Speech Recognition Using Autocorrelation Analysis

RAYMOND F. PURTON

Abstract—**Experiments are described in which word recognition is based on digital autocorrelation analysis followed by computer pattern matching. Incoming speech is split into two frequency bands, and the signals in each band are quantized into two amplitude levels. The two signals are fed to separate autocorrelators, consisting of binary shift registers, digital multipliers, and RC integrators. The low- and high-frequency correlators have, respectively, 10 and 8 outputs which are coded into a 36-bit character, sampled 40 times per second, and fed to a digital computer for recognition. In the computer, master patterns in the form of a 36×30 matrix, are generated for each word of the vocabulary from a number of known utterances of the word. Unknown utterances are then compared with each master pattern in turn, and the best match determined by a simple scoring technique; if desired, master patterns can be "updated" when correct recognition occurs. Master patterns can be formed from either one or several speakers; when formed from a single speaker, and with a vocabulary of 10 words, subsequent utterances by the same speaker are recognized with an average accuracy of 90 percent.**

Manuscript received September 25, 1967. This paper was presented at the 1967 Conference on Speech Communication and Processing, Cambridge, Mass.

The author is with the Plessey Electronics Group, British Telecommunications Research Ltd., Taplow Court, Maidenhead, Berks., England.

Introduction

NUMEROUS experimental automatic speech recognition systems have been described in the literature, but nearly all of these are based on some form of frequency analysis of the speech waveform. The alternative of autocorrelation analysis seems to have been largely neglected as a tool for speech recognition, although autocorrelation vocoders have been described.[1],[2]

It is not claimed that autocorrelation is necessarily any more effective than frequency analysis, and, in fact, for a continuous stationary signal and ideal analyzers, the same information is available from both methods, since the autocorrelation function is the Fourier transform of the power spectral density. For practical analyzers having finite integration times and operating on varying signals, such as speech, the simple Fourier relationship no longer holds, but it is clear that the two methods are still producing basically similar information.[3],[4]

The attractiveness of autocorrelation analysis lies in the fact that it lends itself readily to instrumentation by digital circuits. It is thus compatible with modern microminiaturization techniques, and can be easily integrated with computing and data processing systems. Experiments have, therefore, been carried out using a digital autocorrelation analyzer feeding into a digital computer, with the computer programmed to carry out word recognition by a pattern matching process.

The aims of the work in terms of size of vocabulary and number of speakers are fairly modest, but a word recognition device of even limited capability could have numerous applications. Most of the experiments to date have been limited to a vocabulary of 10 words spoken by single speakers, but it is hoped to extend this to perhaps 50 words and a small group of operators.

The Autocorrelation Analyzer

It is known that infinitely clipped speech, that is, speech signals for which polarity information only is retained, is still quite intelligible.[5] It was decided, therefore, to carry out a single bit (polarity only) coding of the speech signal prior to the autocorrelation process. This simplifies the equipment required to perform the analysis, and also provides a useful preliminary stage of data reduction. However, in order not to discard too much information at this stage, and in particular, to preserve the smaller amplitude higher frequency components of the speech waveform, the speech is first split into two frequency bands by a pair of low- and high-pass filters with a crossover frequency at 1000 Hz; this roughly separates the first formant from higher formants. The signals in each band are then digitized and fed to two separate autocorrelators.

The correlators are required to perform an approximation to the integral

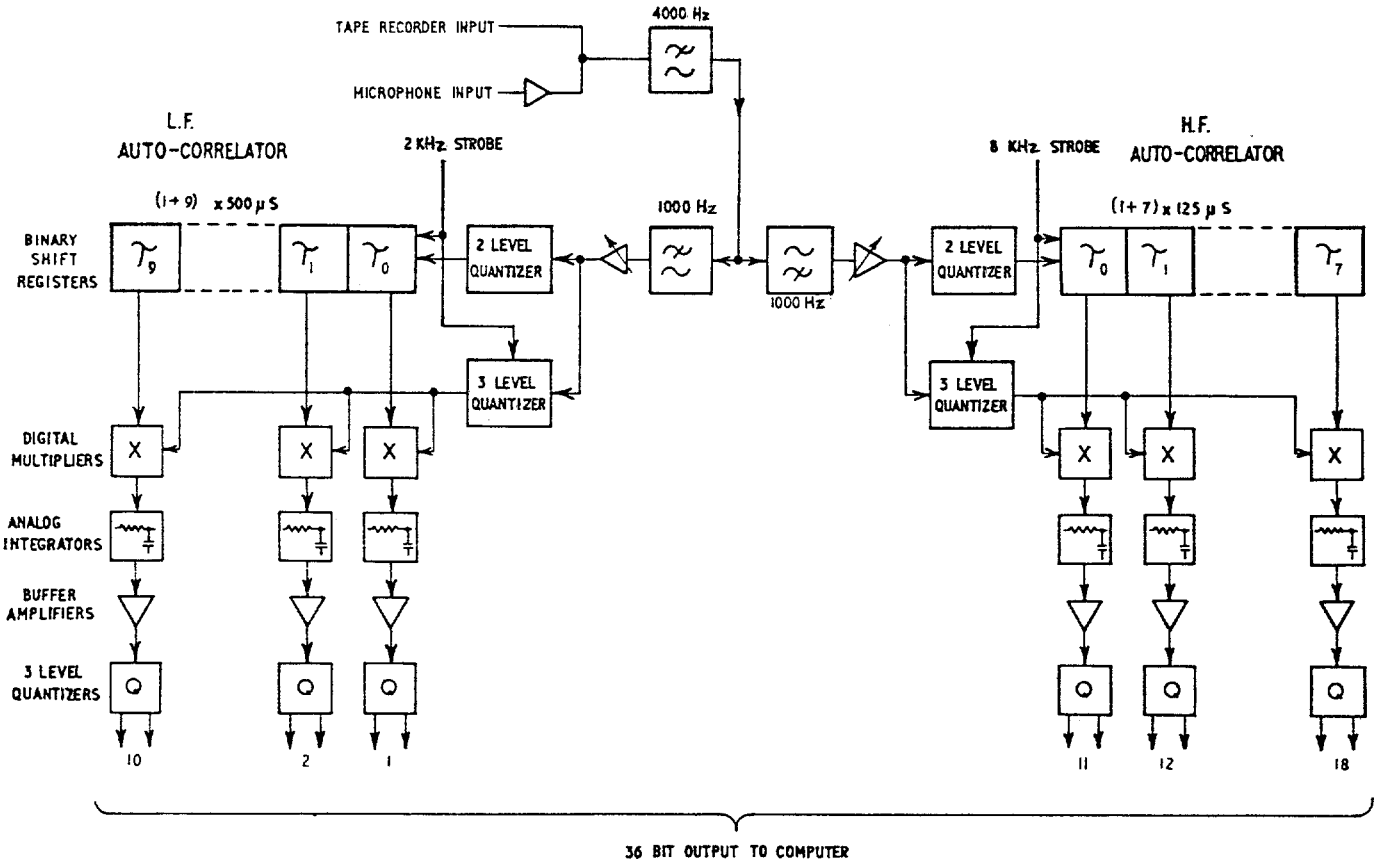

Fig. 1. Block diagram of autocorrelation analyzer.

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} f(t) \cdot f(t-\tau) \cdot dt$$

for a number of values of the delay parameter τ. This is achieved by means of a tapped delay line, with a multiplier and integrator connected to each tap. Fig. 1 shows a block diagram of the arrangement.

Each delay line consists of a digital shift register, which is strobed at a rate appropriate to the frequency band being analyzed. These strobe rates are chosen so that the delay per stage is equal to a half-cycle of the highest frequency of the band; that is, strobing or sampling is at the Nyquist rate. Thus, the low-frequency register is strobed at 2 kHz (delay interval 0.5 ms) and the high-frequency register at 8 kHz (delay interval 0.125 ms). The lengths of the delay lines are determined by the lowest frequency of each band, and have been chosen so that approximately one full cycle of the lowest frequency can be accommodated. The low-frequency register has 10 sections, including one of zero delay, and the high-frequency register has 8 sections. The overall bandwidth of the system is thus 200 to 4000 Hz, corresponding roughly to telephone quality speech.

As mentioned previously, the speech signals are digitized by two-level quantizers before being applied to the delay lines. In addition, undelayed signals are required for applying to the multipliers; these are also digitized, but into three levels, so that low-level noise signals are eliminated by producing zero correlation.

The multipliers are three-level logical gates, the three levels corresponding to instantaneous positive, zero, and negative correlation. The outputs of the multipliers feed RC integrators with a nominal time constant of 22 ms, and after amplification, the slowly varying integrator outputs are again quantized into three levels and coded into two-digit binary signals, in which 01, 00, and 10 represent positive, zero, and negative correlation, respectively, the code 11 being prohibited.

The output of the complete analyzer is thus a parallel 36-bit signal. This signal is fed to a small general-purpose digital computer via an interface unit, and sampled every 25 ms. Sampling is initiated when either shift register is energized for a minimum period of time, and continues for a predetermined number of samples (usually 30) that is, 750 ms. The resulting 36×30 matrix is stored in the computer and can be printed out, either as a 36×30 binary pattern, or an 18×30 ternary (+, −) pattern. An example of the latter for the word "nought" is shown in Fig. 2; in this representation of the pattern, + indicates positive correlation, − indicates negative correlation, and blank spaces represent zero correlation (including zero signal or silence).

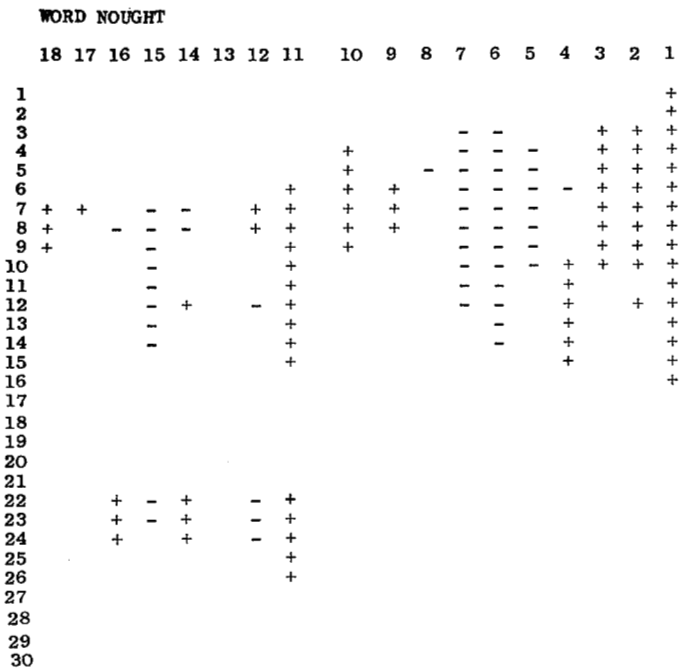

Fig. 2. Ternary printout of the word "nought."

MASTER PATTERN FOR WORD NOUGHT STRENGTH 5
MAXIMUM SCORE 2239

	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
1	00	00	00	00	00	00	00	00	00	00	00	10	10	10	00	02	03	05
2	00	00	00	00	00	00	00	00	01	00	00	20	20	20	00	02	03	05
3	00	00	00	00	00	00	00	00	02	00	00	30	30	20	00	04	04	05
4	00	00	00	00	00	00	00	00	03	01	10	40	40	30	00	04	04	05
5	00	00	00	00	00	00	00	00	04	01	20	50	50	50	00	05	05	05
6	01	00	00	00	00	10	00	01	04	02	20	50	50	50	20	04	05	05
7	01	01	00	10	10	00	01	02	05	03	00	50	50	50	30	04	05	05
8	02	02	00	20	20	00	02	02	05	03	00	50	50	50	20	04	05	05
9	03	01	20	20	30	10	02	04	05	05	00	50	50	50	40	03	05	05
10	03	01	00	20	10	10	00	04	05	04	00	50	50	50	30	04	05	05
11	03	02	00	30	20	20	01	05	04	04	01	50	50	50	22	02	04	05
12	02	02	00	30	20	20	01	05	03	04	01	50	40	50	22	02	04	05
13	03	02	10	50	30	00	03	05	03	04	11	40	40	40	12	02	03	05
14	03	02	10	50	31	00	13	05	03	04	11	40	40	40	12	02	04	05
15	03	01	20	50	30	00	03	05	04	03	00	40	50	40	11	03	03	05
16	03	01	20	50	31	00	03	05	04	03	10	30	40	40	12	03	03	05
17	03	01	20	30	30	00	03	05	03	04	01	40	30	40	01	02	03	05
18	03	00	20	30	20	00	02	04	03	03	01	40	30	40	00	01	03	05
19	02	00	10	30	10	00	02	03	03	01	00	20	20	20	00	01	02	05
20	01	00	00	30	00	00	01	03	03	01	00	20	00	20	00	01	01	03
21	01	00	00	30	00	00	00	03	02	01	00	20	00	20	01	01	01	03
22	01	00	00	30	00	00	00	03	01	01	00	10	10	10	01	00	01	03
23	01	00	00	30	00	00	00	03	01	02	00	20	11	20	01	00	01	03
24	01	00	00	30	00	00	00	03	01	02	00	20	01	20	01	00	01	03
25	00	00	01	20	01	00	10	05	11	02	02	10	10	10	01	00	00	03
26	00	00	01	10	01	00	10	05	10	02	02	10	10	00	01	00	00	03
27	01	00	02	10	01	00	20	04	00	01	01	00	00	00	00	00	00	03
28	01	00	02	00	01	00	20	03	00	00	01	00	00	00	00	00	00	03
29	00	00	00	00	00	00	10	03	00	00	01	00	00	00	00	00	00	03
30	00	00	00	00	00	00	10	03	00	00	01	00	00	00	00	00	00	03

Fig. 3. Master pattern (strength five) formed from five utterances of the word "nought."

Computer Recognition Procedures

The basic recognition procedure consists of comparing the 18×30 binary-coded ternary pattern representing one spoken word with a number of stored reference or master patterns, and determining the best match. Recognition is thus based on complete words, rather than word segments such as phonemes.

Formation of Master Patterns

A master pattern is produced for each word of the chosen vocabulary by combining the individual binary patterns from several known utterances of the word. Usually the utterances are all made by the same speaker, but this is not essential. The combining process is a simple cell-by-cell addition, so that the number in any cell of the master pattern is equal to the number of individual patterns in the reference or training group which contained a 1 in the corresponding cell; the number is thus proportional to the probability of a 1 occurring in that position.

Fig. 3 shows a master pattern formed in this way from five utterances of the word "nought." In this representation, the double columns correspond to the 18 outputs of the two delay lines (1–10 for low-frequency signals, and 11–18 for high-frequency signals). In each double column, the right-hand decimal numbers represent probabilities of positive correlation, and left-hand numbers represent probabilities of negative correlation. By inference, the probability of zero correlation for any double cell is obtained by subtracting the sum of the positive and negative numbers from the master pattern strength, in this case 5. For example, output 4 at sample time 13 has an entry 12; this represents probabilities of $\frac{2}{5}$ for positive correlation, $\frac{1}{5}$ for negative correlation, and, by inference, $\frac{2}{5}$ for zero correlation.

The computer program can handle master patterns up to strength 63, but most of the work has been done on patterns of strength 9 or less.

Scoring Algorithm

The pattern matching procedure consists of comparing each double cell of an unknown pattern with the corresponding double cell of a master pattern, and assigning a score to the unknown pattern equal to the appropriate decimal number (or inferred number) in the master pattern. Thus, in the above example in which the master pattern has an entry 12, an entry of 01 (positive correlation) in the corresponding double cell of an unknown pattern would score 2, an entry of 10 (negative correlation) would score 1, and 00 (zero correlation) would score 2. Similarly, if the master pattern entry is 05, 01 in the unknown pattern would score 5, while 10 and 00 would both score zero.

This is done for each cell of the pattern and the results are summed, so that the whole process is a form of weighted cross-correlation. The total score is then normalized as

$$\frac{\text{Total score} - \text{Minimum score}}{\text{Maximum score} - \text{Minimum score}} \times 100 \text{ percent}$$

where "maximum score" is the score which would result if every double cell scored the highest of the three possi-

bilities, and "minimum score" is the score when every cell scores the lowest of the three.

Time Normalization

Utterances are normally sampled for 750 ms (30 samples), but many words are significantly shorter than this; for example, the utterance represented by Fig. 2 has an active length of 26 samples (650 ms). This means that the lower part of most patterns is blank and does not contribute useful information. To eliminate this blank part of the pattern and to compensate to some extent for varying rates of speaking, an optional facility is provided in the program whereby the active part of each pattern can be stretched or contracted to a predetermined length. This is done by repeating or eliminating samples at appropriate intervals.

Most of the experiments have been carried out with patterns stretched to 30 samples and, in fact, Fig. 3 is a master pattern formed from individual patterns stretched in this way. Unknown patterns would also be stretched to 30 samples before being compared with this master.

Updating

A further optional facility enables a master pattern to be modified or updated by the addition of a correctly recognized pattern. In this way, the master patterns can be adapted to slight changes in a speaker's voice. Alternatively, this procedure could be used to adapt the system to a new speaker, provided his patterns were not too dissimiliar to the existing masters.

Since updating increases the strength of master patterns, the facility includes a halving procedure, whereby at a predetermined strength all entries in the master pattern are halved in value, thereby halving the strength. In this way, the master pattern strengths are kept within predetermined limits, and all added patterns have approximately the same influence.

Updating only takes place if the highest score exceeds the second highest score by a predetermined margin; this margin is usually set at four units.

Results

Most of the tests have been done with 10-word vocabularies, usually the numerals "nought," "one," . . . "nine," but the actual words used are not critical provided they are not too alike. Speech is normally prerecorded on magnetic tape, although a direct microphone input is also available.

To form master patterns, the words of the vocabulary are spoken in sequence, and the complete sequence repeated a number of times, usually five. A master pattern of appropriate strength is then formed and stored in the computer for each word of the vocabulary. "Unknown" utterances, in random order, can now be fed in and recognized, and the results printed out. This printout may consist of a list of words recognized, or, alternatively, a full printout of scores for each pattern can be obtained; an example of the latter is shown in Fig. 4. In this case,

MASTER PATTERN	0	1	2	3	4	5	6	7	8	9	
0	90	82	78	73	85	74	75	69	79	70	0
1	75	88	76	75	81	75	62	73	64	76	1
2	76	84	89	86	82	77	65	73	71	77	2
3	69	76	74	85	75	73	65	71	66	76	3
4	74	78	73	76	82	73	69	81	69	71	4
5	80	86	78	81	84	87	72	73	72	81	5
6	74	73	73	74	75	67	91	72	70	64	6
7	66	66	64	67	68	61	71	81	62	64	7
8	77	75	74	72	77	68	75	67	93	66	8
9	63	75	57	58	67	67	54	65	58	69	1
WORD SPOKEN											WORD RECOGNIZED

Fig. 4. Typical scores for 10 different utterances scored against 10 master patterns, showing one misrecognition.

the numerals "nought," "one," etc. were spoken in sequential order, as shown in the left-hand column, with each row of the table corresponding to one utterance. The 10 central columns show the percentage scores for each utterance against each master pattern, and the right-hand column indicates the words recognized, i.e., the highest score in each row.

In the following results, all patterns were time normalized to 30 samples, and master patterns were of strength five when updating was not used. When updating was used, initial master patterns were of strength three and halving took place at strength six.

Single Speaker Results

When master patterns were formed by one speaker and subsequent utterances of the same speaker scored against these masters, the average recognition accuracy without updating was 89 percent. This was an average from 696 utterances by three male speakers and one female speaker; individual accuracies varied from 78 percent to 99 percent, as shown in Fig. 5.

The experiments with updating of master patterns were done with data from different magnetic tapes and mostly different speakers and so are not directly comparable but, in this case, six speakers (five male and one female) averaged 93 percent on 385 utterances. Again, the details are shown in Fig. 5. Updating thus appears to give a small but useful improvement; however, it must be used with caution, since if incorrect updating occurs, master patterns can become badly distorted.

Multiple Speaker Results

If one speaker's utterances are matched against another speaker's master patterns, the recognition accuracy drops significantly. This was shown by a recognition run in which 150 utterances of each of two male speakers (JHW and PAF) were scored against the other speaker's masters; the average accuracy under these conditions was 59 percent. A third female speaker

SPEAKER	MALE/FEMALE	VOCABULARY	UPDATING OF MASTER PATTERN	RECOGNITION SCORE	RECOGNITION ACCURACY
J.H.W.	M	NOUGHT---NINE	NO	198/200	99%
J.H.W.	M	ALFA---JULIETT	NO	45/48	94%
P.A.F.	M	NOUGHT ---NINE	NO	182/199	91%
J.H.J.	M	ZERO---NINE	NO	116/149	78%
J.A.C.	F	NOUGHT---NINE	NO	79/100	79%
J.H.W.	M	NOUGHT---NINE	YES	153½/158	97%
A.J.G.	M	NOUGHT---NINE	YES	26/26	100%
A.I.H.	M	NOUGHT---NINE	YES	34/36	94%
R.A.J.	M	NOUGHT---NINE	YES	81/88	92%
C.M.H.	M	NOUGHT---NINE	YES	18½/23	80%
J.A.P.	F	NOUGHT---NINE	YES	45/54	83%

Fig. 5. Recognition results for single speakers and 10-word vocabularies.

(JAC) was then added, and all combinations of cross-matching carried out, but the overall average remained virtually unchanged, being now 57 percent.

The multiple speaker accuracy can, of course, be improved by producing composite master patterns from the utterances of several speakers. This has been done for the same three speakers by forming master patterns of strength six from two utterances of each word by each speaker. Average recognition accuracy was then back to 85 percent, compared with an average of 90 percent for these three speakers when scored against their own master patterns.

Conclusions

The results obtained have verified that the autocorrelation analyzer and associated recognition procedures are capable of forming the basis of a limited word recognition system. Significant differences exist in the master patterns of different speakers, but an important advantage of the master pattern technique lies in the ease with which master patterns can be formed for new speakers or new vocabularies.

The average recognition accuracy of about 90 percent with a 10-word vocabulary is, perhaps, rather lower than desirable for a practical system. However, a good speaker can attain a consistent accuracy of 98–99 percent, so results of this order should be achieved by trained operators.

At present, the two main disadvantages are the related ones of relatively long recognition time and large computer storage requirements. These are both proportional to the size of patterns and number of words in the vocabulary. With a 10-word vocabulary and 30 samples per utterance, recognition time is about 3 seconds per utterance, and the master patterns occupy 2700 24-bit words of computer store. These are both excessive for a practical system, but there is some evidence that satisfactory results can be obtained with considerably fewer effective samples per utterance; also, more efficient storage is possible. There is, therefore, hope that recognition in approximately real time (say 1 second per utterance) should be feasible with a vocabulary of at least 30 words.

Acknowledgment

The author wishes to thank British Telecommunications Research Ltd. for permission to publish this paper. He also thanks E. A. Newman of the National Physical Laboratory, Teddington, England, who first suggested the use of twin-channel autocorrelation analysis, and his colleagues J. R. W. Smith and J. H. Warren for very significant contributions to the work.

References

[1] M. R. Schroeder, "Correlation techniques for speech bandwidth compression," *J. Audio Engrg. Soc.*, vol. 10, p. 163, 1962.

[2] H. M. Christiansen *et al.*, "New correlation vocoder," *J. Acoust. Soc. Am.*, vol. 40, p. 614, September 1966.

[3] R. M. Fano, "Short-time autocorrelation and power spectra," *J. Acoust. Soc. Am.*, vol. 22, p. 546, September 1950.

[4] M. R. Schroeder and B. S. Atal, "Generalized short-time power spectra and autocorrelation functions," *J. Acoust. Soc. Am.*, vol. 34, p. 1679, November 1962.

[5] J. C. R. Licklider and I. Pollack, "Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech," *J. Acoust. Soc. Am.*, vol. 20, p. 42, January 1948.

Raymond F. Purton was born in Harrow, England, on May 21, 1923. He received the B.Sc. (Eng.) degree from the University of London, London, England, in 1953.

Since 1953, he has worked at British Telecommunications Research Ltd., Maidenhead, England, now part of the Plessey Electronics Group, and has been mainly concerned with various aspects of transmission systems. He is currently Head of the Transmission Techniques Division at B.T.R., with primary interests in the fields of digital communications and speech processing.

Mr. Purton is a member of the Institution of Electrical Engineers (London).