# Spoken Digit Recognition Using Time-Frequency Pattern Matching

P. DENES
*University College, London, England*

AND

M. V. MATHEWS
*Bell Telephone Laboratories, Inc., Murray Hill, New Jersey*

A study of the machine recognition of the spoken digits zero through nine has been carried out by a digital computer simulation. The spoken utterances were converted to time-frequency patterns of spectral energy. Recognition was done by cross correlating the pattern of an unknown utterance with a test pattern for each digit and selecting the digit having the highest correlation. Time normalization could be applied to all patterns, thus reducing utterances to a standard duration. Six male and one female speakers provided 38 samples of each of the 10 digits. Pauses were made between successive words for segmentation.

No errors were observed recognizing a single speaker using test patterns from his own speech with time normalization. A group of five male speakers and test patterns averaged over the group produced 6% errors with time normalization and 12% without. A 25% rate occurred for the woman matched against male patterns.

The study indicates both the effectiveness and limitations of this simple recognition procedure for limited vocabulary and limited number of speakers. Time normalization improves performance in all cases.

AUTOMATIC speech recognition is probably possible only by a process that makes use of information about the structure and statistics of the language being recognized as well as of the characteristics of the speech sound wave. The digit recognition process described in this paper, on the other hand, utilizes no linguistic information but only methods of detecting acoustic characteristics. It is hoped that by restricting the library of words to be recognized to the relatively small number of 10, the acoustic redundancy of the speech waves will be increased to a level where linguistic information is no longer required for successful recognition. Also, the words were always spoken with sufficiently long silent intervals between them to make the recognition of the beginnings and ends of the words easy; this made the usually difficult problem of segmentation a simple matter, again reducing the need for the use of linguistic information.

This is, of course, by no means the first digit recognition experiment. Previously reported are work by Davis, Biddulph, and Balashek[1]; a recognizer named "Audrey" developed by Dudley and Balashek[2]; studies by Forgie, Groves, and Frick[3,4]; and the work of Shultz.[5] In addition Kersta[6] carried out a program involving matching binary-spectral patterns. All these studies and the experiment described here use some form of short time power spectra as original information. "Audrey" proceeded in two steps, first selecting a frequency pattern, then choosing a digit on the basis of the integrated occurrences of the frequency patterns. Forgie's method involved recognizing certain spectral features, then proceeding to a final choice along the branches of a "decision tree." Shultz also measured certain spectral features, but used these in a Bayes' decision procedure. The present work differs from these in that it involves matching time-frequency patterns with time normalization to compensate for variable speaking rates.

The first step in the recognition procedure consists of producing a reference pattern, in the form of a time-frequency spectrum, for each of the ten digits. The reference patterns are obtained by averaging and normalizing the spectra from a number of utterances of each of the digits. Subsequent utterances, which are to be recognized, are formed into similar patterns and are then matched by a process of cross correlation with each of the reference patterns in turn. The best match is selected by finding the reference pattern which gives the greatest correlation coefficient with the pattern to be recognized. Such a pattern-forming pattern-matching process could be carried out using a digital computer. The problem was also a fairly simple one and was therefore considered a suitable exercise for one of the authors of this paper (P. D.) who wanted to gain some elementary experience in the use of a digital computer for speech research.

The rest of the paper will describe the way in which the acoustic data were converted into digital numbers suitable for processing by the computer; then it will explain the computer program that implemented the digit recognition procedure and this is followed by an account of the experiments that were actually carried out, the results obtained, and an assessment of their significance.

## PREPARATION OF THE ACOUSTIC DATA

A number of different speakers were used to provide the test material. Their utterances of the digits were

[1] K. H. Davis, R. Biddulph, and S. Balashek, J. Acoust. Soc. Am. **24**, 637 (1952).

[2] H. Dudley and S. Balashek, J. Acoust. Soc. Am. **30**, 721 (1958).

[3] C. Forgie, M. L. Groves, and F. C. Frick, J. Acoust. Soc. Am. **30**, 669(A) (1958).

[4] F, C. Frick, "Research on speech recognition at Lincoln Laboratory," Proceedings of Seminar on Speech Compression and Processing, September, 1959, Air Force Cambridge Research Center, Bedford, Massachusetts.

[5] G. L. Schultz, "Investigation procedures for speech recognition," Proceedings of Seminar on Speech Compression and Processing, September, 1959, Air Force Cambridge Research Center, Bedford, Massachusetts.

[6] L. G. Kersta, unpublished work done at Bell Telephone Laboratories.

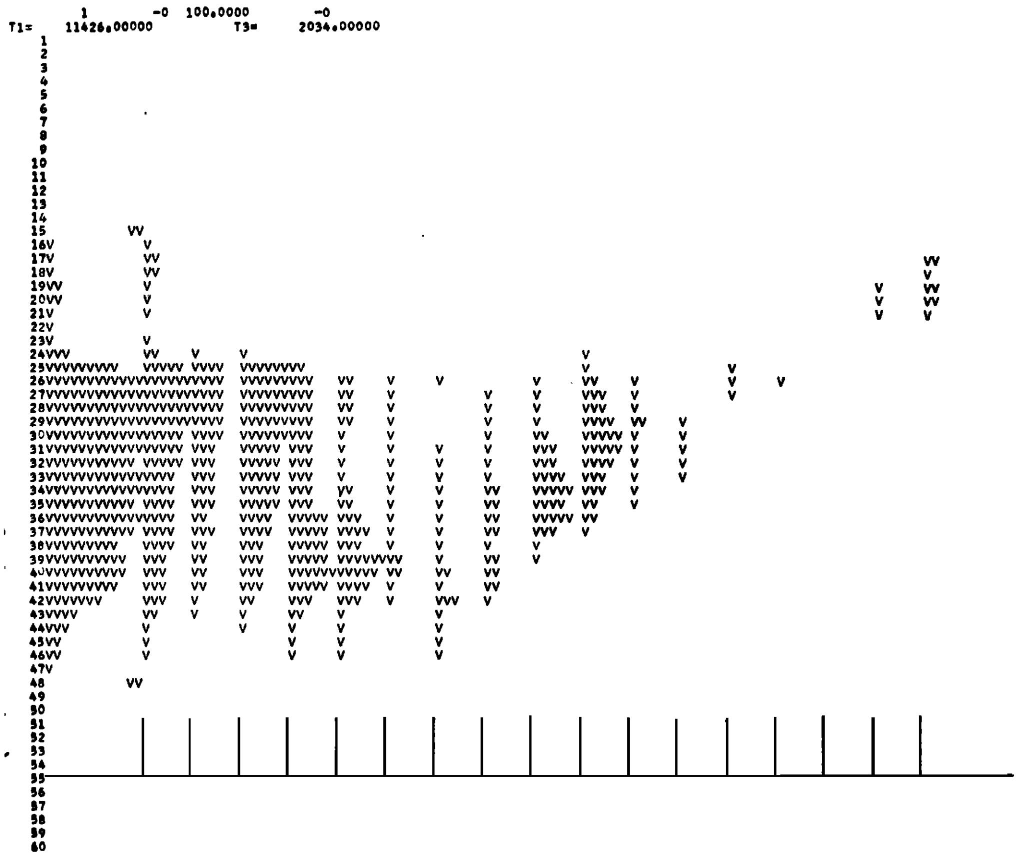


FIG. 1. Digital spectrogram.

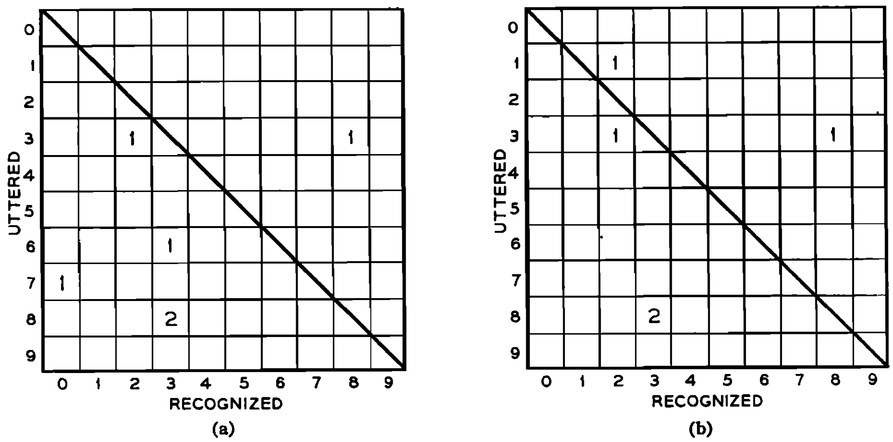
recorded on an Ampex 300 tape recorder. The recordings were made in a sound-treated room in which reverberation and ambient noise were reduced to a resonably low level. Spectral analysis of the test sound waves was obtained by applying the recordings to a filter bank of 17 channels. Each channel consisted of a bandpass filter, a rectifier, and a low-pass filter. The bandpass filters had center frequencies arranged at approximately equal intervals along the Koenig scale and covered the range 200 to 4000 cps. The crossover point between adjacent filters occurred at the 6-db point. The low-pass filters had a cutoff frequency of about 25 cps. The outputs of the channels were sampled sequentially at about 70 times per sec by a multiplexer and the analog samples were quantized into 10 bit binary numbers and the numbers recorded on digital tape suitable for use with an IBM 704 computer. The multiplexer, analog-digital conversion and digital tape recording equipment is described elsewhere.[7] Each utterance of a spoken digit was recorded in a separate block of digital data; this was arranged by manually positioning the analog tape playback heads just before the beginning of the desired digit. Automatic timing equipment then started and stopped the analog and the digital equipment in the proper sequence so as to position the utterance near the beginning of the data block. The digital tape was then edited, the first 60 complete sweeps of the multiplexing switch across the bank of filters being used as data, giving $60\times 17=1020$ samples for each spoken digit. This digital tape recording of the test utterances provided the data input for the computer program.

## PROGRAMS

The computer programs performed three basic functions: tabulating spectral patterns, forming reference patterns, and recognizing utterances. The latter two functions could be carried out either with or without time normalization.

The tabulation of the spectral patterns proved to be a very effective means of monitoring the operation of the recognition programs. As an example a typical pattern of the utterance "zero" is shown on Fig. 1. The 60 time sections are arranged along the ordinate and the outputs of the 17 filter channels along the abscissa as the second through the 18th bands. The number of $V$'s in a band is proportional to the channel output, the maximum energy in the entire spectrogram being set equal to 6 $V$'s. The first band is proportional to the total speech energy obtained by summing over the 17 channels with the constant of proportionality being adjusted to make the maximum energy equal 12 $V$'s.

[7] E. E. David, Jr., M. V. Mathews, and H. S. McDonald, "Description and results of experiments with speech using digital computer simulation," 1958 IRE Wescon Convention Record, Part 7.


UTTERED
RECOGNIZED
(a)
(b)

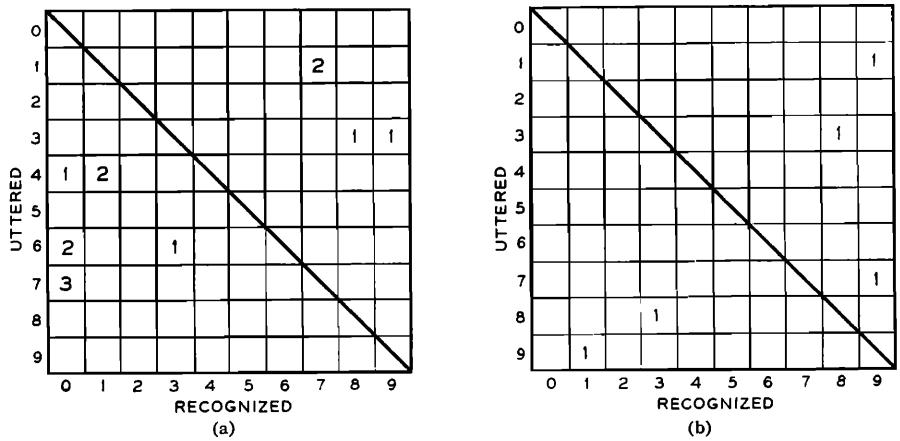
(a)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | | | |
| 1 | | | | | | | | | | |
| 2 | | | | | | | | | | |
| 3 | | | 1 | | | | | | 1 | |
| 4 | | | | | | | | | | |
| 5 | | | | | | | | | | |
| 6 | | | | 1 | | | | | | |
| 7 | 1 | | | | | | | | | |
| 8 | | | | 2 | | | | | | |
| 9 | | | | | | | | | | |

(b)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | | | |
| 1 | | | 1 | | | | | | | |
| 2 | | | | | | | | | | |
| 3 | | | 1 | | | | | | 1 | |
| 4 | | | | | | | | | | |
| 5 | | | | | | | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | | | | |
| 8 | | | | 2 | | | | | | |
| 9 | | | | | | | | | | |

Fig. 2. Recognition with five speaker patterns and time normalization: (a) 99 different utterances by same five speakers, error rate = 6%; (b) 147 utterances used to form patterns, error rate = 3.5%.


UTTERED
RECOGNIZED
(a)
(b)

(a)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | | | |
| 1 | | | | | | | | 2 | | |
| 2 | | | | | | | | | | |
| 3 | | | | | | | | | 1 | 1 |
| 4 | 1 | 2 | | | | | | | | |
| 5 | | | | | | | | | | |
| 6 | 2 | | | 1 | | | | | | |
| 7 | 3 | | | | | | | | | |
| 8 | | | | | | | | | | |
| 9 | | | | | | | | | | |

(b)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | | | |
| 1 | | | | | | | | | | 1 |
| 2 | | | | | | | | | | |
| 3 | | | | | | | | | 1 | |
| 4 | | | | | | | | | | |
| 5 | | | | | | | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | | | | 1 |
| 8 | | | | 1 | | | | | | |
| 9 | | 1 | | | | | | | | |

Fig. 3. Recognition with five speaker patterns: (a) 99 different utterances by same five speakers, no time normalization, error rate = 13%; (b) 20 utterances by women with time normalization, error rate = 25%.

A visual examination of these patterns gives a good idea of the correlations which would be obtained for the utterances and the thresholds needed for the time normalization.

The tabulated patterns are an image of the patterns in the computer memory. However, the former have only seven amplitude levels for each point in time and frequency while the latter represent each point with a 35 digit binary number.

If time normalization were to be used, the beginning and end of each utterance was located, respectively, as the time when the total energy first exceeded a threshold and the time when the energy last fell below this threshold. A new pattern was then formed by stretching the occupied section of the utterance to a standard length of 60 time units. If the pattern is visualized as being painted on a rubber sheet, the normalization consists of stretching the sheet until the distance between the beginning and end is a standard length.

If no time normalization was to be used, then only the beginning of the utterance was located, this was moved to the first time unit, and any time sections beyond the end of the utterance were filled with zeros so as to make the total pattern 60 sections long. Thus each pattern whether normalized or unnormalized consisted of a two-dimensional array containing 1020 points formed by 60 time sections and 17 frequency sections.

The reference patterns were formed by adding together corresponding array points from a group of utterances of the same digit thereby producing a 1020-point reference array. In this and all other operations, the computer was directed by a sequence of punched cards which specified the sequence of digits to be read from the digital tape. Thus, very flexible control was achieved.

After the patterns were summed, the reference array was amplitude normalized by dividing each point by a constant so that the sum of the squared points in the array equaled one. Such a normalization is essential to avoid biasing recognition results toward digits which have more average energy. Arrays for each of the 10 digits were formed simultaneously and all kept in the computer memory at the same time.

Utterance recognition consists of cross correlating the array of an utterance (with or without time normalization as is appropriate) with the reference arrays of the 10 digits and selecting the digit with maximum correlation. The correlation computation multiplies corresponding points in the reference array and the utterance array and sums these products over the entire array. Thereby, full advantage is taken of both the time and frequency composition of the patterns. The computer printout for each utterance consists of not only the recognition choice, but also the correlation coefficient with each of the 10-digit patterns. Thus a measure can be obtained of the margins involved in the choices and the most probable errors.

## EXPERIMENTAL RESULTS

In the experiment described in this paper seven speakers were used to provide altogether 38 utterances of each of the 10 spoken digits. Six of the speakers were men and one a woman. The speakers were asked to pronounce the digits zero to nine and then to repeat this sequence a number of times. The only instruction they received about the way they were expected to speak was that they should pause, however briefly, between the individual digits. The clearness of articulation, inflection and rate of speaking varied very widely from speaker to speaker.

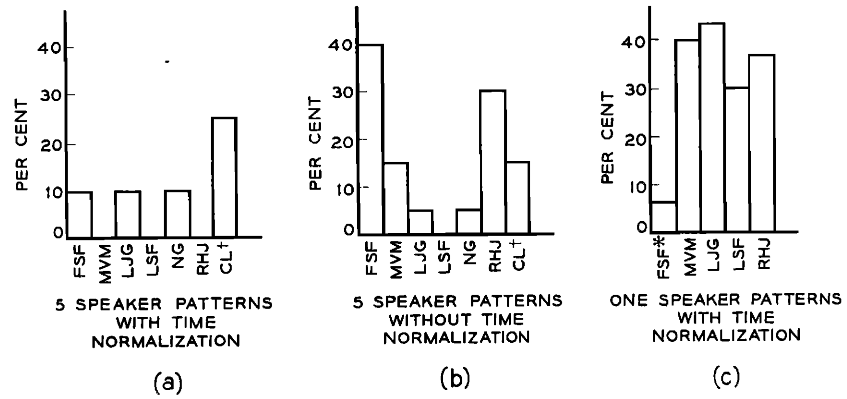
The data were used in a number of separate experiments. First tabulated spectral patterns for all the utterances were examined visually. The examination showed that each digit formed a distinctive pattern whose features correspond to those which would be expected from experience with sound spectrograms. Features such as stops, friction, and formants were apparent. Visual recognition from the tabulations appeared feasible, though no quantitative tests were made. An appropriate threshold with which to determine the beginnings and ends of the utterances was selected by a study of total energy displayed in the first band. The threshold selected located appropriate points for all the digits except "three" and "eight." With "three," an initial /θ/ could neither be included nor rejected uniformly and with "eight" the same difficulty occurred for a final aspiration. Consequently some difficulty was expected in recognizing these digits.

A set of five speaker reference patterns was then formed by summing three utterances of each of the 10 digits by each of the speakers. Reference patterns both with and without time normalization were produced. The patterns were then used to recognize several groups of utterances by the same and different speakers. The recognition results are presented in the form of confusion diagrams on Figs. 2 and 3. For 99 different utterances by the same five speakers, with time normalization, Fig. 2(a), six errors were observed giving an error rate of 6%. For the 147 utterances forming the test patterns, Fig. 2(b), five errors were observed giving an error rate of $3\frac{1}{2}\%$. Because of the small number of errors, these estimates of error rates are, of course, very poor and serve mainly to indicate trends. The confusion diagrams should likewise not be taken very seriously. They do, however, show that a majority of the errors involve either the digits "three" or "eight," thus confirming the conclusions based on the tabulated spectrograms.

Without time normalization the error rate for 99 different utterances by the same speakers, Fig. 3(a) increased to 13%. Also the distribution of errors has changed markedly. Consequently, the time normalization appears to have significantly decreased the error rate. In addition, it tends to spread the errors more uniformly over the different speakers. Figs. 4(a) and 4(b) show the error rate as a function of speaker with and without time normalization. Much greater variation is observed without normalization.

The five speaker patterns were used to recognize the utterances of one man and one woman not in the set of original speakers. For the man, no errors were observed, but as only one utterance each of 10 digits was tested, no conclusions can be drawn. The results of 20 utterances by a woman are given on Fig. 3(b) and show a substantially higher error rate of 25%. Evidently an essential difference exists in the female patterns, though this is not evident from visual examination.

Finally, a set of patterns based on 47 utterances by one speaker with time normalization was formed. No errors were observed when 50 different utterances by the same speaker were tested, but a high error rate of



5 SPEAKER PATTERNS WITH TIME NORMALIZATION

(a)

5 SPEAKER PATTERNS WITHOUT TIME NORMALIZATION

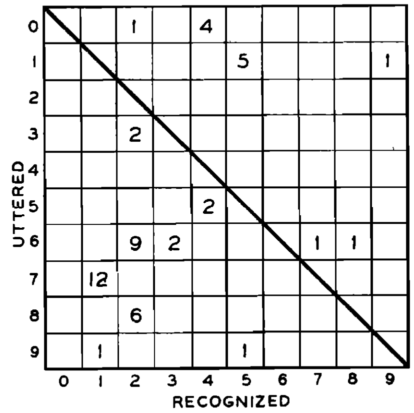
(b)

ONE SPEAKER PATTERNS WITH TIME NORMALIZATION

(c)

FIG. 4. Error rate variation with speaker.

NOTES: † — CL IS WOMAN
* — FSF UTTERANCES FORMED PATTERNS



| UTTERED \ RECOGNIZED | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | 1 | | 4 | | | | | |
| 1 | | | | | | 5 | | | | 1 |
| 2 | | | | | | | | | | |
| 3 | | | 2 | | | | | | | |
| 4 | | | | | | | | | | |
| 5 | | | | | 2 | | | | | |
| 6 | | | 9 | 2 | | | | 1 | 1 | |
| 7 | | 12 | | | | | | | | |
| 8 | | | 6 | | | | | | | |
| 9 | | 1 | | | | 1 | | | | |

FIG. 5. Recognition with one speaker patterns and time normalization: 147 utterances by five speakers, error rate =33%.

33% occurred when the one-speaker patterns were used to recognize 147 utterances by five speakers, four of whom were different, Fig. 5. A plot of error rate as a function of speaker, Fig. 4(c), shows that the error rate is quite uniform for those speakers not in the reference patterns. These results indicate the desirability of having patterns averaged over a number of speakers.

## CONCLUSIONS

Considerations of three general sorts are appropriate in assessing the results of these experiments. First, the performance of the recognition process as simulated should be summarized. Next, certain fairly obvious improvements which do not change the over-all character of the process can be mentioned. Finally, some discussion of the relation of this work to speech recognition as a whole may be profitable.

The performance achieved by the program with time normalization varied from no errors with one speaker recognizing his own utterances to six errors per 100 for a group of five speakers recognizing their own utterances. Performance went from five errors per 150 for the utterances used to form the five speaker patterns to six errors per 100 for the other utterances by the same speakers. Performance for a given process can, perhaps, usefully be considered as a function of the number of speakers, the number of utterances, and the size of the vocabulary. These data are then points on the function for a vocabulary of 10 digits. The most interesting part of this function is, of course, its value for many speakers and many utterances. Unfortunately, both the limited accuracy of the error estimates and the limited number of points on the function preclude estimating this value.

Certain trends in performance, however, are significant. A definite improvement in both error rate and variability of error rate with speaker was achieved by time normalization. Consequently this normalization seems a worthwhile improvement. The error rate reduced substantially when reference patterns were formed from an average of utterances of five speakers rather than from one speaker's utterances. The optimum number of speakers to put in the patterns is thus probably not one. The number is also probably not an exceedingly large one since this may result in "uniformly gray" patterns.

Several minor changes might reduce the error rate considerably. Splitting the "eight" and the "three" each into two separate patterns depending on beginning and ending location should make much more definite patterns and reduce errors involving these digits. The woman's performance with male patterns is quite poor indicating the existence of essential differences in female patterns. This performance could be improved either by having a second set of female patterns, or if the differences are systematic, by a sexual normalization. Such a process may well be connected with frequency normalization. Last, the time normalization was made linearly over the entire digit, whereas a segmentation of the utterance into approximate vowel and consonant segments before normalization would have been very helpful.

The theoretical interest of this experiment to the general field of automatic speech recognition was mentioned in the introduction where it was pointed out that full automatic recognition is only possible when linguistic information is available to supplement the information of the speech sound wave; if the redundancy of the acoustic wave is increased by restricting the vocabulary to be recognized, less linguistic information is required for successful recognition. The way in which the amount of linguistic information available to the recognition process can be traded for the number of words in the vocabulary of a recognizer may give interesting information about the human speech recognition process and perhaps provide useful data for the design of practical recognizers as well. Digit recognition may be regarded as providing a reference point in the mapping out of this relationship, a relationship that will be investigated more extensively as the advantages of using computers, discussed in the next paragraph, become apparent.

There can be little doubt that computers provide considerable advantages for solving many of the problems encountered in speech research. These problems require considerable data storing and processing facilities that are available in the larger computers while being difficult, costly, and time-consuming to provide by conventional electronic means.

The usefulness of the computer is obvious, but how far its advantages can be exploited depends to some extent on the ease with which speech data can be processed into a form suitable for use with the computer, and on the ease with which computer programing can be learned. The present experiment throws some light on these problems. The equipment already available for converting the acoustic data into digital form and for producing the digital tape recording was perfectly satisfactory and could be used for the present experiment without any further modification. Its use did not present any greater problems than are usual in setting up experimental equipment. As for programing difficulties, Fortran[8] was used and this considerably reduced both the difficulties and the time required for programing. Concerning the difficulties of learning to use Fortran, one of the authors (P. D.) had no prior experience of programing. About two weeks were required for learning the elements of Fortran programing, another two weeks for preparing the data tape, and a further two weeks for writing the program, debugging it, and running it. The first results were available six weeks after starting. The present experiment provides further proof that the difficulties of using large digital computers in speech research are not great compared with the advantages gained.

[8] Fortran is an automatic coding procedure for the IBM 704 and 7090 computers. See programer's reference manual on Fortran published by IBM Corporation, New York, New York.

THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOLUME 32, NUMBER 11 NOVEMBER, 1960

# Thermal Relaxation Absorption in Ethylene

JAMES C. GRAVITT*
*Department of Physics and Astronomy, Vanderbilt University, Nashville, Tennessee*

(Received June 29, 1960)

The thermal relaxation absorption in ethylene at temperatures between 0° and 60°C has been obtained by the Tube method for frequencies between 150 kc and 4 Mc per atm. The relaxation time, the transition probability, and the collision efficiency were determined as a function of temperature. The collision efficiency as a function of temperature seems to follow the simple exponential law: $P_{10}\alpha \exp(-c/T^{1/3})$.

## INTRODUCTION

THERE seems to be no conclusive information concerning the relaxation absorption of sound in ethylene ($C_2H_4$). The results so far reported lack the desired agreement between experiment and theory.[1–3] Therefore, the absorption in this gas was determined in order to supplement the previous results and to obtain new data, by an independent method, on a more solid experimental basis. The relaxation times and subsequent calculations obtained here are in substantial agreement with those obtained by McCoubrey *et al.*[4] and McGrath and Ubbelohde[5] who used a dispersion technique to obtain information on collision efficiencies in ethylene and other gases. The present data also extended the measurements to a higher temperature.

There are several proposed theoretical explanations for the mechanism by which energy is transferred in nonelastic molecular collisions[6–9]; however, insufficient data are available to adequately test any of these theories. Therefore, there is a pressing need for experimental information concerning the temperature dependence of the collision efficiency of molecular collisions in gases and vapors. One of the most powerful methods for investigating the collision process in gases and vapors is sonic absorption, which through the agency of velocity dispersion and relaxation absorption yields valuable data concerning the collision process and probability of energy transfer.

The collision efficiency $P_{10}$ as a function of temperature has been obtained for several of the simpler molecules.[10,11] The results for $CO_2$[10a] and $CS_2$[11] appear to satisfy a simple exponential law of the following form:

$$P_{10}\alpha \exp(-c/T^{\frac{1}{3}}),$$

as predicted by Landau and Teller[4]; however, recent measurements in several halogen gases seem to indicate a more complicated temperature dependence for $P_{10}$.[10b] The collision efficiency for ethylene as a function of temperature was determined in an attempt to obtain some information on the problem of energy transfer between a pair of more complex colliding molecules. The procedure for obtaining the collision efficiency from the relaxation absorption is outlined.

The relaxation absorption is related to the relaxation time by[12]

$$\mu = 2\pi\omega\tau R'C_i/[(R'+C_0)C_0+\omega^2\tau^2(R'+C_\infty)C_\infty], \quad (1)$$

* Present address: Midwest Research Institute, Kansas City 10, Missouri.

[1] W. T. Richards and J. A. Ried, J. Chem. Phys. 2, 206 (1934).

[2] W. T. Richards, J. Chem. Phys. 4, 561 (1936).

[3] O. Nomoto, T. Ikeda, and T. Kishimoto, Bull. Koboyasi Inst. Phys. Research 1, 286 (1951).

[4] J. C. McCoubrey, J. B. Park and A. R. Ubbelohde, Proc. Roy. Soc. (London) A223, 155 (1954).

[5] N. D. McGrath and A. R. Ubbelohde, Proc. Roy. Soc. (London) A227, 1 (1954).

[6] L. Landau and E. Teller, Physik. Z. Sowjetunion 10, 34 (1936).

[7] T. D. Rossing and S. Legvold, J. Acoust. Soc. Am. 23, 1118 (1955).

[8] R. N. Schwartz and K. F. Herzfeld, J. Chem. Phys. 22, 767 (1954).

[9] K. F. Herzfeld and V. Griffing, J. Phys. Chem. 61, 844 (1957).

[10] (a) F. D. Shields, J. Acoust. Soc. Am. 29, 450 (1957); (b) F. D. Shields, J. Acoust. Soc. Am. 32, 180 (1960).

[11] J. C. Gravitt, J. Acoust. Soc. Am. 32, 560 (1960).

[12] P. Vigoureux, *Ultrasonics* (John Wiley & Sons, New York, 1951), p. 90.