Some facilities for speech processing by computer

By S. H. Lavington and L. E. Rosenthal*

This paper describes means by which the capabilities of digital processing may be made available to users in the field of speech research. The design of input/output hardware, and a special software package, are outlined and examples are given of their use in current projects.

1. Introduction

At the present time there are several projects in existence involving the analysis of speech waveforms. Such fields of study are: (*a*) development of communication systems that use digitized and/or parametric speech; (*b*) experiments with automatic recognition of spoken commands; (*c*) study of phonetics through models of the vocal system and synthetic speech; (*d*) segmentation of words, and human perception tests. For many of these applications conventional devices for examining speech, such as the fixed-filter spectral analyzer, are inadequate. Additional specialized equipment is required and there is increasing interest in prior digital simulation of these experimental systems as an aid to design. Alternatively digital, rather than analogue, processing of the speech waveform may be used. This method is capable of greater flexibility and accuracy, and new techniques such as zero-crossing analysis readily lend themselves to computer calculations. In the setting-up of digital speech processing facilities, two problems may be distinguished: (1) provision of an input/output device to convert the analogue speech waveform to a numerical representation which may be stored on magnetic tape; (2) development of programs to perform various kinds of analysis on the digital record.

With regard to the first problem, it is found that most speech-processing work requires relatively infrequent use of the input/output hardware compared with computing time, once a selection of test messages or utterances has been digitized and stored. As magnetic tapes may be interchanged between installations, only one central computer need have a speech input/output device as a peripheral; tapes of stored utterances produced on this installation are then made available for other users. One such system using Manchester University's Atlas is described in Section 2 of this paper.

On the programming side, the large amount of work needed on the part of potential users has proved a barrier. There are, however, a number of basic processing procedures that have been found useful for a wide range of speech problems. If these are grouped into a special program, entered via very simple user-statements, then much of the day-to-day programming effort can be eliminated. Such a software package, comprising eight routines for performing such tasks as amplitude-envelope measurements, spectral analysis, and zero-crossing counts, is described in Section 3.

Section 4 outlines the use of the above facilities in two current projects. The first is a system for speech analysis and synthesis based on a parametric acoustical model of the vocal tract. The second is the development of a set of measurements on the speech waveform, to be used in a verbal command recognition system.

2. Hardware facilities

2.1 Design requirements for a speech input/output device

In order to maintain flexibility for a large range of problems, an input peripheral should not significantly reduce the information content of the speech waveform. An analogue-to-digital converter having a sufficient bandwidth and bit-accuracy is thus required. A frequency range extending from about 70 c/s to 8,000 c/s will contain most of the speech information, and an amplitude definition to within $\pm 1\%$ is adequate and does not demand special circuit techniques in the conversion electronics (Lavington, 1964). There is an upper limit to the rate of information transfer between an on-line peripheral and a particular computer, and to determine whether the above requirements can be met it is necessary to consider the organization of input/output devices in an Atlas installation.

When in use, a peripheral has periodically to request the central computer for more information, or to indicate that it has information waiting to be transferred. To initiate the organizational routines for these tasks, an Interrupt digit is set by the peripheral. This causes the central processor to transfer to an *interrupt* phase, and the appropriate routine for that Interrupt is entered. There is a system of priorities governing the case where two or more equipments send Interrupts simultaneously, and there is a limit to the number and type of peripherals that can be handled at any one time without reducing the efficiency of the total system. This limit is a function of the *crisis time* for a particular device, which is generally equal to the time between arrival of two successive Interrupts. To ensure the satisfactory operation of the main system including drums, tape decks, and conventional input/output units, the crisis time for an additional time-shared peripheral should be greater than about 250 microseconds. (The total time spent in actually dealing with an Interrupt from this type of peripheral is about 40 μsec.) For an input bandwidth of 8,000 c/s, Shannon's Theorem (Shannon, 1949) gives an analogue-to-digital converter

* *Department of Computer Science, The University, Manchester*, 13.

sampling period of 62·5 μsec. It is not therefore possible to transfer each amplitude sample as soon as it is produced, and some local storage is necessary. Transfer via a peripheral coordinator unit to the main store is accomplished in half-words of 24 bits, and processing is simplified if the analogue-to-digital converter has a bit accuracy that is a sub-multiple of 24. With regard to the ± 1% tolerance given above, an 8-bit converter would be appropriate. The one-level store in the Manchester Atlas system (Kilburn, Edwards, Lanigan & Sumner, 1962) has a capacity of 112K words and for speech inputs lasting for more than a few seconds, transfer of the record to magnetic tape will be necessary. The transfer rate for one tape-channel is 384,000 bits/sec and is not a practical limitation.

It is seen that an analogue-to-digital converter of the required specification may be connected as a peripheral, provided that some temporary storage is built into the device. The equipment that has been built for the Atlas system is known as the *Speech Converter*. It has a bandwidth of 10 kc/s and incorporates two 48-bit storage registers and this means that, on input mode, six 8-bit amplitude samples of the analogue waveform are transferred to the main computer every 300 μsec (the crisis time). The samples are assembled into blocks of 512 48-bit words, and stored subsequent to a specified address on magnetic tape. On output mode, samples are transferred from magnetic tape to the peripheral, converted into an analogue waveform, and fed to a loudspeaker or tape-recorder.

2.2 User's specification of the Speech Converter

The system employed for analogue-to-digital conversion is the sample-and-hold, current-weighing method, with an input of ± 6·0 volts to the encoder representing the maximum of +127 or −128 amplitude units. A low-pass-filter, having a cut-off of 9·5 kc/s and an attenuation of more than 40 dB at frequencies greater than 10 kc/s, is provided immediately before the sampler. The encoding process, excluding the input amplifier, has a frequency response extending down to D.C.

The input amplifier provided may be set to accept signals either from a moving-coil microphone (order of 5 mV), or from a tape recorder (order of 5 V); the response is linear down to about 20 c/s. This amplifier board can be quickly replaced, should different magnitudes of incoming signal be envisaged. On output from the Speech Converter, the normal amplifier used is capable of delivering 4 watts (peak sinusoidal) to a loudspeaker; the response of this amplifier is linear down to about 35 c/s. The digital logic of the system is made up from standard Atlas boards. The overall noise of the Speech Converter is less than one-tenth of an encoding step (less than 4·7 mV).

Tests of the Speech Converter are provided by an engineer's test program (Mathers, 1964) which checks the correct operation of the control and information digits. In addition a ramp waveform generator of known linearity is used to assess the performance of the analogue encoding circuits. The program gives a print-out of the encoding error and maximum fluctuation due to noise, together with information that facilitates any adjustments which should be made. The Speech Converter has been in operation since December, 1963, and the only significant adjustments that have had to be made concern ageing of the close-tolerance resistors used to define the standard current-sources in the encoding process.

The front panel controls consist of the basic engage/disengage and input mode/output mode buttons, plus the following additional facilities:

(*a*) Provision for remote control of an external tape recorder, or other instrument. At the present time this is used in conjunction with a modified Ferrograph recorder to give a suitable pause to allow the tape to reach a steady speed, before commencing input or output transfers.

(*b*) The "Wait for Input" facility. This allows the input operation to be temporarily halted by inhibiting the Interrupt signal to the main machine, and is useful for operator-control during direct microphone input.

(*c*) Manual Stop, when the mode is input. This terminates transfer, but leaves the device still engaged. It is useful when the exact length of an input message is not known beforehand.

(*d*) Speed selection. The full 10 kc/s bandwidth is uneconomic for some applications, and the sampling rate may be halved to give a 5 kc/s response. The crisis time is then 600 μsec, and a 4·5 kc/s cut-off low pass filter is switched to replace the 9·5 kc/s filter.

(*e*) Input and Output amplifier gain control, and volume meter.

The basic operation and information-transfer of the Speech Converter is under Supervisor control (Mathers, 1964). The user is required to input via paper tape an additional short program in which is specified the title of the magnetic tape to be used and the sequence of input and/or output operations to be performed, giving the length of transfer in each case. The program is used to read and write to tape, and to print an operating record on the line printer. It is time-shared with other programs, and is given high priority by the Supervisor because of its magnetic tape activity. A reel of tape may contain as many as 5,000 addressed blocks (of 512 words), and this corresponds to a continuous transfer of 12·5 minutes of speech for the full bandwidth, or twice this for the lower sampling rate. If transfers of less than 20 blocks are performed, the speech may be held in the one-level store without being written to tape. This could be the case for experiments with a simple verbal command system.

3. Software facilities

3.1 General description

It has been found that a number of basic speech processing routines are frequently used, and it is desirable to minimize the programming effort and computing time needed for each procedure. To enable these requirements to be realized it is convenient if (*a*) the actual programs are available in compiled form, and (*b*) if user-input is reduced to a few simple control parameters. Speech-Processing programs developed at Manchester have therefore been collected into a single package which is stored on magnetic tape in a compiled state. By using the "Compiler Special" facilities in Atlas Autocode (Brooker & Rohl, 1966), this package is given the same status under the Supervisor as any normal compiler. The user need only call for COMPILER SPECIAL, SPP 1 in his job description and the package is brought down from tape and entered. A job therefore consists only of a job description and a few control parameters.

This approach has the disadvantage of wasting a certain amount of store since the entire Speech Processing Package (33 blocks, plus space for the speech data) is brought down although only one or two routines may be needed, but the simplicity of operation and savings in input and compiling time far outweigh the wastage of store. If the package becomes excessively long through future additions, or store is at a premium, it would be possible to reorganize the program such that a central section called down routines as needed. The complications which would be introduced by routines overwriting each other—Atlas Autocode programs are non-relocatable—make this method less attractive unless the package is quite large.

The package, SPP 1, consists of a main program and eight processing routines, and is normally stored at the beginning of the same magnetic tape which holds the digitized speech. The central program reads in the first parameter, and selects a particular routine which reads in further parameters. Several tasks may be initiated by one steering tape, termination being indicated by a routine selection parameter of -1. The individual routines, their parameters, and typical applications, are described below.

3.2 Routines of the Speech Processing Package, SPP 1

Routine r1—Minispec. This routine gives a direct print-out of the digitized speech samples stored on magnetic tape between two specified addresses. The samples, as signed decimal integers in the range -128 to $+127$, are printed 20 to a line-printer line. The program also notes the first-occurring sample to have an amplitude in excess of a pre-set threshold. Two further optional facilities are provided. (*a*) For a specified length of record subsequent to the threshold-crossing address, counts are performed of t, the number of turn-arounds or positions of zero time-derivative; z, the number of zero-crossings; and g, the difference $(t - z)$. These three numbers are referred to as the basic "statistics" of the record. (*b*) An amplitude/time graph is printed of the speech waveform over the interval from 110 samples prior to the threshold-crossing to 220 samples subsequent to this crossing—i.e. a total interval of 16·5 msec for a sampling rate of 20 kc/s. To economize on line-printer output, the waveform is clipped at ± 54 amplitude units, and the curve is plotted as two 55-sample sections per line-printer page. Thus, each page of output should be cut in two and joined end-to-end. Axes are printed and labelled to facilitate this joining.

The program requires four basic parameters, and then six additional parameters for each section to be inspected, as follows:

P1 = sampling rate (20,000 or 10,000).
P2 = amplitude-threshold setting (typically about 15 units in order to allow for moderate background noise).
P3 = length of record subsequent to the threshold crossing over which statistics are to be computed. (May be 0 to 220 samples; 180 has been found useful at 20 kc/s sampling, since this corresponds to an average voiced pitch-period.)
P4 = number of pairs of addresses for examination.

For each section to be examined, the following numbers are required:

Pa, Pb = start at sample Pa in block Pb.
Pc, Pd = end at sample Pc in block Pd.

(The section thus specified may be up to two tape-blocks long (6,143 samples).)

Pe = 1 if statistics are desired; 0, otherwise.
Pf = 1 if a graph is desired; 0, otherwise.

Depending on the facilities requested, the computing requirements are approximately:

time: $\leqslant 1{,}200 \times$ P4 I.C.I.'s, depending on facilities requested, and length of record printed. Note: one I.C.I. (Instruction-counter Interrupt) is equal to 2,048 actual obeyed instructions; the average time for an Atlas instruction is approximately equal to 3 microseconds.
output: 154 lines per block of speech +396 lines per graph.

This routine has been found useful in determining the exact shape of a waveform—for example, in order to check the periodicity—and in examining the initial burst of sound in plosives, etc.

Routine r2—Amplitude bar graph. This routine produces a bar graph of the peak absolute value of the speech envelope as illustrated in **Fig. 1.** Each point on

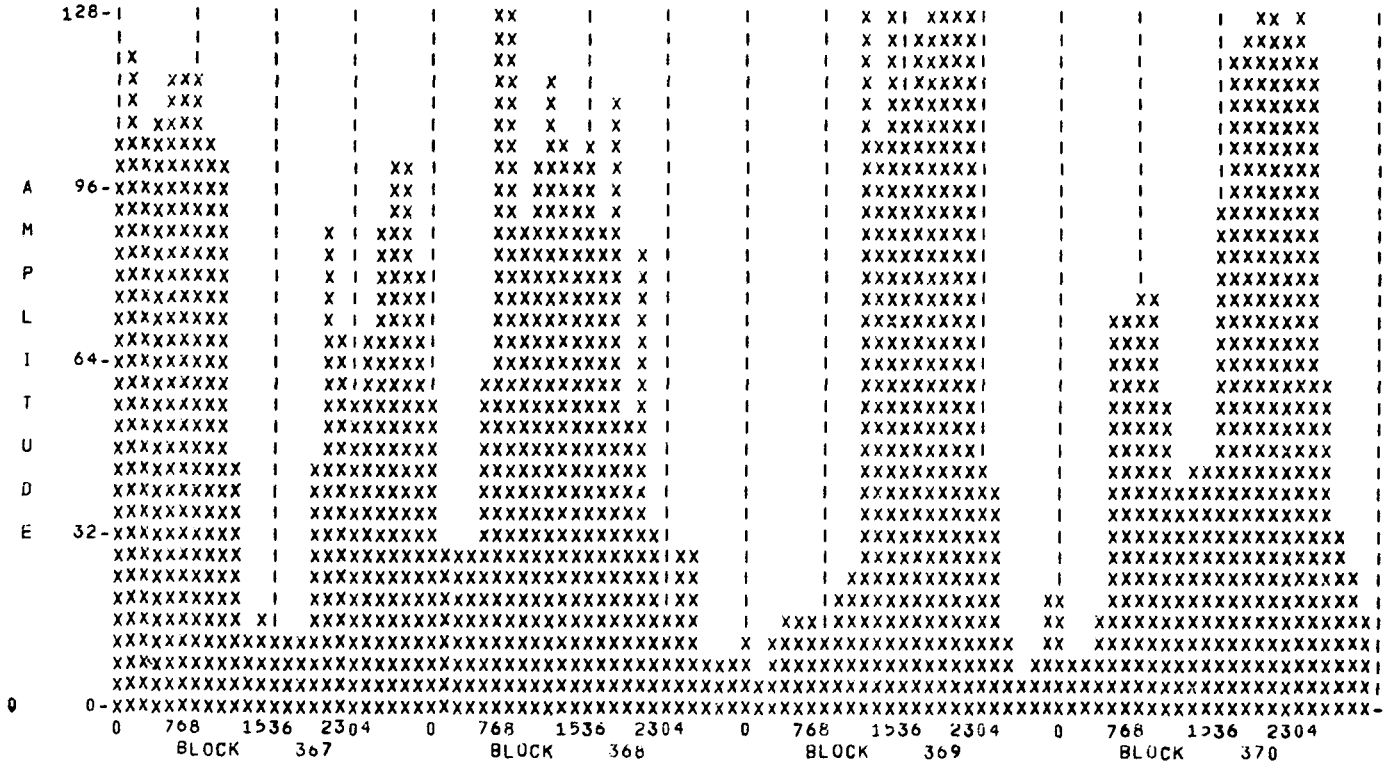

Fig. 1.—Amplitude bar graph of connected speech stored on magnetic tape

the graph represents the maximum absolute value of amplitude over a period of 128 samples of speech.

Two input parameters are required:

P1 = the number of blocks to be graphed.
P2 = the block number of the first block to be graphed.

The computing requirements are approximately

time: 150×P1 I.C.I.'s
output: 16×P1 lines.

This program has been found useful in the location of speech sounds on the magnetic tape containing the digitized speech prior to detailed analysis, and in the production of an easily used visual record of the speech.

Routine r3—General Statistics. This routine performs a small number of statistical measurements on each block of stored speech, sufficient to determine the general nature of the record contained in that block. When the sampling rate is 20 kc/s, five stations starting at address 0 and separated by 600 samples are chosen. At each station the following quantities are determined and printed out: (*a*) the maximum positive amplitude occurring in the 300 samples (15 msec) subsequent to each station; (*b*) the address of this maximum; (*c*) the number of turn-arounds, t, the zero-crossings, z, and the quantity $(t - z)$, occurring in the next 180 samples from each station. For a sampling rate of 10 kc/s, 10 stations are chosen spaced at 300-sample intervals, and the distances over which the statistics are calculated are halved.

In addition, an approximate assessment of the quality of the record is made at each station according to the following rules:

if $t > x$, the record is classed as *high frequency*
if $y < t \leqslant x$ the record is classed as *turbulent*
if $t \leqslant y$, and max. ampl. > 26 the record is classed as *voiced*
if $t \leqslant y$, and max. ampl. $\leqslant 26$ the record is classed as *noise*.

Experience has shown that x and y should have pre-set values of:

$x = 73$; $y = 61$, for 20 kc/s sampling
$x = 47$; $y = 38$, for 10 kc/s sampling.

These rules are not rigorous, but hold for most cases. Finer identification is possible by visual inspection of the statistics.

The program requires two basic parameters, and two further parameters for each section to be inspected, as follows:

P1 = sampling rate.
P2 = number of pairs of blocks.

For each pair, specify:

Pa, Pb = start at block Pa and proceed through to block Pb.

The computing requirements are:

time = 16 + 24(3 — P1/10000) I.C.I.'s per block
output = 5 + 6(3 — P1/10000) lines per block.

This routine has been found useful as a quick guide to utterance-identification. It is also helpful for locating the true beginning of certain initial consonants such as "F", where amplitude information alone is insufficient to determine the commencement.

Routine r4—Micro statistics. This routine computes the same turn-around and zero-crossing statistics of routine r3, but for consecutive sections throughout a block of speech. In addition, adjacent pairs of values for each statistic are averaged to produce a smoother version. As before, the length of record over which each set of statistics is taken is 180 samples for a 20 kc/s sampling rate, giving 16 consecutive sets of smoothed statistics per block. A similar set of parameters to that used in routine r3 must be specified.

The computing requirements are:

time: 16 + 51(3 — P1/10000) I.C.I.'s per block
output: 9(3 — P1/10000) lines per block (plus a six-line spacing between the sets of tables for each pair of block-addresses).

This routine has been found useful for rapid segmentation of speech—e.g. for locating transition sections and any quasi steady-state sections of a diphthong.

Routine r5—Distribution of zero-crossing intervals. This routine notes the interval, measured in samples, between successive zero-crossings of the speech waveform. For a specified length of record the program prints out a table of numbers of occurrences of each interval, covering intervals in the range 1 to 64 (or over) samples (50 μsec to 3·2 msec with the 20 kc/s sampling rate). Each block of speech is divided into consecutive sections of duration equal to the specified length of record, so that in general several tables will be printed out per block. It has been found difficult to assimilate this tabular information and so, in addition, an optional visual representation has been provided. The range of zero-crossing intervals is divided into three sections of interest, as directed by the user, and a histogram is printed out of total number of occurrences in each division. This may be compared with a vowel recognition system under development at Standard Telecommunications Laboratories (Bezdel and Chandler, 1965), which uses zero-crossing data divided into between 6 and 16 sections of interest, or "channels".

The program requires 5 basic parameters, and 4 further parameters for each section to be analyzed, as follows:

P1 = sampling rate.
P2 = upper limit of first division of interval-table.
P3 = upper limit of second division of interval-table.
P4 = upper limit of third division of interval-table.
P5 = number of pairs of block-addresses for processing.

For each pair, specify:

Pa, Pb = start at block Pa, and proceed through to block Pb.
Pc = length of record ($\leqslant$3071 samples) over which interval-distribution is to be computed.
Pd = 1 if a histogram is required; 0 otherwise.

The computing requirements are:

time: $\leqslant$250 I.C.I.'s per block
output: 12 + 10×Pd lines per length of record.

This routine has been found useful as an aid to word-segmentation.

Routine r6—Fundamental frequency detection. This routine detects the fundamental frequency in speech waveforms using a combination of autocorrelation analysis and amplitude peak measurements. Analysis may be performed independently on consecutive, overlapping portions of speech, or a track of the fundamental frequency made and smoothing used to eliminate spurious points.

The method used is as follows:

1. The largest amplitude peak in the section of speech to be considered is found. If this peak is less than 20% of the full encoding magnitude, the waveform is classed as a small signal, and no analysis is performed.
2. A variation of the autocorrelation function (the absolute value of the difference between pairs of samples, rather than the product, is used in order to increase the speed) is calculated, and troughs in this function in the range from periods of 1/60 to 1/300 of a second are located.
3. Each of the troughs in the autocorrelation function is taken as a trial fundamental period measurement, and the speech waveform is examined to check that an amplitude peak lies at this distance from the largest amplitude peak, $\pm$5%. If the amplitude peaks differ by more than 28% the trial period is discarded.
4. The autocorrelation function value at each of the trial period points is weighted by the percentage difference between the amplitude peaks in the speech waveform.
5. The weighted autocorrelation troughs are examined, and if the smallest value is less than 20% smaller than the second smallest, the data is classed as unreliable. This result is usually obtained in unvoiced waveforms or waveforms where the fundamental period is somewhat ambiguously defined.
6. The smallest weighted autocorrelation point is

taken as indicating the fundamental period. A check is made for autocorrelation troughs at twice this frequency, and if such a point is within 25% in magnitude, this value is used. This check has been found necessary in some waveforms where every other voicing cycle is slightly larger than the preceding one.

7. If tracking is desired, the values of the fundamental are stored, and the track inspected for discontinuities greater than 25%. If the track returns to its original value $\pm 5\%$ the spurious point is replaced by an interpolated value. If the data following or preceding the point is unreliable, the point is eliminated.

The analysis is carried out on consecutive sections of 1/30 second of speech, overlapping by 50%. Input parameters required are as follows:

P1 = sampling frequency of the speech.
P2 = control parameter for tracking. A −1 indicates that no tracking is desired.
P3 = sample number of the starting point on magnetic tape.
P4 = block number of the starting point.
P5 = number of sections to be analyzed.

The computing requirements are approximately:

time: 527 × P5 I.C.I.'s
output: P5 lines.

This program has been used as part of a system of speech analysis in connection with investigations of synthetic speech, and in performing pitch synchronous analysis (see routine r7 and Section 4.1). The reliability of the routine has been tested on 62 sections of 31 words spoken by 5 female and 11 male speakers, without tracking. The results differed from those obtained by visual examination of the waveform in the following cases:

(*a*) Two cases where the waveform contained many peaks which were clipped, were called unreliable by the program.
(*b*) One case where the waveform was falling off in magnitude was called unreliable by the program.
(*c*) Two cases occurred where the program gave one-half the frequency indicated by examination of the waveform.
(*d*) One case occurred where the program gave a frequency differing by 18% from that derived by inspection of the waveform.

In all other cases the program gave answers consistent with those obtained by hand. When tracking is used, the type of error obtained above in (*c*) and (*d*) may be eliminated. In working with connected speech as described in Section 4.1, the program produced no failures after tracking and interpolation were performed.

Routine r7—Spectral analysis by Fourier analysis. This routine takes frequency spectral cross sections of portions of the speech waveform by forming the Fourier transform of the autocorrelation function. This process

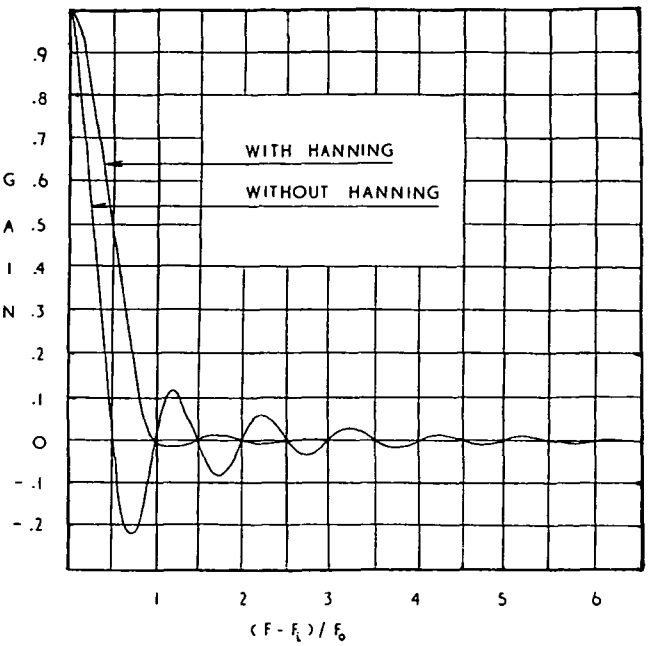

Fig. 2.—Bandpass characteristic of filters in routine r7 where F_i is the centre frequency of the filter, and F_0 is the fundamental frequency

effectively passes speech consisting of repetitions of the record to be analyzed through a series of filters whose bandpass characteristic is shown in **Fig. 2.** It is perhaps worthwhile to note the sources of error inherent in spectral analysis in general, and those arising from the method used, (Blackman and Tukey, 1958).

General sources of error:

1. The uncertainty involved in analyzing a finite portion of a semi-periodic waveform results in a decrease in reliability of amplitude measurements as the bandwidth of the estimates decreases.

2. While it is possible, by overlapping record portions analyzed, to take cross sections of records separated by as little as one encoding sample period, such a procedure does not truly produce additional data. Overlapping may be used to obtain a smoother picture of the changing spectrum, but the basic limitations of the information present prevents the additional measurements from representing independent data.

3. Since the waveform is in a sampled form, the filtering effect of the sampling process, and possible distortion due to aliasing must be borne in mind.

4. A small D.C. offset introduced by the encoding process will introduce a large percentage error in the low frequency estimates. For this reason, the D.C. component of the record to be analyzed must be removed before analysis, as is done in this routine.

The method used has the following limitations:

1. The basic shape of the filters in the process results in side lobes of approximately 20%, with corresponding distortion of the spectrum. Although it is possible to

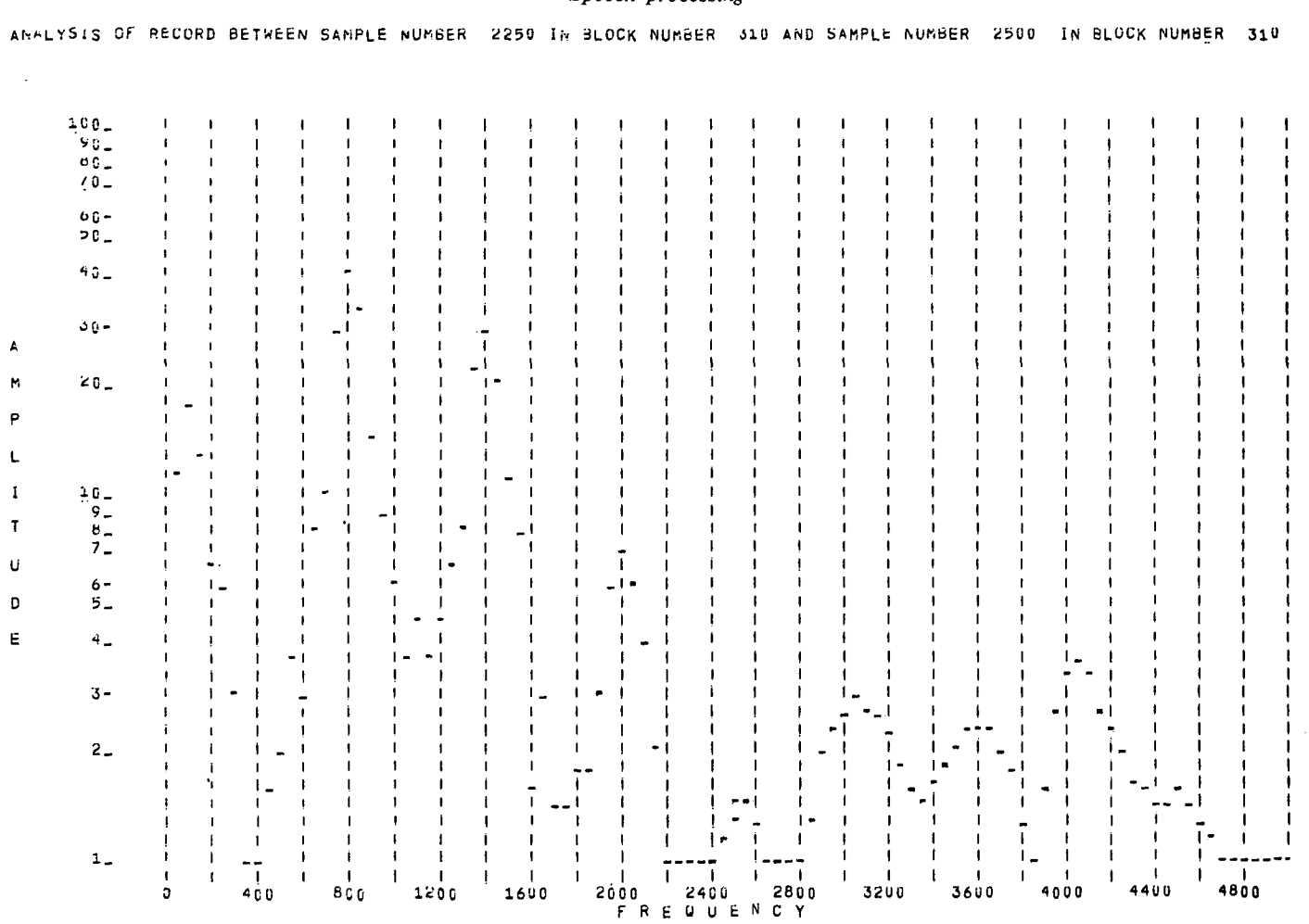

Fig. 3.—Spectral cross section output of routine r7

request Hanning of the filter outputs (see Blackman and Tukey, 1958) in this routine, with resulting improved filter shape (Fig. 2), this process assumes a flat spectrum, and when used on a voiced waveform results in smoothing out of the voicing harmonics. This smoothing has been found advantageous, however, in displaying formant structure.

2. Because the filters are not rectangular, error may be introduced if voicing harmonics fall between filter centres. Pitch synchronous analysis, i.e. the placing of filters centred on each harmonic spike, may be requested in this routine, which uses the method described in routine r6 to find the fundamental frequency, but it should be noted that a very small error in measurement of the fundamental frequency will result in an accumulative slipping of filters away from the voicing harmonics, with corresponding errors introduced in the upper frequency estimates.

It is important to recognize such sources of error, since the general form of the routine may allow unreliable and redundant data to be produced unless care is taken in the specification of the type of analysis desired.

The spectrum calculated by the routine may be displayed on a log-log-lin graph **(Fig. 3)**, and/or the values may be listed in tabular form. The input parameters to the routine are as follows:

P1 = sampling frequency.
P2 = the number of consecutive sections of record to be analyzed.
P3 = sample number of the starting point of the analysis.
P4 = block number of the starting point.
P5 = number of samples per section.
P6 = number of samples overlapped.
P7 = control parameter for tabular results. P7 = 1 indicates that tables are desired.
P8 = control parameter for log-log-lin graph. P8 = 1 indicates that graphs are desired.
P9 = control parameter for pitch synchronous analysis. P9 = 1 indicates such analysis is desired. If automatic determination of the fundamental is not required, P9 should be set equal to the desired fundamental.
P10 = control parameter for Hanning. P10 = 1 indicates that Hanning is desired.

The computing requirements of the routine are approximately:

time: 390×P2 I.C.I.'s.
output: 60×P2 lines.

This routine has been used in a number of investigations of speech properties. Section 4.1 describes one such application.

Routine r8—Spectral analysis by continuous filtering. This routine passes the speech waveform continuously through a set of filters formed from 2nd-order difference equations. The general equation and its band-pass characteristic is given in **Fig. 4**. The routine differs from r7 in the following ways:

1. The waveform is continuously passed through filters, rather than discrete portions of record being analyzed. It is thus possible to read the filter outputs and produce cross sections at frequent intervals without greatly affecting the computing time, except for time consumed in actually printing results. In effect, advantage is being taken of the computations already performed in previous cross sections.
2. The routine is faster than r7 because no cosine calculations are necessary once the filter constants have been calculated.
3. Because variation in the frequencies of the filters during filtering would necessitate frequent re-calculation of the filter constants, and hence a dramatic increase in computing time, pitch synchronous analysis is not provided with this routine.
4. A graphical output with a three-dimensional effect is obtainable by integrating the outputs of the filters, printing an asterisk when the integral reaches a specified threshold, and resetting the integral to zero when such a print-out is made. While the routine will produce, from waveforms of steady spectral distribution, a graph where amplitude is directly proportional to the number of asterisks per unit of time, the variation in the spectral distribution of speech waveforms with time produces deviations which make accurate amplitude measurements from this graph impossible, unless a prohibitively large time scale is used. The graph is intended to facilitate the general analysis of speech by providing a visual display similar to that obtained from standard spectral hardware; where exact data are desired, the cross sections from this routine or r7 should be used.

The input parameters for this routine are as follows:

P1 = sampling frequency.
P2 = distance between cross-sections.
P3 = sample number of starting point.
P4 = block number of starting point.
P5 = sample number of end of analysis.
P6 = block number of end of analysis.
P7 = bandwidth of filters.
P8 = frequency separation of filters.

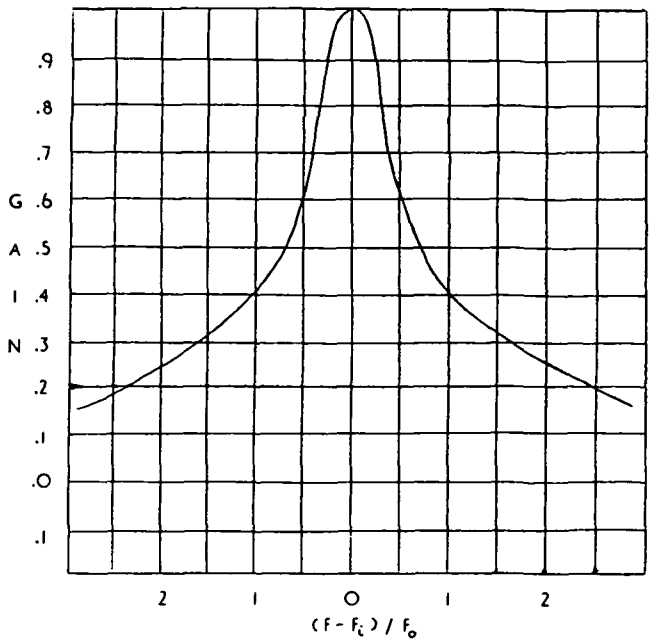

Fig. 4.—Bandpass characteristic of the filter equation used in routine r8, $X_t = 2e^{-bT}\cos(2\,FT)X_{t-T} - e^{-2bT}X_{t-2T} + I_{t-T}$ where X_t is the output of the filter at time t, T is the distance between samples, I_t is the input at time t, and b is the bandwidth, in this case set equal to the fundamental frequency as in Fig. 2 for the sake of comparison

P9 = frequency of lowest filter.
P10 = control parameter for suppression of low frequency range. P10 = 1 indicates that suppression is desired.
P11 = control parameter for tables. P11 = 1 indicates that tables are desired.
P12 = control parameter for time-frequency-amplitude graph. P12 = 1 indicates that the graph is desired.
P13 = number of samples between points on the time-frequency-amplitude graph.
P14 = density scale factor on the time-frequency-amplitude graph. P14 = 128 gives an asterisk every unit of time for a filter output of 128.
P15 = control parameter for log-log-lin graph of cross-sections. P15 = 1 indicates that the graph is required.

The computing requirements are approximately:

time: 50× (number of filters) I.C.I.'s per block of speech
output: 66× (number of cross-sections) + (length of record/P13) lines.

This routine has been found useful in obtaining an overall picture of the changes in the speech spectrum as time passes. It has also been used as a source of input to a system of simulated neural nets in preliminary investigations of the use of such nets in speech recognition.

4. Examples of the use of SPP 1

This section will describe two projects which make use of some of the routines in SPP1. These descriptions are not intended as definitive reports of final results, but rather as indicative of some areas in which the routines have been applied at Manchester.

4.1 Speech analysis and synthesis

Routines r6 and r7 were developed for use in a system of automatic speech analysis and synthesis (Rosenthal, 1965). This system was based on the assumption that the intelligible information in speech may be expressed as values of a few slowly varying parameters, and that these parameters may be measurements of characteristics in the spectrum of the speech waveform. The object of the exercise was to obtain a reduction in the data rate of speech transmission without loss of information. It was desired to produce analysis programs which would form the spectrum of the speech waveform and measure the values of several parameters which would then be used as input to a synthesis program. The synthesis program would produce a waveform whose spectral characteristics were controlled by the parameters, and which would, hopefully, be judged similar to the original by human listeners.

A speech synthesis program was written at Manchester (Mathers, 1964) based on the Parametric Artificial Talker at Edinburgh University (Anthony and Lawrence, 1962). The eight parameters required by this program were measures of the frequency of relatively dense energy bands in the spectrum, energy amplitude measurements, and measurement of the fundamental frequency of the waveform. Two other synthesis programs were developed which required nine and eleven parameters, representing a closer description of the spectrum. Three analysis programs were written which automatically measured the spectrum and produced the appropriate parameters from the speech waveform.

Because of the complete reliance of these programs on the frequency spectrum of the speech, it was essential that an accurate spectral routine be used, and r7 was developed for this purpose. By making the spectral routine general and flexible it was possible to vary the manner and detail in which the spectrum was calculated, and thus determine the most appropriate technique for the problem. Because of the primarily harmonic nature of the voiced spectrum, it was found that pitch synchronous analysis using Hanning produced the most satisfactory results. Occasional distortion of the spectrum, due to peculiarities in the waveform caused by rapid fluctuations in articulation, emphasized the tendency of the analysis program to produce spurious results intermittently, and overlapping spectral sections were therefore taken with smoothing introduced to discard such unreliable data. In general it was found that a great deal of care was needed in specifying the spectral analysis of speech, and in interpreting results, if valid and reliable data were to be obtained (Rosenthal, 1965).

Routine r6 was used to provide a measure of fundamental frequency required by the synthesis program and by the pitch synchronous operation of routine r7.

The three speech analysis-synthesis programs were tested on two speech samples of approximately 3 seconds, each spoken by two male speakers. Computing requirements were 100 seconds per second of speech and 60 blocks of store for the analysis programs, and 8 seconds per second of speech and 32 blocks of store for the synthesis programs. The speech produced by the three systems improved as the number of parameters specifying the spectrum was increased. Some distortion was present even in the output of the eleven-parameter system, caused primarily by rapid shifting of the spectrum in certain voiced-unvoiced transitions, but the overall effect was intelligible and of reasonably good quality. It is probable that further improvement will depend on adjustment to the smoothing procedures rather than an increase in the number of parameters.

4.2 Verbal command recognition system

The procedures that form the basis of Routines 3 and 5 have been employed in conjunction with amplitude-envelope information to develop a set of fast and simple measurements on the speech waveform. These are utilized for speech recognition as follows. Amplitude data is used for approximate utterance-location and for specific tests concerned with initial plosives and nasal consonants. Discrete sections of a word are chosen for detailed inspection from amplitude data and/or by continuously tracking one statistical count. For each section of interest, often corresponding approximately to a formal phoneme, the statistical counts provide a measurement-space of up to five dimensions, although only two dimensions suffice for much classification of sounds. The unvoiced consonants so far tested have tended to form clusters in this space. Vowel-sounds have tended to form a continuum that can be approximately identified with certain physical articulatory variables, such as tongue-movement. For limited command vocabularies for use with a wide range of talker-accents, this continuous vowel-scale may form a useful vector for recognition, when used in conjunction with other consonant information.

A simple 16-word vocabulary has been used to assess the effectiveness of these measurements. The vocabulary consists of monosyllables formed by combinations of one of four initial consonants (S, F, T, N), one of two vowel-sounds ("EE", or "OR"), and one of two final consonants (S, N). A recognition program has been tested on a total of 70 words spoken by 19 speakers (14 male, 5 female; various accents including Cockney, Lancashire, Glasgow, American, and Southern English). Average computing time on the University's Atlas was 1·2 sec per utterance, of which about 0·35 sec was spent in assembling the digitized speech from magnetic tape to the main store prior to analysis. (The speech was previously put onto tape via a tape-recorder or direct microphone input to the Speech Converter.) The main

program was written in Atlas Autocode, and is at present being converted to machine code. Preliminary tests have shown that this will reduce the overall time (including data-assembly) to below 0·7 sec per utterance. The size of the Atlas Autocode object-program is 3,000 machine-instructions.

Of the 70 words tested, one serious misclassification occurred when a "Sorn" was recognized as "Seen". This was due mainly to the simplified method of choosing which finite section of the total "vowel" area should be analyzed. It was seen to be necessary to include some assessment of the "steadyness" of a sound in any future program. When the failed word was subsequently repeated by the same speaker, recognition was correctly achieved. Three other minor errors occurred. In two cases, the Speech Converter was mis-adjusted and an appreciable D.C. shift, coupled with the low overall intensity of a particular speaker, caused a significant reduction in the number of zero-crossings in the case of two initial unvoiced consonants. A later program which includes a D.C. offset test, has successfully recognized these two words. The other "fault" was due to an initial "F" being so poorly articulated that it was indistinguishable from background noise when carefully listened to by subsequent observers.

The first section of the above simple program has been enlarged to accept all the initial consonants that are present in a vocabulary of the digits 0 to 9 (i.e. N, W, E, F, Th, S, T). Results from a test run on 32 voiced and 79 unvoiced consonants spoken by the above group of talkers has produced the following figures. Voiced consonants: all examples correctly classified; average computing time for each consonant is 0·95 sec. Unvoiced consonants: about 62% of all "F"s and 67% of all "Th"s had to be assigned to a general class as "Th/F" sound; all other examples were correctly distinguished; average computing time for each consonant was 1·6 sec. As before, these times include 0·35 sec spent in data-assembly, and are capable of improvement when the programs are machine-coded. The present object-program of the initial consonant section occupies 7,300 machine instructions.

It is possible that the poor distinction between "Th" and "F" indicates a weakness in the measurement system. However, the nature of the clustering obtained for these sounds, and the results of human listener tests support the theory that the basic information-content of these consonants is in fact inadequate for rigorous distinction when considered without the help of context. The same remarks apply concerning distinction between a rather lispy "S" and a "Th"—though classification has been accomplished for all examples so far processed. These problems will be resolved to a large extent by context, in a system used to recognize the total word in the environment of a finite command system format.

5. Conclusion

The hardware and software systems described above have been used at Manchester in research into speech processing by computer. The general-purpose nature of these facilities has made possible their use by other institutions, e.g. communication simulation by Queen Mary College, London University, and investigations of vowel sounds by Standard Telecommunications Laboratories. These projects provide an illustration of the scope and flexibility which may be obtained by a number of research groups from a large computing installation without the necessity for excessive individual investment in hardware and software development.

6. Acknowledgements

Thanks are due to Professor T. Kilburn for his encouragement and provision of the facilities with which this work has been undertaken. The authors also wish to acknowledge the role of Dr. R. W. Mathers, formerly of the Department of Electrical Engineering, in the design of the Speech Converter, and the helpful advice of Dr. F. H. Sumner of the Department of Computer Science.

References

ANTHONY, J., and LAWRENCE, W. (1962). "A resonance analogue speech synthesiser," Fourth International Congress on Acoustics, Copenhagen.

BEZDEL, W., and CHANDLER, H. J. (1965). "Results of an analysis and recognition of vowels by computer using zero-crossing data," *Proc. I.E.E.*, Vol. 112, No. 11, p. 2060.

BLACKMAN, R. B., and TUKEY, J. W. (1958). *The measurement of power spectra*, New York, Dover Publications.

BROOKER, R. A., and ROHL, J. S. (1966). "Further literature on Compilers AA and AB." (Available from the Department of Computer Science, Manchester University.)

KILBURN, T., EDWARDS, D. G. B., LANIGAN, M. J., and SUMNER, F. H. (1962). "One-level storage system," *I.R.E. Trans. on Electronic Computers*, Vol. EC-11, No. 2, p. 223.

LAVINGTON, S. H. (1964). "Some aspects of speech synthesis by computer," M.Sc. Thesis, University of Manchester.

MATHERS, R. W. (1964). "Application of digital computers to road traffic control and to speech synthesis," Ph.D. Thesis, University of Manchester.

ROSENTHAL, L. E. (1965). "Computer analysis and synthesis of speech," M.Sc. Thesis, University of Manchester.

SHANNON, C. E. (1949). "Communication in the presence of noise," *Proc. I.R.E.*, Vol. 37, p. 19.