samples are spaced most closely should next be observed. Finally, the vowel is identified by noting the particular vowel volume in which this portion of the trace lies.

The observations in this study are in terms of acoustical data. To those familiar with the intricacies of articulatory formations, however, the conclusions are probably what one would expect. In fact, a close scruting of the spectrograms in the book *Visible Speech* seems sufficient to show that these are the principles at work.[9]

[9] Potter, Kopp, and Green, *Visible Speech* (D. Van Nostrand Company, Inc., New York, 1947).

Several members of the Bell Laboratories have been of great assistance to the author in carrying out this study. I am indebted to E. E. David, H. K. Dunn, R. K. Kraichnan, and J. C. Steinberg for valuable suggestions about the method and for discussions of the general problem. I should particularly like to thank Mr. A. J. Prestigiacomo, who constructed the magnetic tape repeater and who set up and calibrated the remaining equipment employed in the study. While the author served as experimenter in the free space room, Mr. Prestigiacomo operated the recording and sound analyzing equipment throughout the experiment.

THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOLUME 24, NUMBER 6 NOVEMBER, 1952

Automatic Recognition of Spoken Digits

K. H. Davis, R. Biddulph, and S. Balashek
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received August 11, 1952)

The recognizer discussed will automatically recognize telephone-quality digits spoken at normal speech rates by a single individual, with an accuracy varying between 97 and 99 percent. After some preliminary analysis of the speech of any individual, the circuit can be adjusted to deliver a similar accuracy on the speech of that individual. The circuit is not, however, in its present configuration, capable of performing equally well on the speech of a series of talkers without recourse to such adjustment.

Circuitry involves division of the speech spectrum into two frequency bands, one below and the other above 900 cps. Axis-crossing counts are then individually made of both band energies to determine the frequency of the maximum syllabic rate energy within each band. Simultaneous two-dimensional frequency portrayal is found to possess recognition significance. Standards are then determined, one for each digit of the ten-digit series, and are built into the recognizer as a form of elemental memory. By means of a series of calculations performed automatically on the spoken input digit, a best match type comparison is made with each of the ten standard digit patterns and the digit of best match selected.

INTRODUCTION

It is well-known that speech presented in visible form by any of several speech translators which have been described in the literature can be read by individuals after a requisite period of learning. This general background has guided our efforts to recognize speech by machine methods. We wish to stress, however, that spectrographic presentation of speech is a function of the methods and circuits employed in the analysis and that its interpretation in visual form depends upon sense perception and the learning process. The great majority of situations with which we identify the concept of recognition will be found to involve ultimate human perception. In fact, the element of human perception is difficult to disassociate, in our thinking, from the concept of recognition. In machine recognition with which we are here concerned, however, no human brain with its prodigious memory and its uncanny ability to piece together fragmentary information is involved. The problem is essentially one of speech analysis, of selecting and coding recognition elements in the analysis, and of the design of circuits capable of interpreting the code.

The variability encountered in repeated speakings of a digit, even when uttered by the same individual, is common knowledge. Since design of any successful recognition circuit demands a quantitative knowledge of an inherently variable speech signal, any useful description of this signal must be expressed in statistical terms. Possessing such a statistically quantitative description, a search must then be made for physical characteristics of this statistical speech which show significant factors capable of utilization in recognition type circuits which can be designed and built.

Bearing in mind the statistical nature of speech and the variability which this imposes on any criterion that may prove useful for recognition, a logical program for identification of the digits is inherent in a process we will call pattern matching and which we define as follows. Pattern matching involves comparing some aspect or aspects of data derived from an unknown signal with the corresponding aspect or aspects of a number of known signals considered as standards to determine which one of the standards was intended by the talker when speaking the unknown digit. This definition implies that recognition is relative, in which case best matching procedures are definitely desirable. We are faced with the problem of actually determining

what was in the mind of the speaker, knowing that his ability to express his thought as a sound may differ from that of another speaker or even from his own previous utterance of the same sound. Therefore, we are not looking for something that matches any standard sound exactly, but rather something that resembles one standard more closely than any other.

Undoubtedly, when one deals only, as in the present case, with a limited vocabulary of ten digits, several characteristics of speech could supply recognition criteria. There does not seem to be any *a priori* way to determine whether chosen characteristics are valid for recognition, short of making a considered engineering guess, measuring the chosen criteria in a sufficient range of conditions to assemble statistical data and then carrying through a paper analysis of system reliability. If error margins prove deficient, other physical characteristics of the speech universe must be sought and, either alone or in combination with previously studied factors, restudied for their combined margin against error. It may be safely assumed that as either an enlarged recognition vocabulary or greater margin against error is demanded of the machine, an increased amount of data on the physical characteristics of speech must be extracted and processed by it. Thus, any actual machine will break down and err when it is overloaded, i.e., when the input speech deviates beyond the design centers of circuitry. It is apparent also that choice of the physical characteristics of speech to be explored for their utility in recognition will be influenced considerably by the practicability of circuitry which would be required in order to make use of the characteristics chosen.

With these preliminary remarks regarding the

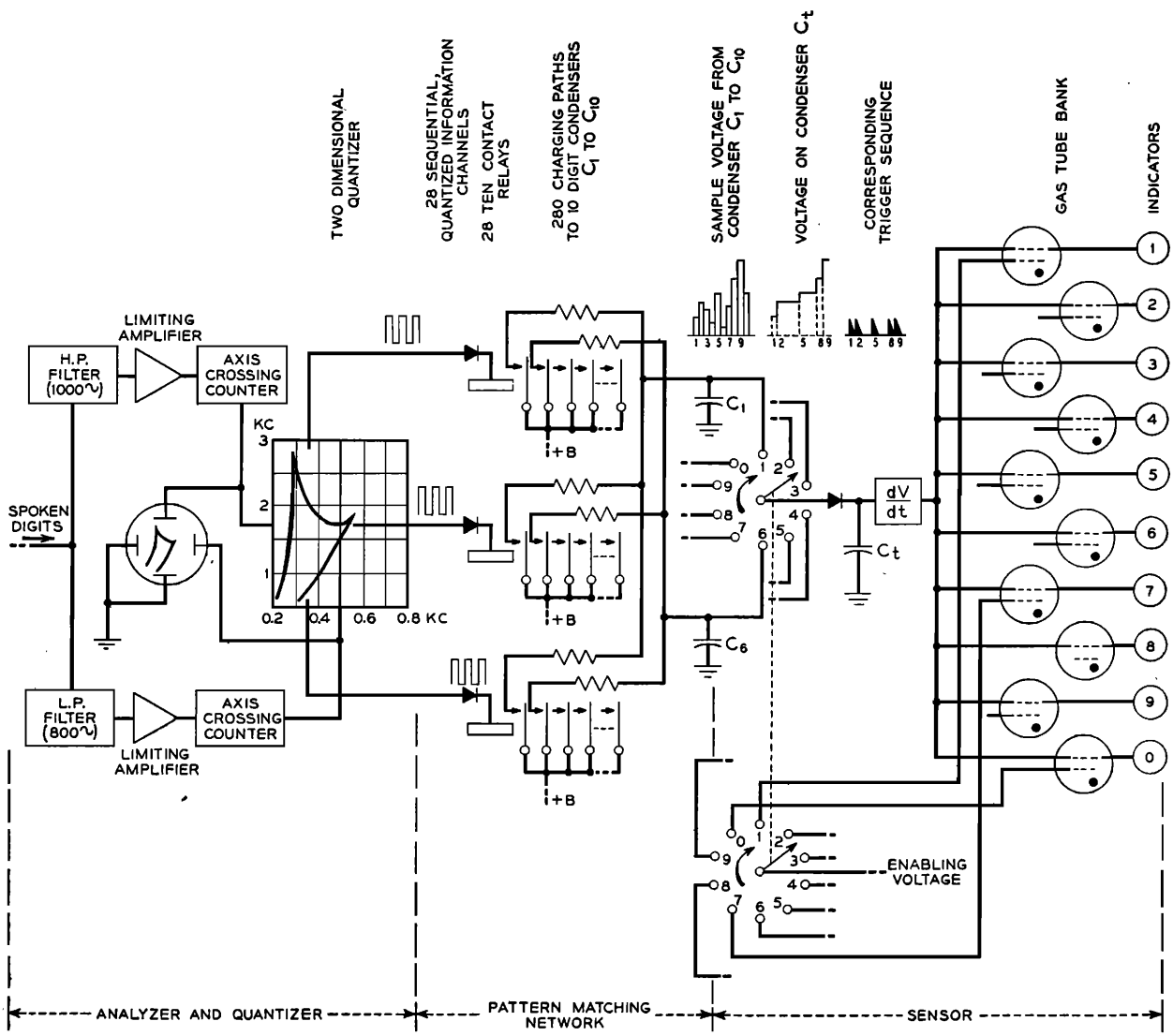

FIG. 1. Block schematic of digit recognizer circuits.

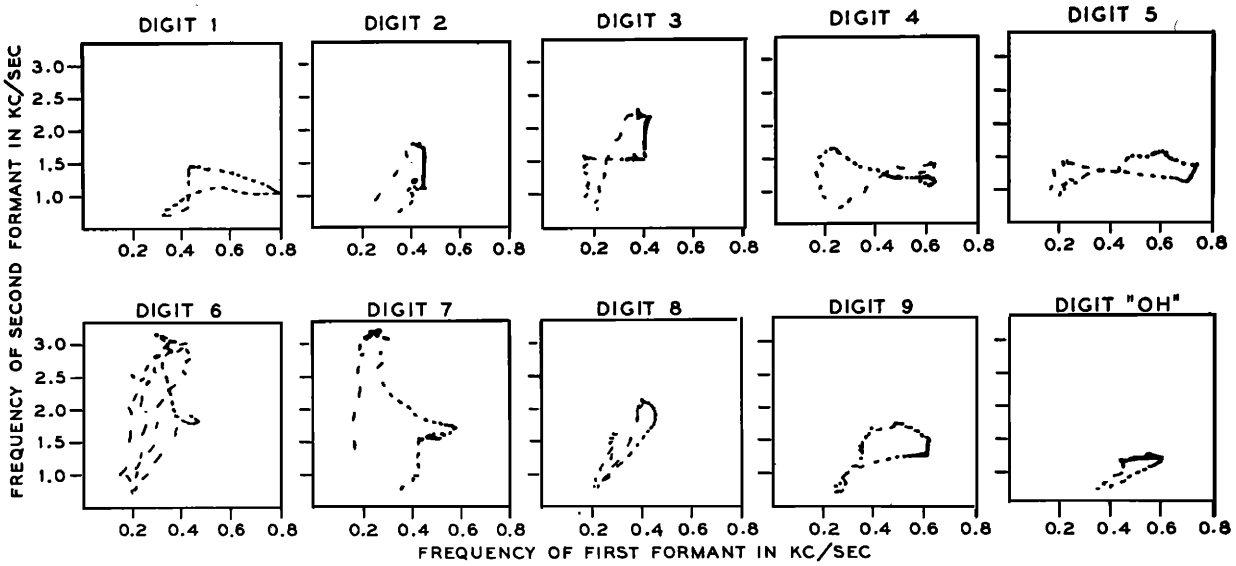

FIG. 2. Photographs of formant 1 *vs* formant 2 presentation of the digits. Trace interruption period = 10 ms. Recognition criteria depend upon significant differences in these shapes and upon their relative duration in the frequency space.

problem of machine recognition of speech, we shall now outline circuits which perform successfully in the limited digit universe with which we are concerned.

CIRCUIT DESCRIPTION

We have purposely limited the speech universe with which we deal to the ten digit series, 1, 2, . . . 9, 0, when the series is spoken by a single talker. Our primary objective has been to study methods of machine speech recognition; thus adaptation to the variability inherent in a larger speech universe has been delayed until problems associated with a single speaker are better understood. That this approach is not without merit is indicated by the error expectancy of the system here described. If a speaker pauses 350 ms between digits, the recognizer will automatically follow his utterance and then will illuminate an appropriately labelled signal indicating which digit of the series has been spoken. An accuracy varying between 97 and 99 percent is obtained when a single male speaker repeats any random series of digits. After preliminary analysis of the speech of any individual, the circuit can be adjusted to function satisfactorily for the speech of that individual. The recognizer is not, however, in its present form, capable of such accuracy on speech of a series of talkers without adjustment for each talker. If no adjustment is made, accuracy may fall to as low as 50 or 60 percent in a random digit series. The errors, however, are usually confined to two- or three-digit pairs.

TABLE I. Maximum and minimum formant frequencies of vowels in repetitive speakings of 33 men and 28 women.

Vowel	Formant 1		Formant 2		Formant 3	
	Max	Min	Max	Min	Max	Min
[u]	480	210	1430	570	3300	1850
[i]	406	190	3100	2000	3900	2600
[ɜ]	652	360	2120	1130	2480	1400
[ɑ]	1040	592	1470	820	3180	2020
[ɪ]	534	206	2700	1710	3400	2340
[ɛ]	760	370	2570	1650	3300	2200
[ʌ]	910	550	1688	880	3250	1950

Formant structure studies of speech have guided the design of this recognizer. It is well established that the formant frequencies, particularly of first and second formants, serve as an important criterion in human recognition of steady state vowels. Coupled with their time changes in dynamic speech, those formant frequencies also provide our most useful quantitative specifications of normal conversation. It is assumed that formant frequencies will also prove of equal relative usefulness in machine recognition of spoken digits.

A functional diagram of the digit recognizer circuitry is sketched in Fig. 1. Recognition involves three separate sequential functions. Speech is first analyzed in a manner to produce a formant 1 *vs* formant 2 two-dimensional plot. A pattern matching network then compares an unknown signal distribution with each of a set of 10 reference distributions to determine which pattern is most like the unknown. Finally an indicator and the necessary associated circuits present this information visually.

Analysis

Speech signals are first divided into two frequency bands. Frequencies above 900 cps appear in the upper path of Fig. 1, while those below 900 cps follow the lower channel. To a first approximation, this routes formant 2 and all higher frequency energy of the digits via the upper path and formant 1 energy into the lower circuit. The signals in these two channels are then each severely limited by circuits designed to remove ampli-

TABLE II. Reference patterns of spoken digits against which an unknown digit is matched.

Squares $f_V(KC)$	Squares $f_H(KC)$	Digits 1	2	3	4	5	6	7	8	9	0
	0.3–0.4	120	128	14		11		62		51	24
	0.4–0.5	70	22		40	23		36			118
0–1.0	0.5–0.6	26			143	46					106
	0.6–0.7	40			21	38					
	0.7–0.8	16									
	0.2–0.3	11		33	52	43	15		19		
	0.3–0.4	27	59	60					15	37	
	0.4–0.5	17	70	33	20	23	25	45		64	
1.0–1.5	0.5–0.6	40			31	100	10	64		97	18
	0.6–0.7	49				99		31		88	
	0.7–0.8					10					
	0.2–0.3			17	22	42	12		15		
	0.3–0.4		46	130			10		53	13	
1.5–2.0	0.4–0.5		32	57			35		133	37	
	0.5–0.6								15		
	0.2–0.3				15		12		25		
2.0–2.5	0.3–0.4								20		
	0.4–0.5								19		
	0.2–0.3						150	120	46		
2.5–3.0	0.3–0.4						34	12	19		

tude characteristics of the signals, even in low-level consonant portions of the digit. Frequency changes of the signal in each band are then tracked at normal speech rates by axis crossing counters. The function of the counters is to locate and indicate, by syllabic rate unidirectional outputs, the frequency of the highest energy in each band. It should be noted that a frequency counter to which a multi-formant speech signal is applied is subject to errors which depend upon frequency separation of the formant energies, relative amplitudes of the formants and the linear instantaneous frequency range built into the detection circuit.[1]

This approach to the recognition problem, by way of two-band analysis using frequency counters, is feasible because of the nature of our digit universe. Table I, compiled from consonant-vowel-consonant word lists[2] measured by Peterson and Barney, shows maximum and minimum formant frequencies encountered in speaking those vowels measured by these authors, which also occur in the digit series. These word lists comprise repetitive speakings by 33 men and 28 women. It will be noted that [u] is the only vowel in which separation into bands above and below 900 cycles does not effectively separate the first formant of any of the measured vowels from its second and third formants. Reference 1 also indicates the relationship of frequencies and amplitudes of second and third formants that must exist if an audio band axis counter is to correctly indicate the frequency of the maximum amplitude formant. Amplitude data from the word lists of reference 2 show that except for the front vowels [i] and [ɪ] of certain speakers, formant 2 will always be correctly indicated in the presence of formant 3. Based on these considerations, 900-cycle high and low pass filters with associated counters were chosen for use in this circuit.

The two syllabic rate signals derived are now plotted against one another, bar 1 information horizontally, bar 2 vertically. The two-dimensional plot can be viewed at this point in the circuit on an oscilloscope monitor screen. Figure 2 is a composite photograph of the screen and illustrates the distinctive bar 1–bar 2 traces obtained when each of the digits is spoken into the system input. These pictures, although obtained with a lower cutoff syllabic rate filter than is used in actual recognition work, are a visual presentation of the basic data that is useful for recognition. The photographs will be recognized as a form of visible speech presentation in which time appears as a non-linear distance along the trace. Such high band, low band traces do not indicate voiced or unvoiced portions of the digit, contain no amplitude information, nor do they show subsidiary resonances as do conventional spectrograms, yet one can readily identify the digits visually, based only upon the shape of the trace they produce. It is the absolute value of the formant frequencies and their relative duration in any position along the trace of this identifying form, which is the basis for our machine recognition.

We shall now consider the problem of deriving recognition information from voltages producing the digit traces. Let us quantize each frequency detector syllabic voltage into steps. Low band information is divided into six 100-cps wide intervals beginning at 200 cps and extending to 800 cps. High band information is similarly quantized into five 500-cps increments, the first between 500 and 1000 cps, the last interval beginning at 2500 cps and including all energy in the signal found above this frequency. Plotting these two syllabic rate signals in rectangular coordinates, we obtain 30 frequency elements each 100 by 500 cycles in area, which we shall call squares. Actual quantization is done by means of successively biased gas tetrodes feeding twin grid tubes to produce the required voltage slots. High and low band slot voltages are then fed to coincidence tubes. One coincidence tube is associated with each of 28 of the 30 squares. Thus 28 individual information channels are made available sequentially. As the speech trace travels about the frequency area, an output will be obtained from a given channel as long as the trace dwells within the particular area associated with that channel.

Some of the channels will never be occupied by any speech sound occurring in the digits. Actually, in subsequently presented data (Table II), only 20 of these areas are ever entered by the trace.

In order to explore each digit pattern for recognition criteria which could be used to select that digit from all other digits, let us associate with each one of the 28 squares a relay arranged to charge one of 28 identical

[1] E. Peterson, J. Acoust. Soc. Am. **23**, 668 (1951).
[2] G. E. Peterson and H. L. Barney, J. Acoust. Soc. Am. **24**, 175 (1952).

condensers through a resistance when the contacts of the relay are closed. All 28 condenser-resistance combinations possess identical time constants so chosen that when a digit is repeatedly spoken into the circuit the total charge on the condenser associated with the square in which the trace dwells longest does not exceed the essentially linear section of the condenser-resistance charging function. Thus we are in a position to explore the occupancy of our frequency space as a digit is spoken. Further, we are in a position to measure the duration or dwell of the trace energy within any and all squares of the frequency area. We can also obtain a statistical pattern of this dwell in each square by simply repeating the same digit into the circuit a statistically sufficient number of times and measuring the final charge distribution over all 28 condensers. In this way, by approximately 100 repetitions of each of the ten digits, the data of Table II were obtained. These data have been normalized for average value and standard deviation. The average value of each digit was made zero, and the corresponding standard deviation adjusted to unity.

The data of Table II show the expected durational distribution in our frequency space when each of the ten digits is spoken into the circuit. A single speaking of any digit may differ somewhat from these average values. However, a measure of the usefulness of any most probable digit distribution in distinguishing its digit from any other most probable digit can be obtained by cross multiplying the charge distribution of the digit in question by itself and by that of each of the other most probable distributions and summing resultant individual informational units. This procedure provides Table III. This table expresses in quantitative terms the similarities or differences between digits. It shows which digit is most likely to be incorrectly recognized and indicates the digit to which error will most likely be made. It also provides a quantitative measure of existence of recognition factors in the input speech data after these data have been operated on by the circuit, and is thus a useful tool for study of circuit parameters under conditions of actual operation. The table shows that, if errors are made, the digit pairs 4 and 0, 5 and 9, 6 and 7, and 1 and 2 will, in that order, most probably err to the other digit of the pair. Of these pairs, only 5 and 9 tend to be interchanged in normal conversation.

Pattern Matching

When a talker speaks an unknown digit, the function of the pattern matching network of Fig. 1 is to find which of the digit reference patterns of Table II is best matched by the new incoming data. Since, in general, the incoming data will not have dwell values in each row of Table II that correspond identically to individual entries for the appropriate digit in the table, it becomes a statistical problem to discover this best match. It is best described in statistical terms by stating that we are looking for the highest relative correlation coefficient between a set of the new incoming data and each of the reference digit patterns.

The correlation coefficient is given by

$$r=\frac{\sum_{i=1}^{i=n}\frac{x_i y_i}{n}-\bar{x}\bar{y}}{\sigma_x\sigma_y},$$

where

$$\bar{x}=\frac{x_1+x_2+\cdots x_n}{n},$$

$$\sigma_x=\left(\sum_{i=1}^{i=n}\frac{x_i^2}{n}-\bar{x}^2\right)^{\frac{1}{2}},$$

with corresponding expressions for $\bar{y}$ and σ_y. Consider x_i to represent sequential channel contributions of an unknown signal to any final digit voltage and y_i the corresponding series of contributions built into the pattern network. Now design of circuits useful for performing the required correlations are very considerably simplified by the following two propositions. Specifically, in determining the pattern standards of Table II from original data, it is required to normalize the standards as regards average value and standard deviation. Now it can be shown that (1) any constant may be added to or subtracted from the individual values of a standard signal without destroying the validity of its correlation with the unknown, and (2) any multiplying factor may be applied to the individual values of a standard signal without destroying the validity of its correlation with the unknown. Application of these propositions makes it possible to adjust the average values and standard deviations of a set of standard signals to a given normal value before computing r. This permits an electrical circuit to give information concerning the relative values of r while only computing the simple term

$$\sum_{i=1}^{i=n}\frac{x_i y_i}{n}$$

for each match measured.

TABLE III. Charge margins resulting from utterances of average digits.

Digit spoken	Resultant relative condenser voltages									
	1	2	3	4	5	6	7	8	9	0
1	1.0	0.75	0.16	0.39	0.59	0.04	0.56	0.02	0.63	0.56
2	0.75	1.0	0.49	0.10	0.13	0.11	0.45	0.21	0.55	0.22
3	0.16	0.49	1.0	0.07	0.10	0.19	0.09	0.66	0.34	0.01
4	0.39	0.10	0.07	1.0	0.52	0.07	0.18	0.04	0.16	0.83
5	0.59	0.13	0.10	0.52	1.0	0.09	0.46	0.04	0.79	0.37
6	0.04	0.11	0.19	0.07	0.09	1.0	0.77	0.47	0.15	0.01
7	0.56	0.45	0.09	0.18	0.46	0.77	1.0	0.18	0.57	0.26
8	0.02	0.21	0.66	0.04	0.04	0.47	0.18	1.0	0.24	0
9	0.63	0.55	0.34	0.16	0.79	0.15	0.57	0.24	1.0	0.12
0	0.56	0.22	0.01	0.83	0.37	0.01	0.26	0	0.12	1.0

The circuit method for obtaining the required product-sums is theoretically simple. Closure of the contact of a relay of Fig. 1 indicates presence of the trace in the allied square. Let us associate 10 contacts with each relay all of which open and close simultaneously. To each contact assign a digit conductance value specified by the horizontal line entry of Table II which corresponds to the square, relay, and digit under consideration. Let each conductance lead to one of 10 identical condensers. When the relay operates and releases, if the time constants are well chosen, charges will build up on each condenser proportional to the expectancy values in the table which we have previously used to determine the conductances. Now associate appropriate conductances with all other relays and connect them in parallel with the corresponding conductance values of the first relay, to these same condensers. The charge delivered to each one of the ten condensers by operation of a single relay will be proportional to the time of closure of the relay contacts and the magnitude of the conductance associated with each one of the ten contacts. Sequential operation of the relays throughout the total duration of the digit then builds up charges across the 10 identical condensers proportional to the vertical product-sums discussed earlier. Designation of the standard best matched to an unknown signal then involves specification of the one condenser in ten carrying the greatest charge.

Indication Circuits

Reference to the wave forms of Fig. 1 will indicate the principle upon which specification of maximum charge is based. The charges on all condensers are transferred sequentially to a holding condenser C_T with which is associated a circuit producing a trigger pulse each time a voltage increase occurs on the holding condenser. This trigger pulse passes to the control grids of 10 parallel gas tetrodes. Synchronously with the condenser sample, an enabling grid of the appropriate tube is activated. A gas tube will then fire each time a voltage increase occurs on C_T, and will simultaneously extinguish any previously fired tube. Thus the tube left conducting when the sample is complete is a specification of the incoming digit.

SUMMARY

Circuitry which has been described is able to recognize, by use of formant 1 *vs* formant 2 plots of unknown sounds matched against similar plots of reference sounds, approximately 98 percent of the random digits spoken by the talker for whom the machine is adjusted.

Best matching techniques are used in a way that normalizes for various lengths of digits and speeds of talking. These techniques also permit the machine to deliver a determination even when the best match is poor. Determinations caused by short unintentional sounds or noises are eliminated by time constants of the sampling circuits. Although it may be desirable in some applications, it is not practical with this circuitry to allow the machine to reject a word or sound not within the accepted vocabulary.