digital computer. This program has the functions of acquisition, graphic display, data print-out, auditory presentation, manipulation and extraction of speech samples. Graphic display of the amplitude-time waveforms of various segments of a speech sample provided an accurate and efficient method of extracting the basic phonemes. Six vowels extracted in this way were experimented in a discrimination test. It was found that even when these vowels were only 10 ms. in duration, the subjects could learn to discriminate them.

The PDP-9 computer was also used to synthesize the letter sounds and to simulate a spelled speech machine. Experiment with three blind subjects indicated that they could read spelled sentences between 60 and 70 words per minute with high

intelligibility after only one hour of contact with this spelled speech code.

A difference coding scheme was used to reduce further the amount of digital memory required to store the basic phonemes. An attempt was also made to find out whether memory storage could be reduced by lowering the sampling rate. This was studied by reducing the bandwidth of the letter sounds in a subjective test using 16 blind students. Also investigated in this experiment were the intelligibility of the letter sounds and the effects of presentation speed and pause between words. Experimental results confirmed that most blind subjects could learn to recognize all 26 synthesized letter sounds after a short period of training, and they could read spelled sentences between 65 and 75 words per minute with an intelligibility score of about 85% correct. Bandwidth reduction reduced the pleasantness and clarity of the letter sounds. It was concluded that for the reduction of memory storage, the difference coding scheme was preferred and the original bandwidth of 6 kHz. should be retained.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ACKNOWLEDGEMENT

The author wishes to express his gratitude to Dr. Michael P. Beddoes, supervisor of this project, for his assistance and interest throughout the course of the work. The author is indebted to Dr. John H. V. Gilbert of the Division of Audiology and Speech Sciences for encouragement and helpful suggestions. The author also benefited from discussion with Dr. Michael S. Humphreys of the Psychology Department.

Special thanks are given to the author's wife, Ling, for encouragement and understanding throughout his graduate study.

Thanks are also due to George Austin for good maintenance of the PDP-9 computer and Rodney George for equipment assistance.

The author wishes to acknowledge also, financial support from the following organizations:

a) operating grants:

A-3290 from the National Research Council of Canada

MA-3971 from the Medical Research Council of Canada

grant from the Vancouver Foundation

grant from the Canadian National Institute for the Blind

b) capital equipment grants:

ME-3782 from the Medical Research Council of Canada

E-3291 from the National Research Council of Canada.

1 Introduction

There are now, either existing or in development, a variety of reading machines which enable the blind to obtain information from printed materials. These machines range from the simple direct translation type that produces either an auditory or tactile code for each letter, to the complex recognition type that talks directly to the operator. Simple machines, such as the Optophone which generates buzz tones and the Lexiphone which produces musical tones modulated in frequency and amplitude, are portable and relatively cheap to produce. But it requires upwards of a year of training to master the code sounds. The ultimate reading rate with these direct translation machines is also limited [1]. Comparable to these machines is the Optocon which presents an image of a character on a 24 by 6 matrix of stimulators, much training is also required to master the tactile codes produced by this machine [2, 3]. Talking machines of various designs have been considered by a number of investigators. With this type of machines, speech is either generated from stored data or synthesized by a set of rules [4] and least learning and effort are required on the part of the user. However, talking machines are very complex and costly [3] and are suitable only for library use in a time-shared mode. A personal-type of machine which strikes a compromise between expense and ease of use is the spelled speech reading machine which produces letter sounds of the alphabet. While being about two to three times more expensive than the Lexiphone, such a machine will appeal to many blind users for one principal reason, i.e. it can be used in a matter of a few contact hours instead of a year or so to master the codes for the simple machines.

One type of spelled speech reading machines called the Cognodictor has been developed by Mauch Laboratories [5]. Letter sounds of a fixed duration were

recorded on different tracks of a film drum. This kind of letter storage has disadvantages of bulkiness, cost, and fixed duration of letter sounds. A better method of storing letter data would be a digital memory substitute for the drum.

Fig. 1 shows the block diagram of a digital spelled speech reading machine under development at the University of British Columbia. It consists of a scanner of the Lexiphone to scan the printed text [1]. This scanner can be pulled either by hand or by a mechanical device along the line of print and signals are obtained

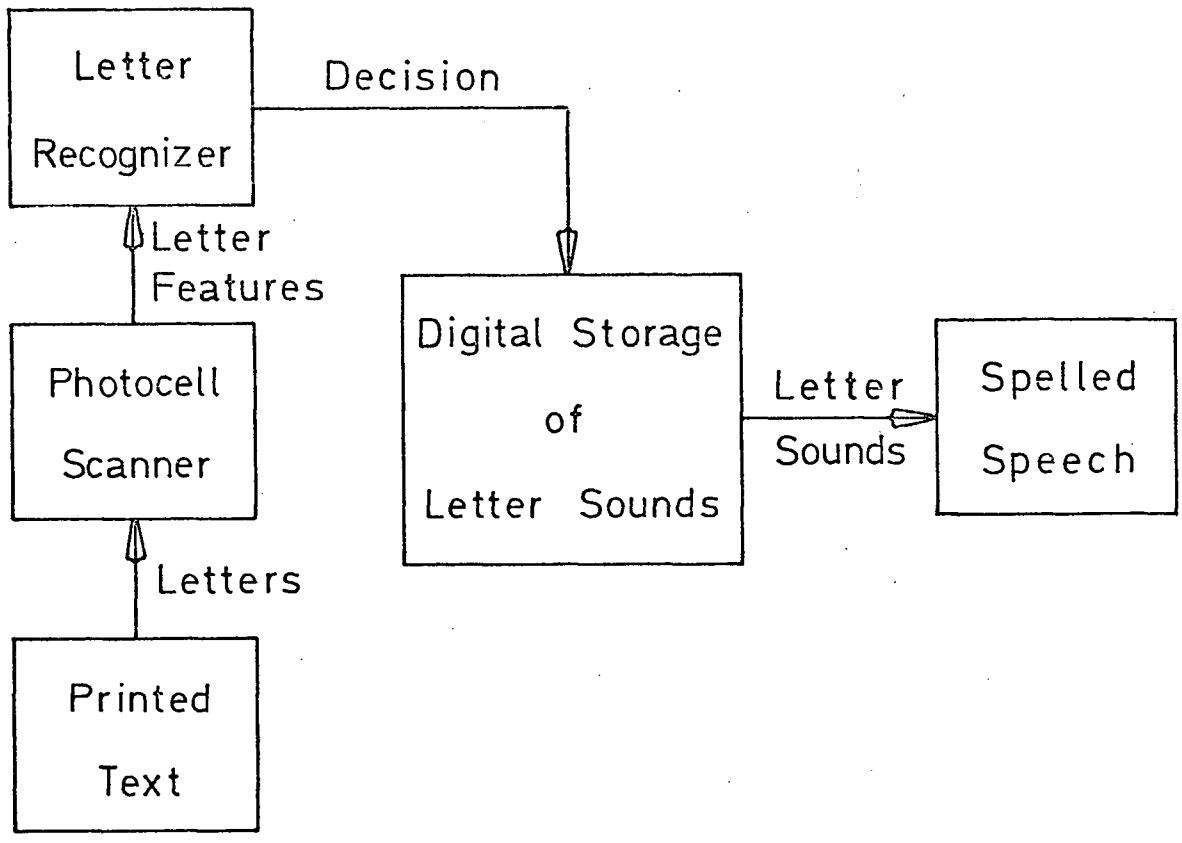

Fig. 1 Block diagram of a digital spelled speech reading machine.

from an array of 54 photocells as they pass over each of the letters in turn. The letter recognizer then makes use of these signals to extract sets of features [6] to identify the letter. After the letter has been recognized, the corresponding letter sound will be triggered and spelled speech utterances of the printed text will be produced.

Although digital storage of letter sounds is preferred to a film drum memory unit, yet at first sight, the number of samples required to produce letter sounds of good quality would appear very large and the cost would be prohibitive. For example, one can easily arrive at an estimate of the cost as follows. At 60 words per minute (wpm.), each letter (roughly 5 letters to the word) will occupy 0.2 sec. on the average; at 12.5 kHz. sampling using 9 bit samples and allowing 20% pause time between letters, it requires 468,000 bits to store the entire set of 26 letter sounds. Read-only memory will cost about one cent a bit, thus the memory will cost $4,680. In view of this, a synthesis process was sought and a number of methods were investigated to reduce the amount of digital storage and cost in constructing a digital spelled speech reading machine for the blind. In the synthesis of letter sounds, the acoustic properties of some letter sounds were also studied and were used as guidelines in the development of a high speed digital spelled speech code.

This investigation was carried out with the aid of a PDP-9 digital computer and its graphic display and interface accessories. The plan is shown in Fig. 2. First a segmentation computer program was developed to study the properties of letter sounds and to extract a set of basic phonemes from the letter sounds after sampling and digitization. An experiment was then conducted to study discrimination learning of six vowels extracted by the segmentation program. Computer software was also developed to synthesize letter sounds and to generate the digital spelled speech code by concatenation of the basic phonemes. A pilot experiment was conducted

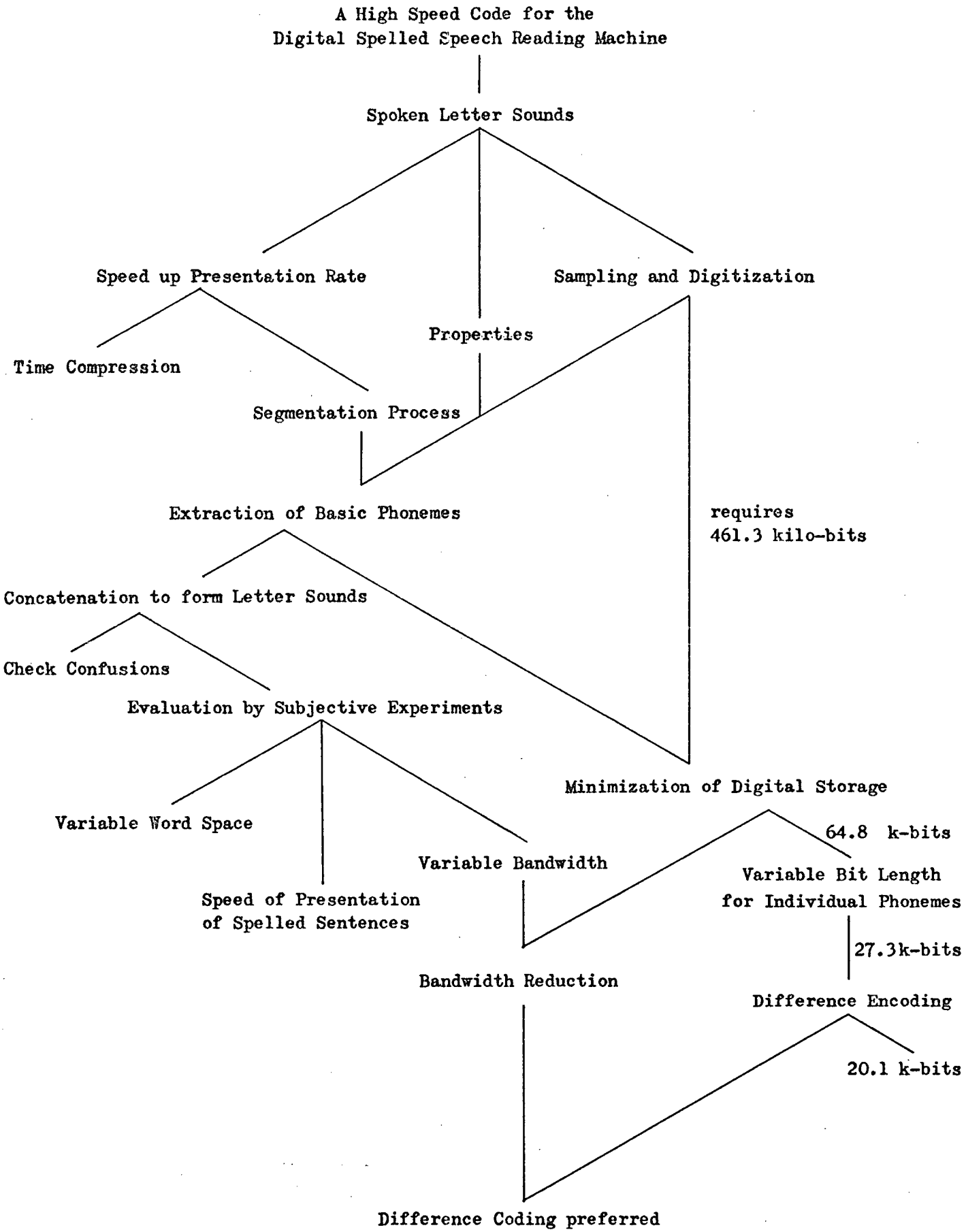

Fig. 2 Plan of study.

using three blind subjects. Memory storage of the basic phonemes was reduced by providing variable bit lengths for the individual phonemes. Further reduction was accomplished by a difference coding scheme. The possibility of bandwidth reduction was studied in an intensive subjective experiment which also investigated the effects of presentation speed and variable pause between words. This thesis is concluded with a summary, discussions and suggestions.

2 Extraction of Speech Segments

One cannot spell out words at a rate much faster than 50 wpm. In order to speed up presentation rate of spelled speech, letter sounds must be processed either by compression or by extraction of sound segments [7]. The compression process speeds up the delivery of speech by discarding alternate segments from the original speech sample. The remaining segments are then joined together to form compressed speech. Thus the average frequencies and pitch of the original speech sample are unchanged. Unfortunately this method discards segments irrespective of their importance to the original speech sample, and thus intelligibility of compressed speech deteriorates especially when the speed-up ratio is high. In the case of spelled speech, this compression process also brings in a lot of confusions among some letters, in particular, confusions among B, D, G, P and T, and confusions between M and N, F and S [8]. In the segmentation process, a speech sample is first sampled, digitized, and stored in the computer. The waveform of this speech sample can then be displayed on a screen and various segments of it can be accurately and efficiently extracted and also joined together when required. Since the segmentation process is far more precise than the compression process, it was adopted throughout this study to analyze speech sounds and to extract speech segments for the synthesis of letter sounds.

2.1 Computer System for Processing Speech

Fig. 3 shows the computer system used for processing speech signals. It consists of a PDP-9 computer as the central processing unit and a number of input and output accessories. The PDP-9 computer has a memory size of 16,000 (16 K) words and each word contains 18 bits. The teletype is used for typing in and printing out program instructions, statements and data; it is also the chief means of commanding the digital computer. Occasionally, console switches are also used to control the computer. DEC tapes are magnetic tapes for program and data storage and retrieval. The paper-tape unit is used for reading in and punching out program instructions, statements and data. A precision display unit is used to display processed speech signals or data stored in the computer, it has a display area of 9.25" square. Data points can be plotted on its screen in a waveform pattern in a square array of 1,024 by 1,024 (10 bits by 10 bits) points. Low-pass filters are used to limit the bandwidth of the sound signal entering the analog-to-digital (A/D) converter and the signal coming out from the digital-to-analog (D/A) converter. An A/D converter is required to interface the sound wave with the digital computer because a digital computer cannot handle sound wave (an analog signal) directly. Similarly, the D/A converter is used to convert the digitally processed information into useful sound wave again. Both the A/D converter and the D/A converter have 12 bits of resolution. The input signal to the low-pass filter can come from a tape-recorder or a microphone or some other devices and the output signal of the other low-pass filter can be recorded on a tape-recorder or directly put into loudspeaker, headphones, x-y plotter, etc.

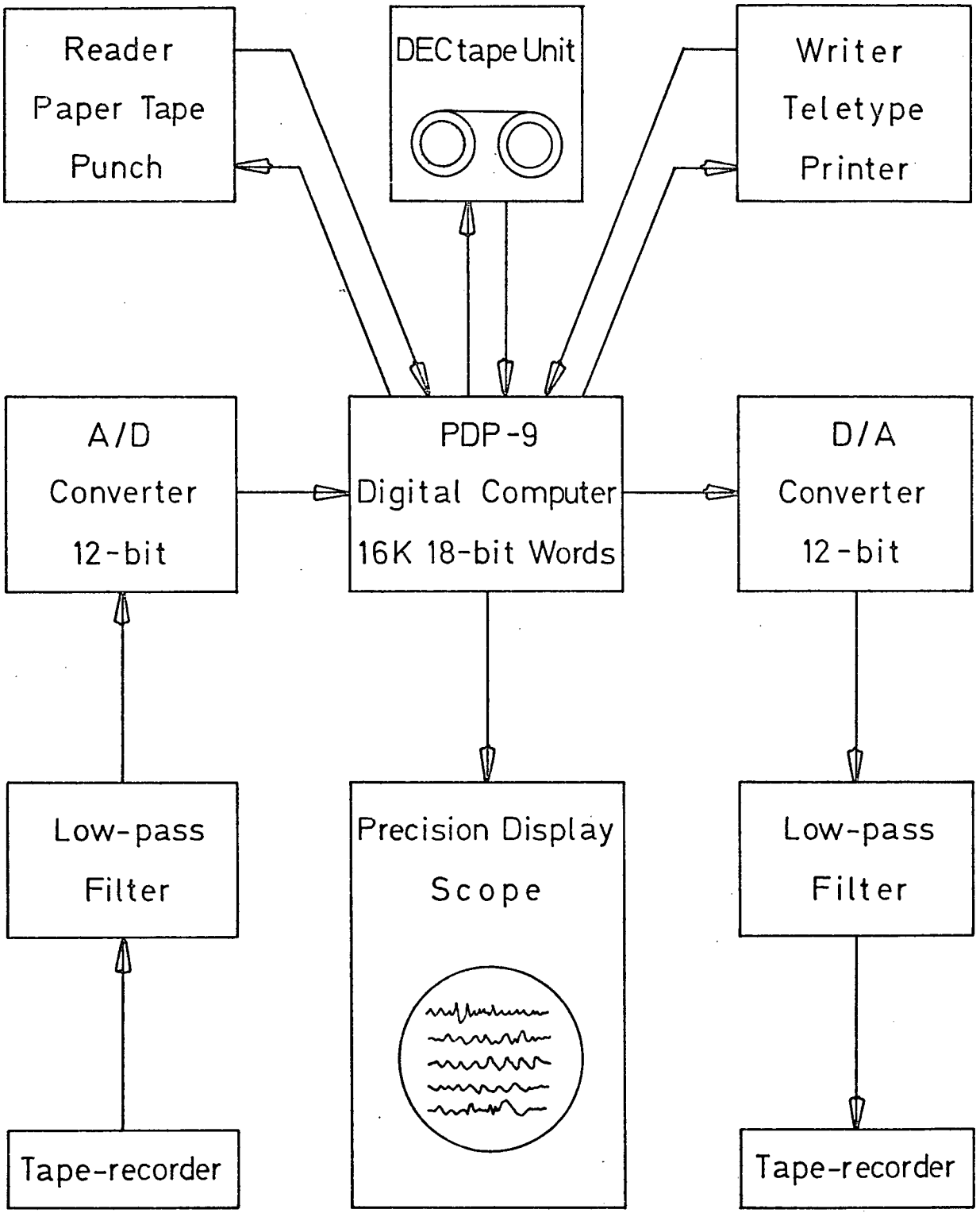

Fig. 3 Computer system for processing speech.

2.2 The Segmentation Program

Fig. 4 shows the flow chart of the segmentation program developed to process speech signals. The input signal is first sampled and then stored in the data buffer. The sampling frequency can be specified by typing in the wanted value from the teletype. The amount of data that can be stored depends on the memory size of the computer. For a sampling period of 80 μs., a memory size of 5 K (i.e. 5,000) words is required to store a speech sound that lasts 400 ms. (the average duration of a word spoken at 150 wpm.). The PDP-9 computer has a memory size of 16 K words. Setting aside 8 K words for program instructions and subroutines, it still has 8 K words of memory for storing speech data. Thus for the sampling period of 80 μs., it can store 640 ms. of speech sounds. Making use of the long word length of 18 bits of the computer, a software technique was also developed to effectively enlarge the available memory storage size by a factor of two. This was done by packing two 9 bit data samples in the 18 bit word. The actual values of these samples were obtained afterwards by an unpacking subroutine. In sampling the incoming signal, a threshold was set up to detect the beginning of the speech signal automatically. This threshold was adjusted to a level which was just higher than the noise level inherent in the input system. Once the speech sample has been stored in core memory, its data will be under complete control of the computer and can be processed in a number of ways. Some important features of this program will be described below.

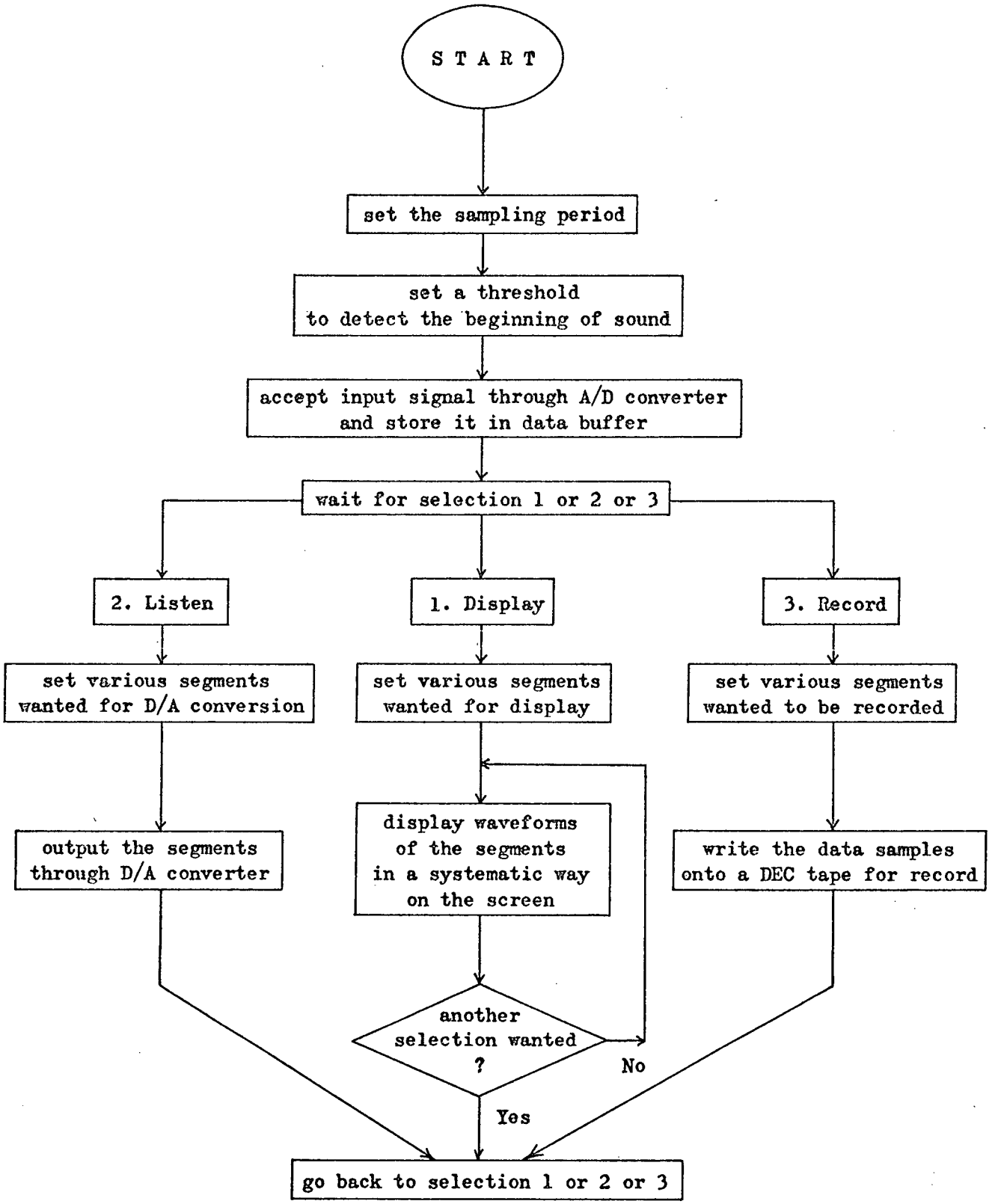

Fig. 4 Flow chart for the segmentation program.

1. Display

With the aid of a precision oscilloscope (DEC model 30), either the whole amplitude-time waveform of the speech sample or various portions of it can be displayed on the screen for visual inspection. One such waveform of the word 'split' and some segments of it displayed on this oscilloscope are depicted in Fig. 5 and Fig. 6 respectively. Not shown in the flow chart is a scheme that can amplify and expand any part of the waveform. This yields a very accurate and efficient method of segmentation and enables the operator to examine any part of the waveform in greater detail.

2. Listen

The whole word, or any segment or a combination of segments of it in any specified order, can be presented to a listener. A recording of various segments of the word 'split' to form other words was prepared and demonstrated to a number of listeners. There is also a subroutine which can repeat these sounds a number of times.

3. Record

Various segments of the speech sample can also be recorded on a DEC tape. This gives the flexibility of calling out the recorded segments when required later on.

Other features of the segmentation program include subroutines to command the teletype to print out data, or to punch data on a paper-tape so that data can be transferred to another computer or device.

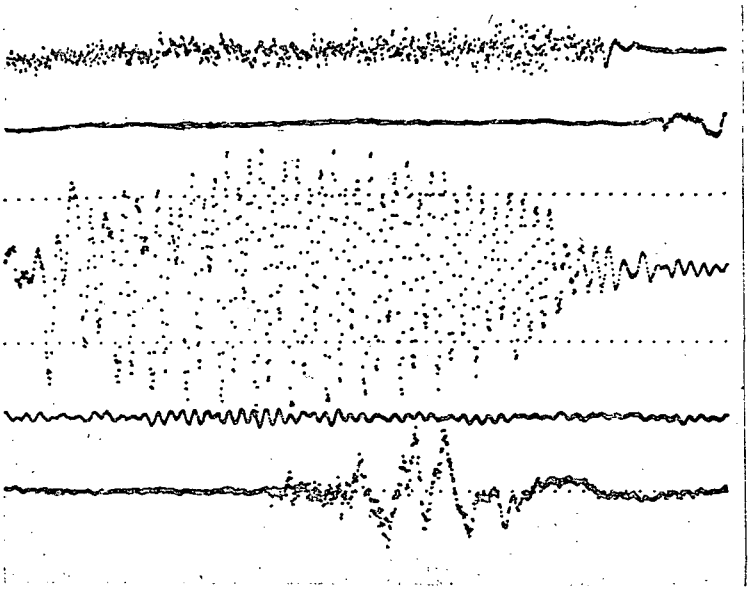

Fig. 5 Waveform of the sound of the word 'SPLIT'

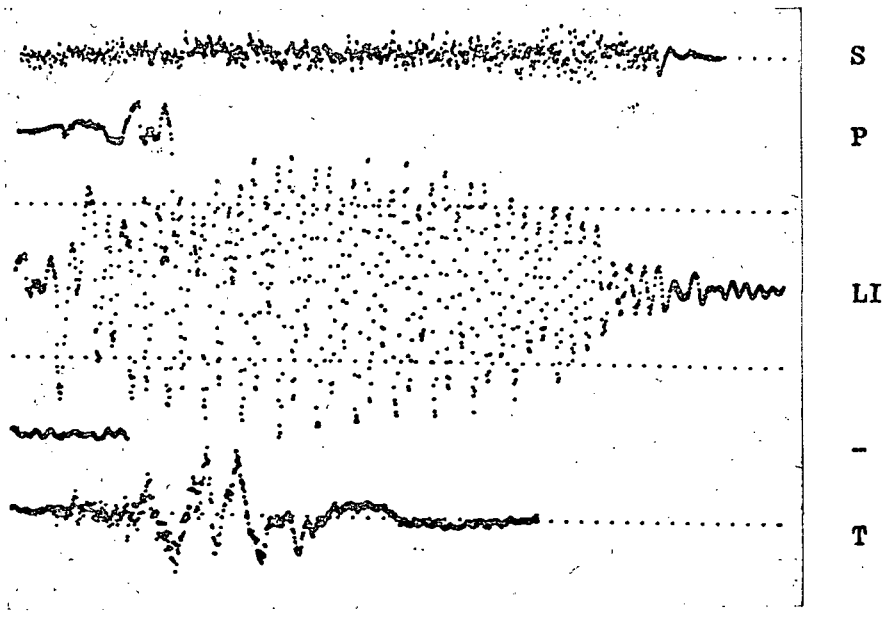

Fig. 6 Segmenting the waveform into different parts.

3 Discrimination of Vowels of Very Short Duration [9]

One way of speeding up the presentation of spelled speech is to shorten the duration of vowel sounds. This experiment investigated the effect of short duration on the discrimination of the six vowels to be used in the synthesis of letter sounds. Also, this experiment was conducted with the aim of gaining some experience in psychological acoustics experiments with speech stimuli.

In this experiment, the discrimination of six vowel sounds (/ɑ/, /ɛ/, /e/, /o/, /u/ and /i/)* of 10 ms. duration was studied. The question "Can we perceive a vowel if only 10 ms. of it is heard?" has interest both because if the answer is yes then we may actually be able with sufficient training to identify a vowel specified in this minimal way; and, alternatively, it may be permissible to produce a close replica of a natural vowel sound by repeating a period of the vowel say ten to twenty times and in this way economize the storage of data needed for the vowel.

The entire experiment was controlled by the PDP-9 computer. Vowels of equal pitch and intensity level were generated. Both phonetically trained (PT) and untrained (UPT) subjects were used. Rapid learning took place and the PT subjects showed much better discrimination than the UPT subjects. The results also indicated that subjects could be trained to correctly discriminate these 10 ms. vowels. Confusion matrices of the last four learning blocks indicated that /i/ and /u/ sounded very much alike when they were short. The pattern of the test scores was discussed with reference to pure tone perception.

* A list of phonetic symbols and key words can be found in Appendix I.

3.1 A Review of Vowel Perception

Although there has been extensive study of the duration of vowels [10-18], relatively less work has been done relating the durations of vowels to their recognition [19-22]. Recently an investigation has been made to determine the recognition threshold of some vowels as a function of temporal segmentations [23]. It was found that the median recognition threshold varied from vowel to vowel from about 10 to 30 ms. However, the results of this experiment were not in good agreement with those previously obtained by others [24-26] who reported that vowel fragments less than 10 ms. could be recognized. In other studies [21, 22], it was reported that vowels could be correctly identified at durations of the order of 30 ms. Despite the above results, discrimination learning of short durations of vowels has virtually been neglected. Using the computer to generate the stimuli, this experiment shows that both PT and UPT subjects can learn to discriminate vowels of 10 ms. A reason for using two sets of subjects was to anticipate what the effect of training might have on performance: the PT subjects represented one extreme; the UPT subjects represented the novices. We tend to identify the performance of the PT subjects with blind people who are characteristically acute listeners.

3.2 Procedure

3.2.1 Preparation of Vowels

Since the fundamental frequency limen is about 0.3 to 0.5% of the fundamental frequency [27], and intensity discrimination may be acute with short sounds [21], precise control of intensity and fundamental frequency of the vowels is necessary and this was done by a computer.

The stimuli were obtained from six sustained vowels, /ɑ/, /ɛ/, /e/, /o/, /u/ and /i/. Since irregularities of amplitude and pitch may occur when a vowel is sustained by a human speaker (e.g. see [28]), a computer controlled method was employed to extract a basic segment (the pitch-period of the voice) of the vowel waveform. This segment was then repeated a number of times to simulate the vowel waveform. The scheme of basic segment extraction was done by a segmentation program described in Chapter 2 [7]. This scheme allowed the operator to detect and extract accurately a desired segment of the vowel waveform. The sustained vowels were first low-pass filtered at 8 kHz. and sampled at a rate of 20 kHz. The waveforms were then displayed on the screen of the display unit. A pitch-period of 7.65 ms. duration (i.e. fundamental frequency = 131 Hz.) occurred in all the sustained vowels. A basic segment of each vowel with this pitch period was then extracted. The starting point of this basic segment was taken to be the zero-crossing before the major peak in the period of the vowel waveform. After these basic vowel segments had been extracted, synthesized vowels were generated and presented to both PT and UPT listeners for identification. When all these vowels were correctly identified, they were recorded on a tape. Subsequently, these synthesized vowels were played and their intensities were equalized by measurement with a rms. voltmeter (Hewlett Packard 3400A). From these synthesized vowels, a second basic segment of each vowel was extracted and

formed the basic segment of the synthesized vowels used in this study. All the synthesized vowels were low-pass filtered at 8 kHz. before presentation to the subjects. The first three formants of these resulting vowels were measured with a variable band-pass filter (Krohn-Hite model 3342R) and the rms. voltmeter. The formant frequencies in Hz. were: /ɑ/, F1 = 760, F2 = 1050, F3 = 2500; /ɛ/, F1 = 580, F2 = 1900, F3 = 2450; /e/, F1 = 510, F2 = 2050, F3 = 2700; /o/, F1 = 530, F2 = 820, F3 = 2420; /u/, F1 = 270, F2 = 660, F3 = 2350; /i/, F1 = 260, F2 =2200, F3 = 2950.

3.2.2 Experimental Design and Testing Procedure

To determine the duration of vowels to be employed in this study, a pilot experiment was conducted. This pilot experiment was also used to find out the reaction of PT and UPT subjects to this experiment. The results indicated that it was possible to discriminate among six 10 ms. vowels used in this study and that there might be a great difference between PT and UPT subjects. As a result, vowels of 10 ms. duration were used and two groups of subjects were employed. Six university students who had no training in articulatory phonetics, formed the group of UPT subjects. Four graduate students and two faculty members all from the Division of Audiology and Speech Sciences formed the group of PT subjects. The graduate students had had about one year of training in articulatory phonetics and the faculty members had had about five years of teaching experience in phonetics and speech sciences.

The design of this experiment was similar to that of House et al. [29]. The test materials consisted of six different blocks of 36 vowel stimuli each. These 36 stimuli were composed of the six tested vowels occurring six times in a block in a constrained manner so that each vowel followed itself and every other vowel in the whole block. To minimize order effect, each of the six subjects of the two groups was assigned

to a given row in a 6 X 6 Latin square. At the end of the sixth block the first two blocks were presented to the subject again.

This experiment was conducted in a quiet room. Prior to the presentation of a vowel sound, a 100 ms. "ready" signal of 1 kHz. was generated (see Fig. 7). After hearing the vowel sound, the subject was required to associate it with one of the six needle positions (indicated by numbers from 1 to 6) corresponding to different deflections on the meter in front of him. He was required to write down the number in the response period of 4.5 sec. on a response sheet provided. The feedback signal would then deflect the needle to the position with which the vowel was to be correctly associated. After this, the ready signal would again be heard before the next vowel was presented, etc. To ensure uniformity of stimuli presentation, both indicating signals and all vowel stimuli were generated by the computer. Both the 1 kHz. signal and the vowel stimuli were recorded on one channel of a stereo tape-recorder (Tandberg model 64). The signal which monitored the meter was recorded on another channel. Both the 1 kHz. signal and the vowel stimuli were presented to the subjects through a loudspeaker (Ampex F2044 speaker amplifier). Prior to the test, 6 to 10

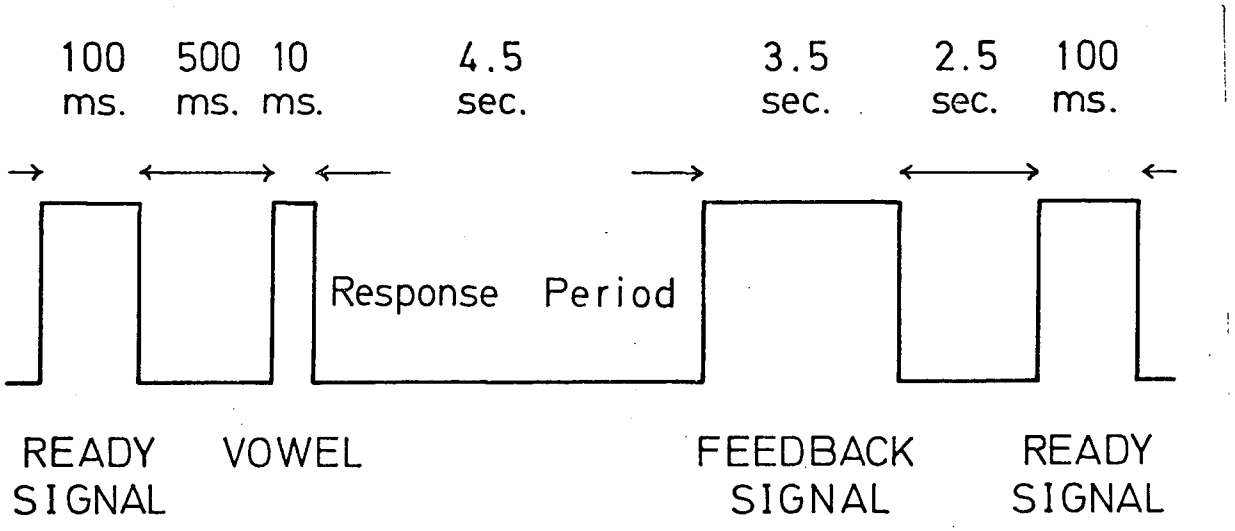

Fig. 7 Presentation of the vowel stimuli.

stimuli of a block were presented to the subject to familiarize him with the testing procedure and to adjust the sound to a comfortable listening level. He was also told the six different vowels used in this experiment. To the UPT subjects, key words (father, set, chaotic, notation, pool and beet) were used to illustrate the vowels, explanation was also provided when there was doubt. To avoid obvious relations between the stimulus and the number put on the meter, the numbers were changed from block to block following another 6 X 6 Latin square. Thus deflection of the needle of the meter was the same for the same vowel throughout the whole test, but the numbers put on the meter were changed from one block to another. There was a rest period of three to four minutes between blocks and each subject spent about 1 hour and 20 minutes for this experiment. All subjects were paid and were encouraged to try their best by giving them a bonus if they got a good average percent correct discrimination.

3.3 Results and Discussions

The means and standard deviations of the percent correct discrimination scores for both groups of subjects are shown in Table 1. Graphical displays of the test scores are also shown in Fig. 8. The large standard deviations indicate that initially there was quite a big spread among the test scores of the different subjects. Deviations among the scores of the PT group however, were not great in later blocks of discrimination learning.

It must be emphasized that even though the vowels were only 10 ms. in length, they did sound like vowels to the PT subjects after several exposures. In fact, some of the vowels, particularly /α/ and /ε/, were recognized by most of the PT subjects on first hearing them. During the test, they also mimicked the vowels and

Block	PT Subjects		UPT Subjects	
	Mean	SD	Mean	SD
1	51.4	17.48	25.0	10.27
2	69.0	17.23	38.4	13.56
3	76.9	23.61	61.1	12.73
4	85.2	13.68	66.7	16.11
5	89.8	7.46	68.5	21.60
6	91.7	6.00	69.0	21.84
7	94.9	5.88	76.9	19.08
8	95.8	5.96	74.1	15.83

Table 1 Means and standard deviations of the percent correct discrimination scores of PT and UPT subjects as a function of test blocks.

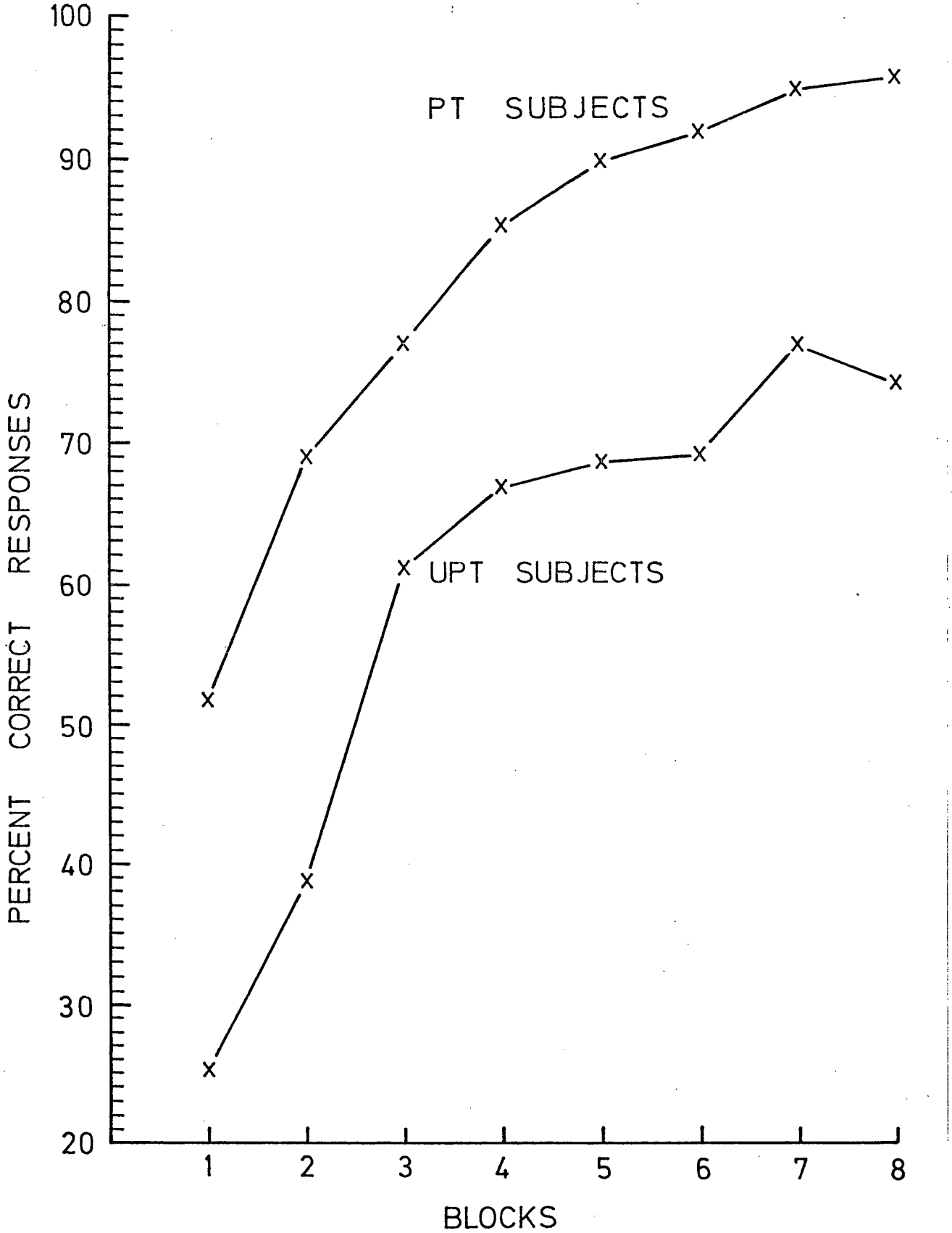

Fig. 8 Discrimination scores for PT and UPT subjects as a function of test blocks.

tried to map them into their own vowel system. To the UPT subjects, the vowels sounded like clicks.

Large differences were obtained in the performance of the two groups of subjects. The scores of the UPT group ranged from 15 to 25% below those of the PT group. Rapid learning took place in the first four learning blocks after which the scores rose steadily. The PT subjects approached perfect discrimination of the six vowels and there were four subjects (including one UPT subject) who reached the 100% correct discrimination scores towards the end of the experiment.

An analysis of variance was performed on the discrimination scores shown in Table 1. The results of this analysis are shown in Table 2. Significant differences were obtained in learning blocks ($p < 0.001$) and groups ($p < 0.05$), but not their interaction ($p > 0.10$).

An alternative way to examine the data is to see how the different vowels affected the test scores of the subjects. For this, discrimination learning curves for the different vowels were plotted and are shown in Figs. 9 and 10. The PT subjects could learn the vowels /ɑ/, /ɛ/, /e/ and /o/ to 100% correct discrimination. The UPT subjects also had high scores for these four vowels. Both groups of subjects had lower scores in /i/ and /u/. Confusion matrices for the last four blocks are shown in Tables 3 and 4 respectively for the PT and UPT subjects. These matrices reveal that /i/ and /u/ sound very much alike when they are short. This kind of /i/ and /u/ confusion has also been observed before by Powell and Tosi [23].

Clarification concerning the scores of the different vowels can be achieved by reference to the perception of short tones. Bürck, Kotowski and Lichte (described in [30]) found that the duration of a tone required to produce the experience of a definite pitch decreased from about 50 ms. to 11 ms. as the frequency of the short tone increased from 50 Hz. up to 2 kHz. Beyond 2 kHz., duration increased with

frequency. For a tone of 250 Hz., about 20 ms. was required to produce a definite pitch and only about 12 ms. was required for a tone in the range of 1 to 4 kHz. These figures indicate that the first formant is the most severely affected formant when the vowels have a duration shorter than 12 ms. Since intensity has a strong influence on the perception of short vowels and the first formant is the most intense formant, the lower the frequency range of this formant is, the more difficult it is to perceive the short vowels. Since /ɑ/ and /ɛ/ have the highest first formant frequencies among the six tested vowels, their discrimination might be expected to be best. This is in agreement with the results. Both /e/ and /o/ have a first formant frequency lower than that of /ɑ/ and /ɛ/, their scores were correspondingly lower. The first formants of /i/ and /u/ (260 and 270 Hz. respectively) lie in the lowest frequency range among the six vowels. As a result, the scores of /i/ and /u/ were poorest and these two vowels were easily confused because of the lack of a good perception of their first formants. This suggests that discrimination of vowels of a very short duration is like the perception of short tones, and for a short duration, vowels with a higher first formant frequency are better discriminated.

Source of Variation	df	SS	MS	F	p
Between Subjects	11	3192.3			
Group (G)	1	1488.4	1488.40	8.74	<0.05
Subjects within groups	10	1703.9	170.39		
Within Subjects	84	4365.7			
Block (B)	7	3069.0	438.43	24.86	<0.001
Interaction (BG)	7	62.3	8.90	0.50	>0.10
B X Subjects within groups	70	1234.4	17.63		

df = degree of freedom, SS = sum of squares, MS = mean square, F = statistic, p = probability.

Table 2 Analysis of variance of discrimination scores.

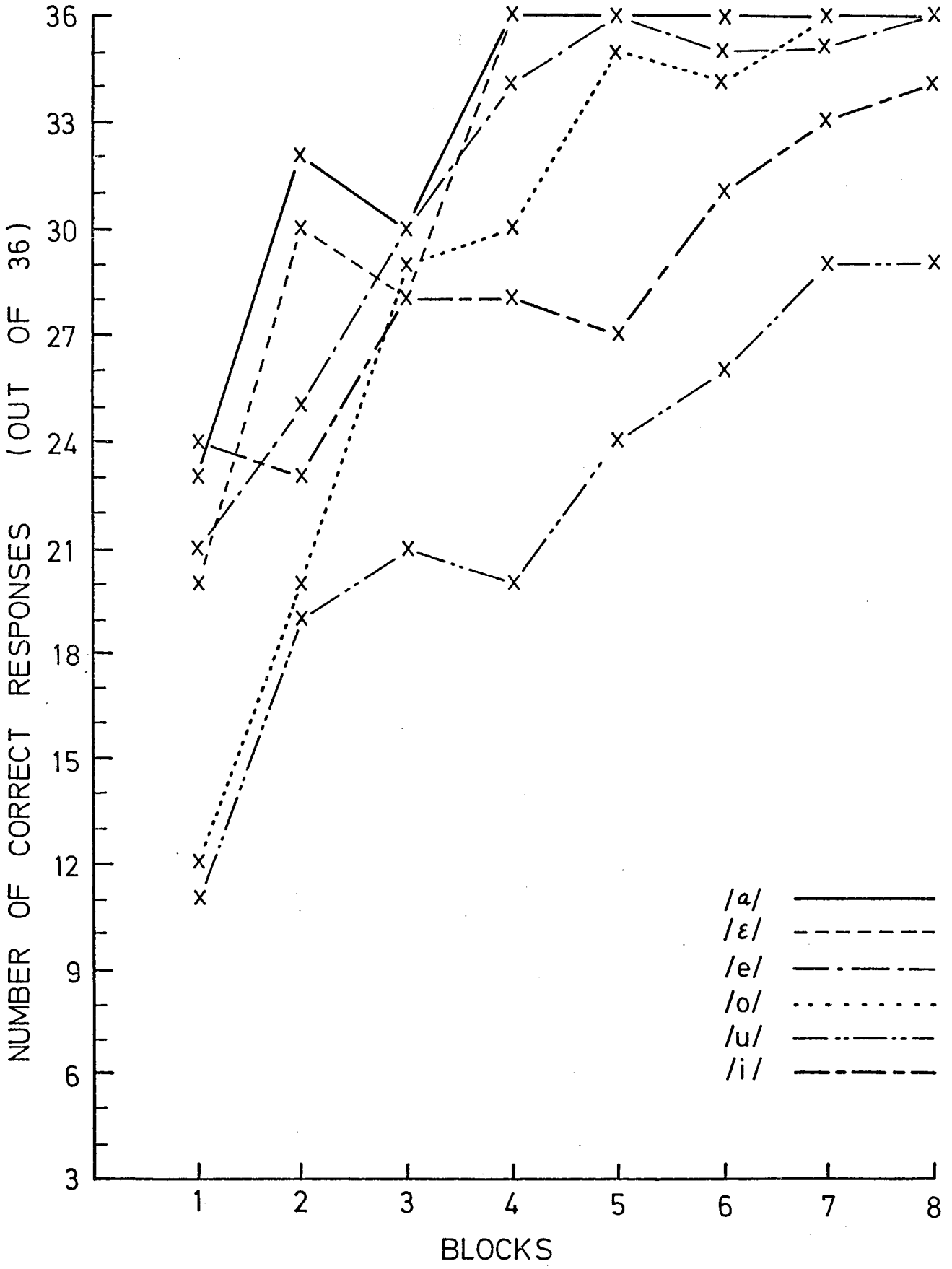

Fig. 9 **Discrimination scores of PT subjects for the six tested vowels.**

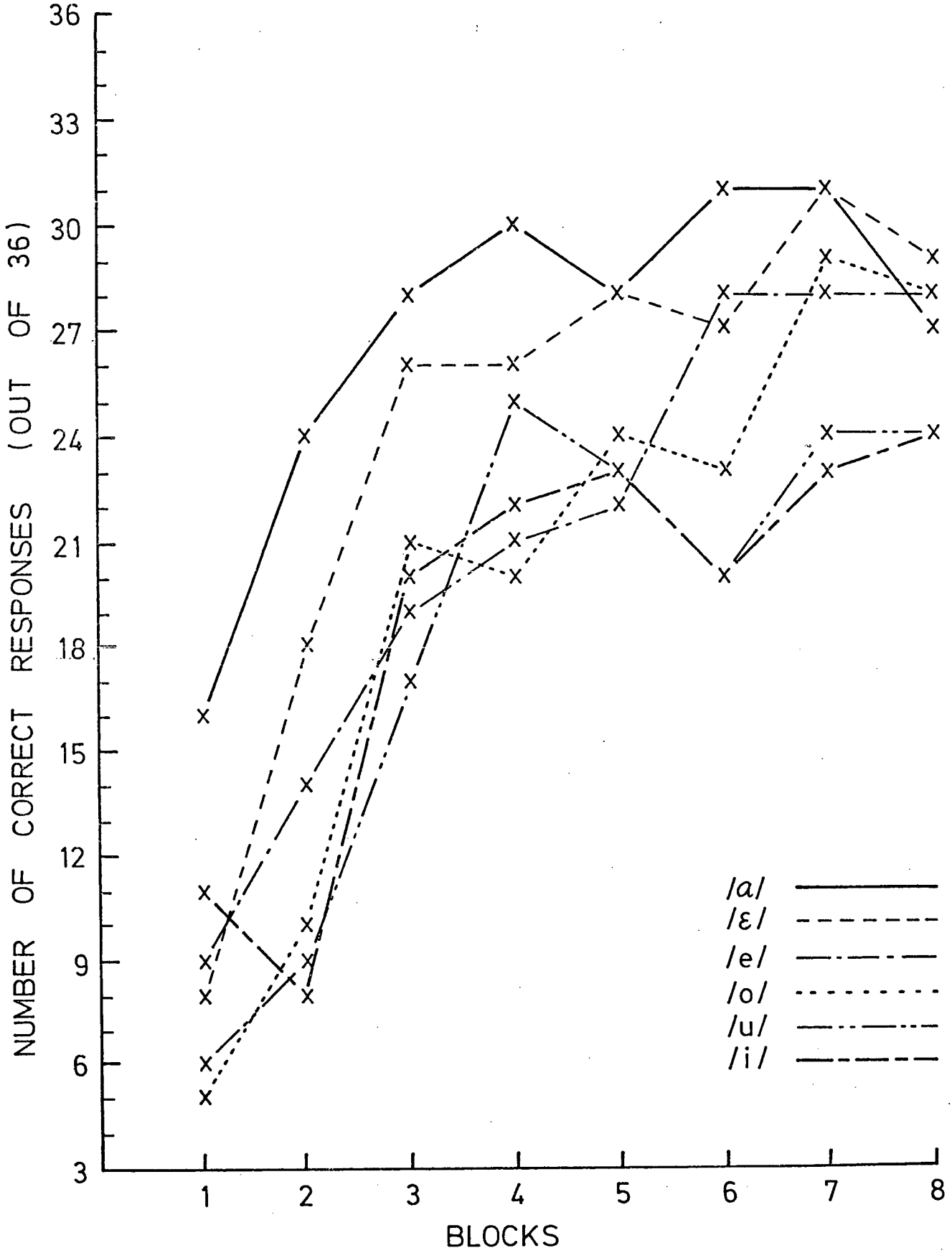

Fig. 10 Discrimination scores of UPT subjects for the six tested vowels.

Stimulus	Response					
	α	ɛ	e	o	u	i
α	144					
ɛ		144				
e			142	1	1	
o				141	3	
u				3	108	33
i				2	17	125
Total	144	144	142	147	129	158

Table 3 Confusion matrix of PT subjects for the last four blocks (144 vowel stimuli).

Stimulus	Response					
	α	ε	e	o	u	i
α	117	22	5			
ε	16	115	10	2	1	
e		9	106	15	7	7
o	1	2	13	103	13	12
u			3	7	91	43
i	3		12	12	26	91
Total	137	148	149	139	138	153

Table 4 Confusion matrix of UPT subjects for the last four blocks (144 vowel stimuli).

4 Preparation of Letter Sounds and Spelled Speech Experiment*

In order to reduce the amount of memory required to generate letter sounds, redundant phonemes were eliminated and a selected set of basic phonemes was extracted by the segmentation process. Letter sounds were then synthesized by concatenation of these basic phonemes. Also, vowels and vowel-like sounds have quasi-periodic waveforms. These sounds were reproduced satisfactorily by repeating over and over again a very small segment (corresponding to the pitch period of the voice) extracted from the original waveforms. In this way, the amount of stored data should be drastically reduced. Through a subjective experiment, it was observed that spelled speech artificially generated according to the above algorithm was assimilated at a rate which was close to the upper bound calculated using the concepts of spoken speech.

* This work was presented at the International Conference on Speech Communication and Processing, Boston, April 24-26, 1972 [31].

4.1.1 Basic Phonemes and Letter Sounds

It is noted that letter sounds of the alphabet contain many redundant phonemes, e.g. phoneme /i/* in letters B, C, D, E, G, etc. To minimize the number of samples required to represent letter sounds and thus minimize memory storage, this kind of redundancy must be eliminated. Table 5 shows a list of 18 basic phonemes selected in such a way that distinct sounds of the whole set of 26 letters of the alphabet could be generated by concatenation of these basic phonemes.

Consonants	/b/	/s/	/d/	/dʒ/	/k/	/p/	/t/	/v/	/w/
Vowels, Liquid and Nasals	/i/	/e/	/ɛ/	/ɑ/	/o/	/u/	/ℓ/	/m/	/n/

Table 5 List of basic phonemes.

The basic phonemes were obtained by extraction from naturally spoken letter sounds using the segmentation program described in Chapter 2. First, several samples (varying from 5 to 15) of each letter sound were recorded in succession in a quiet room. A Tandberg (model 64) tape-recorder and a Brüel & Kjaer condenser microphone (model 4145 with preamplifier 2619) were used. All materials were spoken by the author. Next, the best sample of each letter sound was chosen. Each best sample was then prefiltered by a low-pass filter (Krohn-Hite model 3342R), sampled by the computer, and stored in core for extraction by the segmentation program.

* A list of phonetic symbols and key words can be found in Appendix I.

In the sampling process of sound, the higher the sampling rate, the more perfect will be the reproduced sound. The number of bits/sample must also be high. But a higher sampling rate and a higher number of bits/sample will give a proportionately higher number of bits required for storage. According to the sampling theorem, the sampling rate should be $\geq 2\ W$ for a signal having frequency components bandlimited $0 - W$ Hz. In this study, a 12.5 kHz. sampling rate was chosen because it was shown that speech intelligibility was only slightly deteriorated when low-pass filtered around 6 kHz. [32]. Experimenting with different number of bits/sample, it was observed that 9 bits/sample was quite adequate for good letter sounds. As a result, all phoneme data was low-pass filtered at 6 kHz. and sampled at 12.5 kHz. (i.e. at 80 μs. intervals) with 9 bits/sample.

Synthesized letter sounds are presented in Table 6. In the assignment of letter sounds, some consideration had been given to reduce confusions among letter sounds. /ɛv/ was assigned to letter F so that voicing (/v/) at the end of the letter would make it more distinct from letter S (/ɛs/). For the same reason, the pitch of phoneme /m/ in letter M was made higher than that of phoneme /n/ in letter N. Distinctions among stop consonants deserve a greater attention because of their high frequency of occurrence in the language and their ease of being confused. As a first step toward making voiced phonemes (stop consonants /b/, /d/ and affricate /dʒ/) more distinct from voiceless phonemes (stop consonants /k/, /p/ and /t/), a slightly stronger than normal aspiration was given to all voiceless stops. The next step was to find out the proper silent intervals which should be put before the bursts of these stop consonants. It was observed that a voiced stop seemed to merge with the preceding sound and voiceless stops sounded like their voiced cognates when short silent intervals were used. In an experiment to measure the silent intervals of stop consonants [33], it was found that the silent intervals

of stop consonants were much longer than those of their voiced cognates. The average difference in silent intervals between voiced and voiceless stop consonants was about 30 ms. Based on this result, the silent interval of voiceless stop consonants was made 30 ms. longer than the voiced ones. This additional amount of silent interval for voiceless stop consonants proved in pilot experiments to be essential especially when letters were presented at a high rate in spelled speech.

Letter	A	B	C	D	E	F	G	H	I
Sound	ei	bi	si	di	i	ɛv	dʒi	eidʒ	ɑɛe
Letter	J	K	L	M	N	O	P	Q	R
Sound	dʒei	kei	ɛl	ɛm	ɛn	ou	pi	kiu	ɑ
Letter	S	T	U	V	W	X	Y	Z	
Sound	ɛs	ti	iu	vi	dɑbiu	ɛks	wɑɛe	sɛ	

Table 6 Letter sounds synthesized by concatenation of basic phonemes.

4.1.2 Savings in Memory Storage

In terms of bits of digital samples required, this method of synthesizing letter sounds from the constructed set of basic phonemes gives a tremendous saving in memory space and thus construction cost is minimized. Table 7 shows a set of figures comparing the number of bits required by naturally spoken letter sounds uttered at a rate of about 50 wpm. and the number of bits required by the 18 basic phonemes. These figures show that the selected basic phonemes occupy only one seventh of the normal memory storage required by naturally produced letter sounds.

Mode	Spoken Letter Sounds	Basic Phonemes
Duration of Sounds (sec.)	4.1	0.576
Memory Space (kilo-bits)	461.3	64.8

Table 7 Comparison of memory space required by spoken letter sounds and memory space required by the basic phonemes.

4.2.1 Preliminary Experiment with Blind Subjects

To test the acceptability of concatenated letter sounds and the listening speed blind subjects could listen to spelled speech, an experiment was conducted using a PDP-9 computer to simulate the digital spelled speech reading machine. Letter sounds were synthesized by concatenation of the proposed 18 basic phonemes. They were also low-pass filtered at 6 kHz. to eliminate unwanted harmonics resulted from the sampling process. Three blind subjects aged between 14 and 19 were used. They all had a knowledge of grade II Braille. Before the actual test, the subjects had a practice session of one hour to familiarize themselves with the letter sounds. The subjects were tested one at a time. During the first half hour or so, letter sounds only were presented to the subjects. For the first fifteen minutes, they were given control of the keyboard of the teletype and could listen to any letter sounds they wanted by striking the corresponding keys. After this, a quiz of identification of letter sounds was given to them. This short quiz indicated by that time they had had already practically no difficulty in distinguishing the 26 letter sounds. When this was finished, single words and sentences at various speeds were presented. Presentation rate was varied by computer control of the silent intervals between letters (20 - 64 ms.) and words (88 - 282 ms.) corresponding to 50 - 80 wpm. As different authors have different ways of specifying presentation rate, a standard sentence "If you want to know reading speed divide twelve hundred by the time in seconds needed to read this sentence" was used to determine the reading rate. This sentence contains a reasonable distribution of letters and comes up with an average word length of 4.4 letters. The maximum speed that a subject could listen to spelled sentences was also determined in this practice session. When this practice session was over, the subjects had a rest period of ten minutes after which the actual test was given.

The test material consisted of selected lists of phonetically balanced sentences [34]. Each list had an average length of 78 words contained in ten sentences. Each sentence was stopped at two places corresponding to the middle and end of the sentence for the subjects to respond after presentation. Three lists were tested for each listening speed and the subjects were instructed to omit the words if they could not catch them. The whole test took about one hour and twenty minutes.

4.2.2 Experimental Results

The intelligibility scores of this test are shown in Table 8. This table indicates that the blind subjects could read spelled sentences between 60 and 70 wpm. with a high average intelligibility score of 90% correct. Also the subjects had no difficulty in recognizing these artificially generated letter sounds after a short period of training.

Subject	A		B		C	
Listening Speed (wpm.)	60	70	50	60	60	70
% Correctness	96.59	93.61	80.41	82.65	89.33	89.39

Table 8 Average % correctness scores of three blind subjects.

4.3 An Upper Bound to Spelled Speech Assimilation

In order to have an estimate of the maximum listening rate blind people could attain with spelled speech, a heuristic comparison with rapidly presented speech will be made. It is believed that, for compressed speech, 275 wpm. is a speech rate beyond which comprehension begins to decline sharply [35]. When isolated phonetically balanced words were presented to listeners at this rate, Foulke and Sticht [36] found the correctness score of intelligibility was 91%. A speech rate of 275 wpm. corresponds to a syllable rate of 6.55 syllables/sec. when one word is counted as 1.43 syllables [37]. Each spelled speech letter contains about one syllable. If one word is taken to contain an average of 4.4 letters and each letter sound is represented by one syllable, then 6.55 syllables/sec. is equivalent to a spelled speech rate of 89 wpm. However, letter sounds have to be strung together to form words for perception of meaning, this will slow down the listening rate somewhat for spelled speech to be intelligible. For similar test materials (phonetically balanced sentences), two blind subjects of the present experiment achieved a rate of 70 wpm. with a comparable intelligibility score of 91.5% correct. Although listening rate can be increased with learning, an increase of more than 10 wpm. is doubtful. Thus it is estimated that for an average blind subject, the maximum acceptable listening rate for good intelligibility of spelled speech is around 80 wpm. In order to increase this rate still further, some kind of contractions is necessary and this will be discussed later.

5 Minimizing the Amount of Memory Required to Store the Basic Phonemes

Two parameters, the sampling rate and the number of bits contained in a sample, were considered. When these parameters are reduced, the data storage will be reduced accordingly. Reduction of the number of bits per sample was accomplished by providing the minimum number of bits required by each individual phoneme. Further reduction was achieved by an encoding scheme. Reduction of the sampling rate was studied by varying the bandwidth of the basic phonemes.

5.1.1 Bit Reduction by Considering the Individual Basic Phonemes

It is well-known that intensities differ a great deal from one phoneme to another. Thus only a small number of bits per sample will be required by those phonemes which have relatively low intensities.

The set of basic phonemes was examined by displaying the amplitude-time waveform of each individual phoneme on the screen of the precision display unit of the PDP-9 computer. A computer program was developed so that the magnitude and size of the waveforms could be modified. In the manipulation of the phoneme data, the maximum of the waveform was defined as the greatest amplitude in bits per sample occurring in the waveform. In case the maximum was only slightly bigger than a certain bit range, the data was modified to fit into this bit range so that one bit per sample of the data could be saved. After each modification of the data, comparison by listening was made to ensure the modified phoneme sounded as good as the original one. In this manner, the resulting phoneme data amounted to 27.3 k-bits instead of 64.8 k-bits for 9 bits per sample. Distributions of maxima of the modified data of the basic phonemes are shown in Table 9.

Phoneme	Maxima (bits)	Phoneme	Maxima (bits)
b	5	i	4
s	3	e	5
d	5	ɛ	4
dʒ	4	ɑ	5
k	5	o	5
p	5	u	5
t	5	ɭ	4
v	4	m	5
w	5	n	4

Table 9 Distributions of maxima of the basic phonemes.

5.1.2 Magnitude Difference Encoding Scheme

When the sampling rate is high, the magnitude differences between successive samples of phoneme data are much smaller than the actual magnitudes of the data. Thus storing only the magnitude differences will save a lot of storage space. The original magnitudes of the data samples can be recovered simply by the addition of successive magnitudes. The operation of this scheme resembles differential pulse code modulation. The following equations and Table 10 will illustrate how this scheme works.

$$\text{Actual data:} \quad D_a = \sum_{i=1}^{n} A_i, \qquad i = 1, 2, 3, \ldots\ldots, n$$

$$\text{where} \quad n = \text{number of data samples.}$$

$$\text{Encoded data:} \quad D_e = \sum_{i=1}^{n} E_i$$

$$\text{where} \quad E_i = A_i - A_{i-1}$$

$$\text{and} \quad A_0 = 0.$$

$$\text{Decoded data:} \quad D_d = \sum_{i=1}^{n} C_i = D_a$$

$$\text{where} \quad C_i = E_i + C_{i-1} = A_i$$

$$\text{and} \quad C_0 = 0.$$

Samples (i)	1	2	3	4	5	6	7	8	9	10
Actual Data (A_i)	0	3	6	8	10	7	4	1	-1	2
Encoded Data (E_i)	0	3	3	2	2	-3	-3	-3	-2	3

Table 10 Example of magnitude difference encoding scheme for $n = 10$.

Computer software had also been developed to encode the data and decode it afterwards. Results of these procedures were verified by data print-out and graphic display of the waveforms. When the encoded waveforms were compared with the original ones, it was found that except for those phonemes (/s/ and /d_3/) which had strong high frequency components and /α/ which had a high intensity, this encoding scheme yielded a substantial reduction of the maxima of most of the other phonemes. Distributions of the maxima of the encoded data are shown in Table 11.

Phoneme	Maxima (bits)		Phoneme	Maxima (bits)
b	4		i	2
s	4		e	3
d	4		ε	4
d_3	4		α	5
k	4		o	3
p	3		u	3
t	3		ℓ	3
v	2		m	3
w	5		n	2

Table 11 Distributions of maxima of the basic phonemes after difference coding.

The encoded data requires a storage of 20.1 k-bits compared with 27.3 k-bits before encoding. This gives a further saving of 7.2 k-bits.

Since the decoding process merely involves simple addition, it is an error-free process. With this encoding scheme, basic phonemes of better quality could be extracted by raising the sampling rate. This is because the magnitude differences between successive samples will decrease with the rise of sampling rate.

5.2 Bit Reduction by Decreasing the Sampling Rate

By reducing the bandwidth of the basic phonemes, bit reduction can be achieved by lowering the sampling rate. As it has been found [38] that certain phonemes occupy a smaller bandwidth than the others, it is possible to reduce the bandwidth from 0 - 6 kHz. to a smaller range at least for a number of phonemes.

Reduction in bandwidth was investigated experimentally. This will be described in the next chapter. This experiment was designed to study simultaneously the effects of some other factors which affected intelligibility and reading speed of spelled speech. The bandwidths studied were: 0 - 3 kHz., 0 - 4 kHz., 0 - 5 kHz. and 0 - 6 kHz. If it was found that the range 0 - 3 kHz. was good enough for all the phonemes, then data storage should be cut down by half. Unfortunately this method cannot be combined with the magnitude difference encoding scheme because a decrease in sampling rate results in an increase in magnitude differences, otherwise bit storage could be cut down more effectively.

6 Further Experiment to Study Intelligibility of Letter Sounds and Effects of Presentation Speed, Bandwidth Reduction and Pause Time Between Words

Using three blind subjects, a preliminary experiment described in Chapter 4 indicated that spelled sentences presented between 60 and 70 wpm. were highly intelligible. It also indicated that the synthesized letter sounds could be easily learned. In order to confirm these results, a more intensive study was made and an experiment employing 16 blind subjects was conducted. This experiment investigated four factors, namely, the intelligibility of letter sounds, the presentation speed, the bandwidth of letter sounds, and the pause duration between words. The investigation of pause duration stemmed from the hypothesis that a pause proportional to the length of word might be helpful in decoding long words. The results of this experiment agreed well with those of the preliminary one. The intelligibility of the constructed set of letter sounds proved to be strongly resistant to bandwidth reduction and subjects could learn to recognize all 26 synthesized letter sounds of the alphabet even when the bandwidth was reduced by half. However, intelligibility scores of spelled sentences decreased significantly with either reduction in bandwidth or increase in presentation speed. A pause proportional to the length of the preceding word did not give significantly better intelligibility than a pause of fixed duration.

6.1 Experimental Design and Testing Procedure

The investigated factors were divided into a number of treatment levels. Presentation speed was divided into four levels: 45, 55, 65 and 75 wpm. The bandwidths studied were 0 - 3 kHz., 0 - 4 kHz., 0 - 5 kHz. and 0 - 6 kHz. Two variations of pause duration after each word were studied: in one case the pause duration was fixed and in the other case, the pause duration was made linearly proportional to the length of the preceding word. A Greco-Latin square design [39] was used in this experiment, the plan of this design is shown in Appendix II. This design eliminated order effects in presentation speed and bandwidth. Four lists of phonetically balanced sentences [34] were used. Each list had a length of 79 words contained in ten unrelated sentences. Sixteen blind subjects, aged between 14 and 27, were used and were divided into eight groups for the experiment. Most of the subjects were students of the University of British Columbia. Most of them had a good knowledge of Braille (reading speed above 100 wpm.). Only those who had been blind for only several years were poor in Braille (30 - 60 wpm.). All subjects were individually trained and tested.

This experiment was divided into three sessions of about one and a half hours each. At the beginning of the first session, the entire set of letter sounds in random order was presented to the subject for identification. Four subjects were tested for each bandwidth. The result of this part of the experiment furnished the subjects' identification scores of the synthesized letter sounds before training. Following this, the subject was taught to recognize the letter sounds at 6 kHz. bandwidth. Subsequently the subject was given control of the keyboard of the teletype and could listen to any letter sounds by striking the corresponding keys. In this way, the subject could compare and contrast those letters which were easily

confused. A short quiz of letter identification was given to the subject from time to time. As the experiment progressed, the bandwidth was gradually reduced to 3 kHz. and words and sentences were also introduced at a gradually increasing speed. By the end of this session, most subjects could recognize all the letter sounds and had experienced all four bandwidths and four speeds. The second session was emphasized on sentence reading. Letter sounds were also reviewed from time to time. At the end of the second session, the subject was given a letter test to evaluate letter distinctiveness and learning effect. Sentence tests were conducted in the third session. Prior to each test, practice sentences were given to familiarize the subject with the test condition. Test sentences were presented in the same manner as described in section 4.2.1 of Chapter 4. There was a rest period of five minutes between lists.

6.2 Experimental Results

6.2.1 Letter Identification before Training

In order to find out how natural and distinct the synthesized letter sounds are, at the beginning of the first session, letter sounds in random order were presented to the subjects for identification. The results of this part of the experiment, in the form of confusion matrices, are shown in Tables 12 - 16. The total scores (out of 104) for each bandwidth are: 57 at 3 kHz., 70 at 4 kHz., 64 at 5 kHz. and 64 at 6 kHz. The average score is 63.75, i.e. 61.3 % correct.

Stimulus \ Response	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	3				1																					
B		4																								
C			1				1													2						
D		2		2																						
E		2			2																					
F						3													1							
G							4																			
H								4																		
I									3									1								
J							4																			
K							1			1						1				1						
L						1						1		2												
M														4												
N														4												
O															4											
P																4										
Q																4										
R	1																	3								
S	1					1								1					1							
T																2				2						
U		2			2																					
V		1																		2		1				
W									3														1			
X																								4		
Y																									4	
Z						1													1							2

Table 12 Confusion matrix of letter sounds at the 3 kHz. bandwidth.

Stimulus \ Response	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	3				1																					
B		2														2										
C			3																	1						
D		2		2																						
E	1				3																					
F						4																				
G							3													1						
H								4																		
I									4																	
J							1			3																
K											3									1						
L						1						3														
M						1							1	2												
N														3					1							
O															4											
P																3				1						
Q																3				1						
R																		4								
S																			4							
T																				4						
U		1			3																					
V							1									2						1				
W									2														2			
X																			1					3		
Y																									4	
Z						1					1	1							1							

Table 13 Confusion matrix of letter sounds at the 4 kHz. bandwidth.

Stimulus \ Response	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	2				2																					
B		1		1	1											1										
C			1																	3						
D				2												2										
E					4																					
F	1					3																				
G							3													1						
H								4																		
I									3									1								
J							3			1																
K					1						1					1				1						
L												3		1												
M														4												
N														4												
O															4											
P																4										
Q																2				2						
R																		4								
S																			4							
T																				4						
U					3																1					
V			1	1			1									1										
W									3														1			
X																			1					3		
Y																									4	
Z	1																									3

Table 14 Confusion matrix of letter sounds at the 5 kHz. bandwidth.

Stimulus \ Response	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	2				2																					
B		4																								
C			3																	1						
D				3												1										
E		1			2																1					
F						3													1							
G							3													1						
H								4																		
I									4																	
J							3													1						
K							1			1						1				1						
L												1		3												
M	1													3												
N														4												
O	1				1										2											
P																4										
Q																4										
R									1									3								
S																			4							
T																1				3						
U					2																2					
V		1		1																		2				
W									2														2			
X																			1					3		
Y																									4	
Z	1					1																				2

Table 15 Confusion matrix of letter sounds at the 6 kHz. bandwidth.

Stimulus \ Response	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	10				6																					
B		11		1	1											3										
C			8				1													7						
D		4		9												3										
E	1	3			11															1						
F	1					13													2							
G							13													3						
H								16																		
I									14									2								
J							11			4										1						
K					1		2			2	4					3				4						
L						2						8		6												
M	1					1							1	13												
N														15					1							
O	1				1										14											
P																15				1						
Q																13				3						
R	1								1									14								
S	1					1								1					13							
T																3				13						
U		3			10																3					
V		2	1	2			2									3				2		4				
W									10														6			
X																			3					13		
Y																									16	
Z	2					3					1	1							2							7

Table 16. Combined confusion matrix of letter sounds for the four bandwidths.

The combined scores shown in Table 16 indicate that some letters could be identified more easily than others. This table also indicates letters having the same phoneme sound at the beginning (such as /dʒ/ in letters G and J, and /ɛ/ in F, L, M, N, S etc.) or the same phoneme sound at the end (such as /i/ in B, C, D, etc.) are more easily confused among themselves. Naturalness and distinctiveness can be measured according to the identification scores of this test. Letters ranked in this way are shown in Table 17 below. From this table, it can be seen that

Score	16	15	14	13	11
Letter	H Y	N P	I O R	F G S T X	B E

Score	10	9	8	7	6	4	3	1	0
Letter	A	D	C L	Z	W	J K V	U	M	Q

Table 17 Combined identification scores of letter sounds in descending order of correctness (perfect score: 16).

synthesis (using /ɑɛe/) of diphthong /ɑI/ which occurs in both Y and I was very successful. The low scores of U and Q indicate that direct combination of /i/ and /u/ does not give a good sound of diphthong /ju/ or /Iu/ which occur in U and Q respectively. The low scores of some other letters were mainly due to confusions, e.g. letter M was most of the time misidentified as N (see Table 16), letter J misidentified as G, letter U misidentified as letter E, etc.

Although the identification scores of some letters (particularly U, M and Q) seem disappointingly low, it will be shown that after a short period of training, the majority of the subjects could identify all the letters without error.

6.2.2 Letter Identification after Training

During the progress of the experiment, it was observed that many subjects could easily learn to recognize all the letters correctly. In the letter test towards the end of the second session, five sounds of each letter in random order were presented to the subject for identification. Results of this test are shown in Table 18. In this test, 12 subjects had perfect scores, 3 subjects made only 1 mistake (out of 130 responses) and 1 subject made 8 mistakes. Of the total number of 11 mistakes made, 9 were made at 3 kHz. and 2 at 6 kHz. The extremely high average correct identification score of 99.5% suggests that all letter sounds can be learned to perfect recognition after a short period of training. It also indicates that the intelligibility of the constructed set of letter sounds is highly resistant to bandwidth reduction.

Stimulus \ Response	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	80																									
B		80																								
C			80																							
D		2		77												1										
E					80																					
F						80																				
G							80																			
H								80																		
I									80																	
J										80																
K											80															
L												80														
M													78	2												
N														80												
O															80											
P																80										
Q																2	78									
R																		80								
S						2													78							
T																1				79						
U																					80					
V			1																			79				
W																							80			
X																								80		
Y																									80	
Z																										80

Table 18 Combined confusion matrix of letter sounds for the four bandwidths after learning.

6.2.3 Test on Sentence Reading

Results of this test are summarized in Tables 19 and 20. It can be seen that the scores vary widely from subject to subject. A detailed analysis of the data showed that the subjects who were bad in Braille (particularly those who became blind in their adulthood), were also relatively bad in spelled speech; their results resembled those obtained with sighted subjects†. This indicates that those who can read Braille well will also be good in reading spelled speech. Also there were a number of exceptional subjects who could read spelled speech comfortably at a speed of 80 wpm.

Table 20 shows that the average scores of the blind subjects are very high, ranging from 83 to 94.15% correct. It must be borne in mind that although it was an intelligibility test, a certain level of comprehension had already taken place because the subjects were tested on sentences rather than on single words. Occasionally the subjects were asked what the sentences were about and from their answers one knew that very often they understood the sentences even in cases when one or two words were unintelligible.

---

† A spelled speech experiment was conducted some time ago using a highly motivated group of sighted subjects [8]. The same testing procedure and the same type of testing materials (PB sentences) were used. For approximately the same presentation speed (55 wpm. in the present experiment with blind subjects and 54 wpm. in the experiment with sighted subjects), the average percent correct score of the blind subjects is 12.93% higher than that of the sighted subjects'.

		L_1	L_2	L_3	L_4
I_1	G_1	80.38	77.22	82.91	87.97
	G_2	95.57	79.11	85.44	76.58
	G_3	96.20	96.20	84.18	91.77
	G_4	88.61	93.04	96.20	99.37
I_2	G_5	84.18	79.11	90.51	90.51
	G_6	94.94	91.77	89.87	86.71
	G_7	92.45	98.10	89.87	90.51
	G_8	81.65	82.91	90.51	95.57

Table 19 Scores of spelled speech experiment in percent correctness.

List	L_1	L_2	L_3	L_4
I_1	90.19	86.39	87.18	87.92
I_2	88.31	87.97	90.19	90.82
Average	89.25	87.18	88.69	89.87

Speed (wpm.)	a_1 (45)	a_2 (55)	a_3 (65)	a_4 (75)
I_1	93.51	89.87	87.66	81.65
I_2	94.78	91.30	86.87	84.34
Average	94.15	90.59	87.27	83.00

Bandwidth (kHz.)	b_1 (3)	b_2 (4)	b_3 (5)	b_4 (6)
I_1	85.60	87.34	90.19	89.56
I_2	88.61	89.87	89.56	89.24
Average	87.11	88.61	89.88	89.40

Table 20 Summary data of spelled speech experiment in percent correctness.

An analysis of variance of the data is shown in Appendix III. This analysis indicates that: bandwidth and list of testing materials are both significant ($p<0.05$), the effect of speed of presentation is highly significant ($p<0.01$), but effects of row position and interval of pause between words are not significant. The bandwidth effect is significant because lowering the bandwidth reduces the pleasantness and clarity of letter sounds. Since different lists of testing materials have different sentences containing words differing both in length and familiarity, it is not surprised to see the list effect being statistically significant. As expected, presentation speed is highly significant because intelligibility decreases sharply with increase in speed of presentation. The insignificance of row effect suggests that the order effect of subject groups in this test is not significant. As far as the interval of pause between words is concerned, although a pause proportional to the length of the preceding word gives some improvement over a fixed interval (average score: variable interval, 89.32% correct; fixed interval, 88.17% correct), its effect is not significant to the 0.05 level.

Since presentation speed and bandwidth are the two factors which are most interested in this experiment, further tests were made to probe the nature of differences among treatment means of these two factors. The results of Newman-Keuls' tests are shown in Appendix IV and can be summarized schematically as follows:

wpm.	45	55	65	75
45	—	**	**	**
55		—	**	**
65			—	**
75				—

kHz.	5	6	4	3
5	-			**
6		-		*
4			-	
3				-

** $p<0.01$ * $p<0.05$

Thus it can be concluded that the various speeds of presentation differ significantly from one another while for the different bandwidths, only 3 kHz. differs significantly from 5 kHz. and 6 kHz. During the progress of the experiment, the subjects also reported that the 3 kHz. letter sounds were less pleasant than those having higher bandwidths.

6.2.4 Word Length and Intelligibility

In this spelled speech experiment, a large number of errors occurred in those words which contained a large number of letters. In view of this, a calculation was made based on the number of errors observed at different word lengths (the number of letters contained in a word). The effect of this factor on the entire experiment is shown below.

Word Length (letters)	4 or less	5	6	7	8
% Correctness	92.13	83.58	80.21	76.84	62.50

Table 21 Overall percent correctness of the words according to word length.

From Table 21, it can be seen that the word length has a very great effect on the correctness of the response. The longer the word, the more likely that an error will be made. It must be pointed out that this is true even in the case where a longer pause was given to decode the long words in the test. In some cases, the subjects reported insufficient time for the perception of the long words from the letters, and while they were still pondering on the long words, words of the subsequent order followed and disruption occurred. In other cases they forgot the letters which occurred at the beginning or in the middle of the long words. There are two main reasons why long words were less intelligible. First, long words occur less frequently in the English language and so they were less familiar to the subjects. Second, some subjects perceived the long words letter by letter and so they forgot some of the letters as the word length exceeded their memory limit. This second factor will become less prominent as the subjects become more skilful in decoding spelled speech. It is expected that through practice, they will eventually perceive chunks of letters (e.g. in the form of syllables) at a time and not a letter at a time.

7 Summary, Discussions and Suggestion

7.1 Summary

With the aid of a PDP-9 digital computer and its graphic display and interface accessories, a segmentation program was developed to examine the properties of letter sounds, to extract various segments of a speech sample and to display their waveforms on a precision oscilloscope, and to present the processed speech signals to a listener. In order to cut down the amount of memory required to store the letter sounds, redundant parts of spoken letter sounds were eliminated and a set of 18 basic phonemes was chosen to synthesize the letter sounds. Further saving was accomplished by storing only a pitch period of each vowel or vowel-like phoneme. In this way, a very big memory reduction was possible and the amount of memory needed for good quality letter sounds synthesized by concatenation of the basic phonemes was 64.8 kilo-bits compared with 461.3 kilo-bits required to store a set of naturally produced letter sounds. By taking into account the amplitudes of individual phonemes, the 18 basic phonemes could be stored with 27.3 kilo-bits.

Two other methods of further reducing the memory storage were investigated. The first method involved a difference coding scheme which stored only the differences between successive samples. This method brought the memory storage down to 20.1 kilo-bits. The other method was studied by reducing the bandwidth of the letter sounds. Subjective experiments indicated that there was a statistically significant difference between intelligibility score at the 3 kHz. bandwidth and scores at 5 and 6 kHz. bandwidths. Also, the quality of the letter sounds at 3 kHz. bandwidth was reported to be worse than that between 5 and 6 kHz. Thus it was concluded that the difference coding scheme should be chosen to preserve the quality of the synthesized letter sounds.

An experiment was conducted to study the discrimination of six vowels and indicated that subjects could learn to discriminate these speech stimuli which were only 10 ms. in duration.

In the synthesis of letter sounds, some effort was given to make the synthesized letter sounds easily distinguishable. The distinctness of the letter sounds was tested in a pilot experiment using three blind subjects. This test indicated that blind subjects could read spelled sentences between 60 and 70 wpm. at high intelligibility after only one hour of contact with the letter sounds. The subjects also demonstrated that they could learn to recognize all 26 synthesized letter sounds after a short period of training. This was confirmed in a more intensive experiment which was designed to investigate simultaneously the effects of bandwidth reduction and rapid presentation of the spelled speech code and also the effect of giving a longer pause to compensate for a longer time needed to decode long words. Sixteen blind subjects were used in this experiment and the PDP-9 computer was programmed to simulate the digital spelled speech reading machine. Test sentences were typed in and spelled sentences were generated at the output and presented to the subjects. The results of this experiment indicated that an average young blind subject could read spelled sentences between 65 and 75 wpm. with an intelligibility score of about 85% correct. A longer pause did not increase substantially the intelligibility of long words. Also, reduction of the bandwidth reduced the intelligibility. Thus for preserving good quality of the letter sounds, a bandwidth between 4 and 6 kHz. should be used. Unfortunately, the difference encoding scheme and bandwidth reduction could not be combined to lower the memory storage and thus it was concluded that for this developed set of letter sounds, the difference coding scheme was preferred and a 6 kHz. bandwidth should be retained.

7.2 Discussions

7.2.1 Spelled Speech and Elderly Subjects

It is obvious that spelled speech is of no use to blind people who do not know the spelling of words. Although this would eliminate the group of young children from using the machine to full advantage, this machine can nevertheless be used to train their spelling. So far, the results on spelled speech were obtained from young blind subjects. In order to see whether elderly people can make good use of this machine, two elderly blind subjects (one 51 years old and the other 58 years old) were used to explore this possibility. They were trained in the same manner as the 16 subjects except that each had only two sessions instead of three, and only the 6 kHz. bandwidth was used. This is because they (and possibly other elderly blind subjects) had problems in recognizing the letter sounds even at the 6 kHz. bandwidth. This might have resulted from aging of their auditory system. When isolated words were presented, they could recognize them up to 40 wpm. But when sentences were presented, they could read only at about 20 wpm. This is in great contrast with the results obtained with young blind subjects. Although both elderly subjects found this mode of communication very interesting, it seems unlikely that they can be trained to read at a much faster rate, and certainly not the rate attained by young blind subjects. Whether they would like to use this machine for pleasure reading will have to be explored, but definitely this machine can help them to read short sentences or words like denomination of paper money, bank account, bills, names and addresses on envelope, telephone numbers, etc.

7.2.2 Other Speech Aids for the Blind

It has been shown that spelled sentences can be read at a rate much faster than that can be achieved with the Lexiphone code. Also, it requires only a short period of training to recognize all the letter sounds. Although synthesized speech would be a much better and more natural mode of machine-to-man communication, it remains to see whether cost and complexity of speech synthesis are justified. How expensive, how convenient and how natural will synthesized speech be? Will it be more expensive than employing a paid reader or using the Talking Books? It is very difficult to give definite answers to these questions at the present moment and indeed more ingenious works in this field have yet to come.

One other type of speech aid for the blind now under development produces a language-like auditory code called "Spelltalk" [40, 41]. Similar to spelled speech, Spelltalk has one fixed sound for each printed letter. The phonetic system of this code is based on the frequency of sound occurrence in the English language. Thus phoneme /I/ is chosen for letter I because in the English language, letter I has the following frequency of sound distribution: /I/, 68%; /αI/, 26%; and others, 6%. Although this phonetic system will produce a number of intelligible words because of its resemblance to the English language, there are many words which will sound unexpectedly strange and difficult to pronounce and guess at especially when presented at a high rate (e.g. /kɔengɛ/ stands for the word change, /ʒIkʒ/ for kick and /bɛeʌtj/ for beauty, etc.). A fair amount of training is required to understand Spelltalk and further evaluation is necessary to find out whether blind people can be trained to understand this kind of machine language at a high rate.

7.3 Suggestions: Contracted Spelled Speech

One way of increasing the reading rate of spelled speech is to include contractions of spelled speech (i.e. spoken words) like the contractions used in Braille. It is reckoned that inclusion of a large vocabulary of words will come back to the problem faced with the talking machine, i.e. a big increase in construction cost and complexity. However, with a limited vocabulary of spoken words formed by concatenation of the basic phonemes used for the letter sounds of the alphabet, the cost would not be increased appreciably. In Table 22, a suggested vocabulary for contracted spelled speech is shown. This vocabulary consists of some of the most frequently used words [37]. The way these words are formed from the basic phonemes is also presented in Table 22. These words were also synthesized by the computer and presented to several listeners. Most of the words were recognized after only a short period of training. Most of these words also occur in Grade II contracted Braille. Based on the frequency of occurrence of these words, it has been estimated that a listening rate of about 100 wpm. could be expected for a mixture of spelled speech and these spoken words. From Table 22, it can be seen that some spoken words do not sound exactly like natural speech because of limitation of selected basic phonemes. These words however can be improved by constructing more vowels (such as /I/, /ə/ and /ɔ/) without increasing appreciably the amount of memory required.

Word	the	of	and	to	in	that	it	is	for	be
Sound	dɛ	ov	ɛnd	tu	in	dɛt	it	is	vo	bi
Word	was	as	you	with	on	by	at	this	are	we
Sound	wos	ɛs	iu	wid	on	bɑɛe	ɛt	dis	ɑ	wi

Table 22 List of words to be synthesized from basic phonemes.

REFERENCES

1. Beddoes, M. P., and Suen, C. Y.

"Evaluation and a method of presentation of the sound output from the Lexiphone- a reading machine for the blind"

IEEE Trans. Bio-Med. Eng., 18, 85-91, 1971.

2. Linvill, J. G.

"Development progress on a microelectronic tactile facsimile reading aid for the blind"

IEEE Trans. Audio and Electroacoustics, 17, 271-274, 1969.

3. Nye, P. W., and Bliss, J. C.

"Sensory aids for the blind: a challenging problem with lessons for the future"

Proc. IEEE, 58, 1878-1898, 1970.

4. Cooper, F. S., Gaitenby, J. H., Mattingly, I. G., and Umeda, N.

"Reading aids for the blind: a special case of machine-to-man communication"

IEEE Trans. Audio and Electroacoustics, 17, 266-270, 1969.

5. Smith, G. C., and Mauch, H. A.

"Summary report on the development of a reading machine for the blind"

Mauch Laboratories Summary Report to the Prosthetic and Sensory Aids Service, Veterans Administration, July 1971.

6. Beddoes, M. P., Fletcher, T. R., and Suen, C. Y.

"A spelled speech reading machine for the blind"

Paper presented at the International Electrical, Electronics Conference and Exposition, Toronto, Oct. 4-6, 1971.

7. Suen, C. Y., and Beddoes, M. P.

"Some applications of a small digital computer in speech processing"

J. Acous. Soc. Am., 50, 107, 1971. (A)

Expanded version of this paper is now in press, in "Time-compressed Speech: Anthology and Bibliography" by Sam Duker, Scarecrow Press, New Jersey.

8. Suen, C. Y.

"Towards an improved method of presenting the Lexiphone code and spelled speech"

M.A.Sc. Thesis, University of British Columbia, May 1970.

9. Suen, C. Y., and Beddoes, M. P.

"Discrimination of vowels of very short duration"

Perception & Psychophysics, 11, 417-419, 1972.

10. Parmenter, C. E., and Treviño, S. N.

"The length of the sounds of a Middle Westerner"

Amer. Speech, 10, 129-133, 1935.

11. Lehmann,W. P., and Heffner, R-M. S.

"Notes on the length of vowels (VI)"

Amer. Speech, 18, 208-215, 1943.

12. Black, J. W.

"Natural frequency, duration, and intensity of vowels in reading"

J. Speech Hear. Dis., 14, 216-221, 1949.

13. House, A. S., and Fairbanks, G.

"The influence of consonantal environment upon the secondary acoustical characteristics of vowels"

J. Acous. Soc. Am., 25, 105-113, 1953.

14. Zimmerman, S. A., and Sapon, S. M.

"Note on vowel duration seen cross-linguistically"

J. Acous. Soc. Am., 30, 152-153, 1958.

15. Tiffany, W. R.

"Sources of variation in vowel quality"

J. Speech Hear. Res., 2, 305-317, 1959.

16. Peterson, G. E., and Lehiste, I.

"Duration of syllable nuclei in English"

J. Acous. Soc. Am., 32, 693-703, 1960.

17. House, A. S.

"On vowel duration in English"

J. Acous. Soc. Am., 33, 1174-1178, 1961.

18. Sharf, D. J.

"Vowel duration in whispered and in normal speech"

Language and Speech, 7, 89-97, 1964.

19. Siegenthaler, B. M.

"A study of the intelligibility of sustained vowels"

Quart. J. Speech, 36, 202-208, 1950.

20. Tiffany, W. R.

"Vowel recognition as a function of duration, frequency modulation and phonetic context"

J. Speech Hear. Dis., 18, 289-301, 1953.

21. Schwartz, M. F.

"A study of thresholds of identification for vowels as a function of their duration"

J. Aud. Res., 3, 47-52, 1963.

22. Fujisaki, H., and Kawashima, T.

"The influence of various factors on the identification and discrimination of synthetic speech sounds"

The 6th International Congress on Acoustics, Tokyo, Japan, 1968.

23. Powell, R. L., and Tosi, O.

"Vowel recognition threshold as a function of temporal segments"

J. Speech Hear. Res., 13, 715-724, 1970.

24. Peterson, G. E.

"The significance of various portions of the wave length in the minimum duration necessary for the recognition of vowel sounds"

Ph.D. Dissertation, Department of Speech, Louisiana State University, 1939.

25. Gray, G. W.

"Phonemic microtomy: the minimum duration of perceptible speech sounds"

Speech Monographs, 9, 75-90, 1942.

26. Joos, M.

"Acoustic phonetics"

Language Monograph, 24, No. 2 Suppl., 77-78, 1948.

27. Flanagan, J. L.

Speech analysis synthesis and perception

Academic Press Inc., New York (1965), p. 214.

28. Thomas, I. B., Hill, P. B., Carroll, F. S., and Garcia, B.

"Temporal order in the perception of vowels"

J. Acous. Soc. Am., 48, 1010-1013, 1970.

29. House, A. S., Stevens, K. N., Sandel, T. T., and Arnold, J. B.

"On the learning of speechlike vocabularies"

J. verb. Learn. verb. Behaviour, 1, 133-143, 1962.

30. Stevens, S. S., and Davis, H.

Hearing - its psychology and physiology

John Wiley, New York (1937), p. 102.

31. Suen, C. Y., and Beddoes, M. P.

"Output sounds of a digital spelled speech reading machine for the blind"

Proceedings of the International Conference on Speech Communication and Processing, Boston, April 24-26, 1972.

32. Miller, G. A.

Language and communication

McGraw Hill, New York (1951), p. 64.

33. Slis, I. H., and Cohen, A.

"On the complex regulating the voiced-voiceless distinction I"

Language and Speech, 12, 80-102, 1969.

34. "1965 revised list of phonetically balanced sentences (Harvard sentences)"

IEEE Trans. Audio and Electroacoustics, 17, 239-246, 1969.

35. Foulke, E.

"A review of research on time compressed speech"

Proceedings of the Louisville Conference on Time Compressed Speech, pp. 3-20, Oct. 1966.

36. Foulke, E., and Sticht, T. G.

"The intelligibility and comprehension of time compressed speech"

Proceedings of the Louisville Conference on Time Compressed Speech, pp. 21-28, Oct. 1966.

37. Dewey, G.

Relativ frequency of English speech sounds

Harvard University Press, Cambridge (1950), p. 45.

38. Fletcher, H.

Speech and hearing in communication

D. Van Nostrand Co. Inc. (1953), p. 87.

39. Winer, B. J.

Statistical principles in experimental design

McGraw Hill, New York (1962), pp. 575-577.

40. Bellavia, D. C.

"A prosthetic reading aid for the blind"

Ph.D. Dissertation, Bio-Medical Engineering, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 1970.

41. Longini, R. L.

"Spelltalk: a new approach to reading machine output for the blind"

AFB Research Bulletin, No. 24, 153-157, March 1972.

Appendix I List of Phonetic Symbols used

Consonants		Vowels, Liquid, Nasals and Diphthongs	
Phonetic Symbol	Key Word	Phonetic Symbol	Key Word
b	bee	i	beet
s	see	e	chaotic
d	deed	ɛ	set
dʒ	jade	ɑ	father
k	case	o	notation
p	pea	u	pool
t	tea	I	it
v	vela	ə	the
w	wide	ɔ	for
f	fife	ʌ	up
tʃ	church	ɭ	elder
g	get	m	empty
j	you	n	end
ʒ	vision	ɑI	eye
		ju	you
		Iu	mute

Appendix II Plan of Spelled Speech Experiment with 16 Blind Subjects

		L_1	L_2	L_3	L_4
I_1	G_1	a_3b_3	a_4b_1	a_1b_4	a_2b_2
	G_2	a_1b_2	a_2b_4	a_3b_1	a_4b_3
	G_3	a_2b_1	a_1b_3	a_4b_2	a_3b_4
	G_4	a_4b_4	a_3b_2	a_2b_3	a_1b_1
I_2	G_5	a_3b_3	a_4b_1	a_1b_4	a_2b_2
	G_6	a_1b_2	a_2b_4	a_3b_1	a_4b_3
	G_7	a_2b_1	a_1b_3	a_4b_2	a_3b_4
	G_8	a_4b_4	a_3b_2	a_2b_3	a_1b_1

I: Interval of pause between words.

I_1 = fixed interval = 4.4 T_L

I_2 = interval increases linearly with word length = m L_W + 2 T_L

where T_L = pause between letters

L_W = word length

and m = 2.4 T_L/4.4 = 0.545 T_L.

G: Subject group, there were two subjects per group.

L: List of testing materials, lists 13, 16, 22 and 45 of PB sentences were used.

a: Speed of presentation, four speeds were used, viz. 45, 55, 65 and 75 wpm.

b: Bandwidth of letter sounds, four bandwidths were used, viz. 0 - 3, 0 - 4, 0 - 5 and 0 - 6 kHz.

Appendix III Analysis of Variance: Data Analysis of Spelled Speech Experiment

Source	df	MS	F
Between Subjects	15		
I Interval of pause	1	13.141	
R Row	3	138.270	
IR	3	82.474	
Subjects within groups	8	84.077	
Within Subjects	48		
a Speed of presentation	3	225.890	71.688**
b Bandwidth	3	14.682	4.659*
L List of testing materials	3	13.182	4.183*
aI	3	5.182	
bI	3	8.891	
LI	3	11.224	3.562*
(AB)'	3	24.157	
(AB)'I	3	26.990	
Error	24	3.151	

** $p<0.01$ * $p<0.05$

Appendix IV Results of Newman-Keuls' Test of Spelled Speech Scores

Speed (wpm.)	a_1 (45)	a_2 (55)	a_3 (65)	a_4 (75)
a_1 (45)	—	45**	87**	141**
a_2 (55)		—	42**	96**
a_3 (65)			—	54**
a_4 (75)				—

Bandwidth (kHz.)	b_3 (5)	b_4 (6)	b_2 (4)	b_1 (3)
b_3 (5)	—	6	16	35**
b_4 (6)		—	10	29*
b_2 (4)			—	19
b_1 (3)				—

** $p<0.01$ * $p<0.05$

PUBLICATIONS

Suen, C.Y. and M.P. Beddoes
Some applications of a small digital computer in speech processing (in) "Time-Compressed Speech-Anthology and Bibliography": S. Duker. Scarecrow Press (in press).

Suen,C.Y. and M.P. Beddoes. "Output sounds for a digital spelled speech reading machine for the blind". Proc. 1972 International Conference on Speech Communication and Processing, 1972.

Suen, C.Y. and M.P. Beddoes. "Discrimination of vowel sounds of very short duration". Perception and Psychophysics, 11, 417-419, 1972.

Beddoes, M.P., T.R. Fletcher and C.Y. Suen. "A spelled speech reading machine for the blind". Proc. International Electrical, Electronics Conference, 1971.

Suen, C.Y. and M.P. Beddoes. "Some applications of a small digital computer in speech processing". Paper presented at the 81st Meeting of the Acoustical Society of America, Washington, 1971.

Beddoes,M.P. and C.Y. Suen. "Evaluation and a method of presentation of the sound output from the Lexiphone - a reading machine for the blind". IEEE Trans. Bio-Medical Engineering, Vol. BME-18, 85-91, 1971.

Suen, C.Y. "Derivation of harmonic equations in non-linear circuits". J. Audio Engineering Soc., 18, 675-676, 1970.

Yu, P.K. and C.Y. Suen. "Analysis of the Darlington Configurations". Electronic Engineering, 40, 38-39, 1968.

Suen, C.Y. "Characteristics of the Darlington composite transistor". Int. J. Electronics, 24, 373-380, 1968.