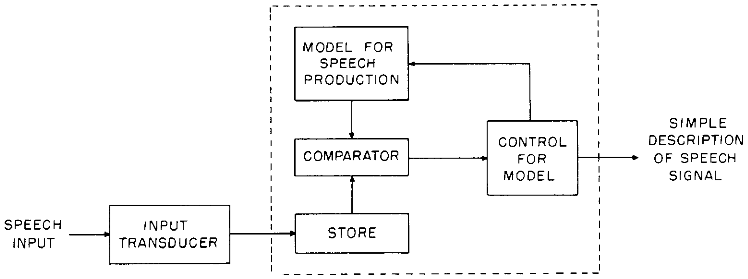
XIV. SPEECH COMMUNICATION

Prof. M. Halle
Prof. K. N. Stevens
Dr. T. T. Sandel
C. G. Bell
G. W. Hughes
Jane B. Arnold
P. T. Brady
O. Fujimura
J. M. Heinz
C. I. Malme
F. Poza
G. Rosen

A. AUTOMATIC RESOLUTION OF SPEECH SPECTRA INTO ELEMENTAL SPECTRA*

In any speech recognition scheme or bandwidth compression system we are faced with the problem of extracting signals that have a low information rate from the speech wave. These signals must preserve enough data so they can be used as a basis either for reconstructing an intelligible version of the original speech or for phonetic or phonemic recognition. In the search for these so-called information-bearing elements, we must take into consideration whatever is known concerning the properties of the speech wave, the perception of speech, and the acoustical theory of speech production.

This report suggests one approach to the problem of extracting low-information-rate signals from the speech wave, and describes how one version of the proposed method has been implemented and tested. The general procedure is shown schematically in Fig. XIV-1. All components of the system enclosed by the dashed line were simulated on a digital computer.



MODEL FOR SPEECH PRODUCTION
COMPARATOR
CONTROL FOR MODEL
SIMPLE DESCRIPTION OF SPEECH SIGNAL
SPEECH INPUT
INPUT TRANSDUCER
STORE

Fig. XIV-1. Block diagram of proposed general approach to automatic reduction of speech wave to low-information-rate signals.

The speech is passed first through a peripheral element or "transducer" whose output is then stored in a set of storage registers. The input transducer performs the function of a set of filters. Also built into the device is a "model" of the speech-production process. This model, suitably controlled, can generate outputs in forms

*This research was supported in part by the U.S. Air Force (Air Force Cambridge Research Center, Air Research and Development Command) under Contract AF 19(604)-2061.

that are compatible with the original stored speech data. It might, for example, consist of a set of equations relating vocal-tract configurations and excitations to output spectra, or it might (as it does in the experiments described here) consist of a set of stored elemental spectra that can be assembled in various combinations to synthesize speech-like spectra. The "comparator" computes a measure of the difference between the input speech data and the data derived from the model, and the comparator output tells the control section to synthesize (according to some systematic plan) new speech data from the model until a minimum error is obtained. The device then reads out the data that describe the settings of the model that produce a best match in the comparator.

Five operations, therefore, are performed in the computer: (a) storage of the speech data processed by the input transducer, (b) synthesis of speech data at the command of (c) a control system, (d) calculation of a measure of the difference between the input speech spectra and the speech spectra computed from the model, and (e) display, in some form, of the settings of the model that yield minimum comparator output. Details of each of these operations for one operating version of the scheme are discussed here.

1. Speech Input System

Sampled speech data are introduced into the computer in spectral form. Figure XIV-2 shows a block diagram of the equipment. Speech is recorded on one channel of a two-channel magnetic tape loop and is played back through a bank of 36 single-tuned filters. The center frequencies of the filters range from 150 cps to 7025 cps and are selected so that the half-power points of adjacent filters are coincident. The bandwidths are constant at 100 cps for center frequencies up to 1550 cps and then increase to 475 cps for a center frequency of 7025 cps. Outputs of the filters are selected in sequence by a stepping switch that steps after each cycle of the tape loop. Thus the loop is played 36 times to obtain a complete spectral analysis of the speech sample. The selected filter output is full-wave rectified, smoothed, and logarithmically amplified before being converted from analog to digital form. A commercial analog-to-digital

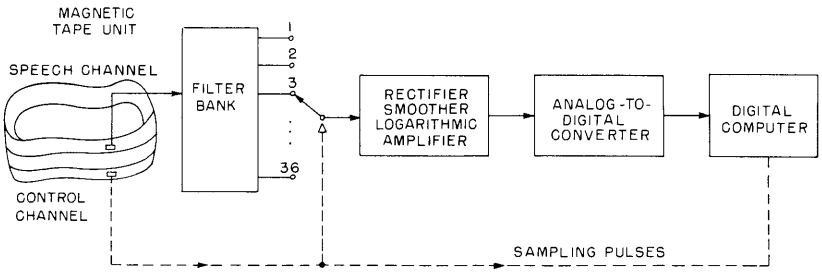


Fig. XIV-2. Input system for speech analysis with digital computer.

encoder performs this conversion.

The second tape channel contains recorded control pulses. A pulse train of positive polarity in which the pulses occur every 10 msec is used to indicate times at which the data are to be sampled. A train of opposite polarity marks the end of the tape loop and initiates the stepping switch. These control pulses enter two light-pen flip-flop registers of the TX-0 computer, so that the sampling is under the control of the computer.

The computer is programmed to search the light-pen flip-flop registers for sample pulses and to transfer data from the encoder when a sampling pulse appears. The filter outputs are encoded into six bits and are read into the computer's live register. Data are rearranged in the computer so that three samples are stored in each 18-bit memory word and each group of 12 words contains outputs of the 36 filters at one sample time. Successive groups of 12 words contain speech spectra at successive 10-msec intervals. With the present 4096-word memory, 2111 words are used for data storage, and thus 1.75 seconds of speech can be processed. The program provides a punching routine that allows the data to be punched out on paper tape for later use. In addition, several error-checking routines are built into the program to maintain the accuracy of the read-in process.

2. Model for Speech Production

The model of speech production that we have used in the present experiment is based on the acoustical theory of the vocal tract. The speech wave is generated by excitation of the vocal tract by one or more sources. The acoustical properties of the vocal tract are described in terms of a transfer function T(s), which we define as the ratio of the transform of the sound pressure P(s) measured at some distance from the speaker's lips to the transform S(s) of the source velocity or pressure. Thus P(s) = S(s) T(s). For voiced sounds, the source consists of a series of pulses of air. For a particular speaker talking at a given level, the waveform of each pulse is relatively invariant, and hence the spectrum envelope of the source is probably not dependent upon the configuration of the vocal tract. For many consonant sounds, the source is noise-like or of transient character, and seems to have a relatively smooth or flat spectrum.

The transfer function T(s) is characterized by a series of poles and zeros and can be written in the form

$$T(s) = K \frac{\left(s - s_1\right)\left(s - s_1^*\right)\left(s - s_2\right)\left(s - s_2^*\right) \dots}{\left(s - s_a\right)\left(s - s_a^*\right)\left(s - s_b\right)\left(s - s_b^*\right) \dots}$$

where s_1, s_1^*, s_2, s_2^*, ... are the complex frequencies of the zeros, and s_a, s_a^*, s_b, s_b^*, ... are the complex frequencies of the poles. For vowel configurations, T(s) has only poles and no zeros. If we set $s = j\omega$ and take the real part of the logarithm of T, we find each pair of poles and zeros represented by a single additive term:

$$\log T(j\omega) = \log \frac{K_a}{\left(j\omega - s_a\right)\left(j\omega - s_a^*\right)} + \log \frac{K_b}{\left(j\omega - s_b\right)\left(j\omega - s_b^*\right)} + \dots$$

$$- \log \frac{K_1}{\left(j\omega - s_1\right)\left(j\omega - s_1^*\right)} - \log \frac{K_2}{\left(j\omega - s_2\right)\left(j\omega - s_2^*\right)} - \dots$$

Each of these terms represents a simple resonance curve, corresponding to a conjugate pair of poles in the left half of the s-plane. A curve is added if it represents a pole and subtracted if it represents a zero.

If, for the moment, we restrict our consideration to vowel sounds, then we can, according to the foregoing result, construct the spectrum envelope of a vowel by adding a group of resonance curves and a curve representing the source spectrum plus a simple radiation characteristic. Thus if a catalog of simple resonance curves is available in the model box shown in Fig. XIV-1, then we can construct from a set of three of these

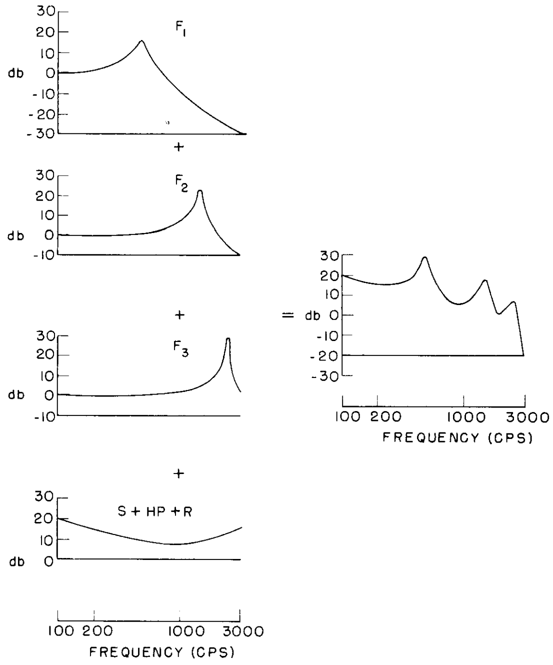


Fig. XIV-3. Illustration of method for constructing vowel spectrum envelope from four elemental spectra. Curves labeled F_1, F_2, and F_3 are simple resonance curves; the fourth curve represents source spectrum plus higher pole correction plus radiation characteristic.

curves the transfer function of a vowel up to approximately 3000 cps. In our program, we store a catalog of 24 such curves, with resonant frequencies from 150 cps to 3000 cps. If we add another curve, which represents the glottal spectrum plus the radiation characteristic plus a correction to account for the omission of higher poles, then we can construct a complete vowel spectrum envelope. The last curve added will probably be relatively invariant for a given speaker who uses a given voice effort, but may vary somewhat from speaker to speaker. To enable the model to follow this variation, we provide it with a catalog of six glottal spectra, which results in a total of 30 stored curves.

In Fig. XIV-3 three simple resonance curves labeled F_1, F_2, and F_3 are shown, together with a fourth curve representing the source spectrum (S) plus the correction for higher poles (HP) plus the radiation characteristic (R). The sum of these four curves yields the vowel spectrum envelope shown at the right of Fig. XIV-3. In effect, we are generating a vowel spectrum envelope by selecting four numbers; three of these define the resonance curves or formant frequencies, and one is a property of the particular talker and does not change very rapidly with time. For many consonants, a similar principle perhaps could be used, but then spectral zeros might have to be introduced, and different source spectra must be used. The experimental studies thus far have involved only vowel and vowel-like sounds.

3. Control and Comparator Sections

The task of the control and comparator sections of Fig. XIV-1 is to assemble spectra from the catalog of elemental curves stored in the model and to compute the error between each synthesized spectrum and the particular speech spectrum that is being examined. The aim is to determine which set of elemental curves yields the best fit with the speech spectrum. Various measures of error can be used to determine how well one curves fits another. The measure used in most of the present studies is the integral of the magnitude of the difference curve, the average difference being normalized to zero. We have also tested a variation criterion, the variation being the integral of the magnitude of the first derivative of the difference curve.

Since approximately 8×10^4 possible combinations can be constructed from the elemental spectra, and since it is not feasible to compute the error for each one of these combinations, a strategy must be devised for obtaining the best fit from tests of a much smaller number of combinations. The strategy that we have adopted is based on the finding that minimization of the criterion with respect to one variable seems to be relatively independent of the values of the other variables. In most of our initial experiments we have assumed that for each sample there is no a priori information concerning the correct formant frequencies and voicing spectrum. For this case, the

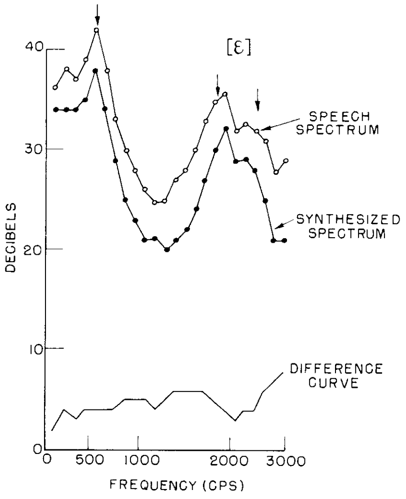


Fig. XIV-4. Upper curve represents typical speech spectrum stored in computer; curve of synthesized spectrum was generated from elemental curves by the procedure described.

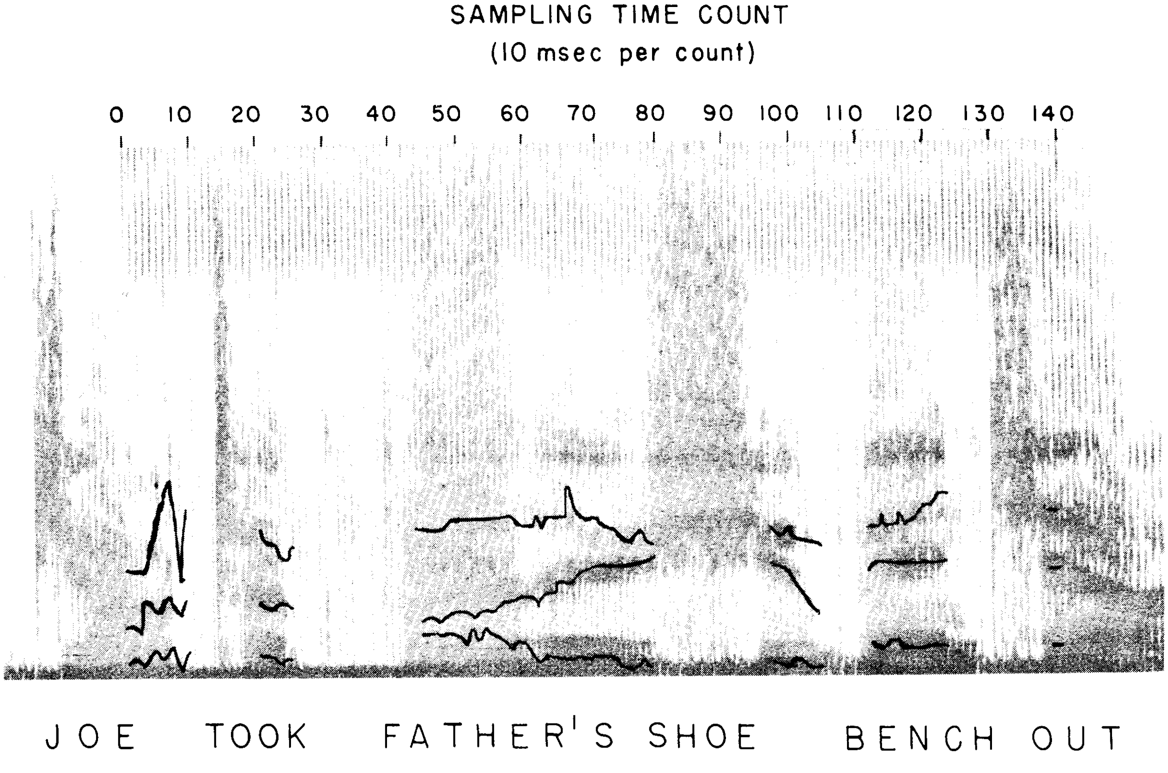


Fig. XIV-5. Spectrogram of sentence "Joe took Father's shoe bench out." Formant frequencies determined by the procedure described are shown by black lines plotted on spectrogram.

procedure that we have used for determining the component curves for a particular vowel spectrum is: (a) The lowest frequency resonance curve is selected from storage, and the error between this curve and the input spectrum is evaluated. Similar calculations are made for successively higher resonance curves up to 850 cps, and the particular curve (which we label F_1) that yields the smallest error is selected. (b) A second elemental resonance curve at 550 cps is added to curve F_1, and an error is again computed. Similar calculations are made for resonance curves up to 3000 cps, and the curve (F_2) that yields the smallest error is selected. (c) With curves F_1 and F_2 fixed, step (b) is repeated, a third minimum error is found, and curve F_3 is selected. (d) The same procedure is followed to determine the glottal spectrum (GS). (e) With curves F_2, F_3, and GS fixed at the values given by steps (a)-(d), step (a) is repeated, and a revised value of F_1 is obtained. Similar reiterations are carried out for curves F_2, F_3, and GS. (f) Following this process, identifying numbers for curves F_1, F_2, F_3, and GS are printed out. (g) The foregoing steps are repeated for the next sampled spectrum of the input speech.

If a first approximation to the correct curves is available, say from calculations on a previous speech sample, then steps (a) through (d) can be omitted.

An example of the closeness of fit obtained is shown in Fig. XIV-4. The original speech spectrum for the vowel /ɛ/ is shown, together with the synthesized spectrum that yielded the smallest error by the procedure described. The difference curve is also shown. The arrows indicate the resonant frequencies for the curves finally selected. The spectrogram of Fig. XIV-5 for a typical speech sample shows the accuracy with which the formant curves are matched during the vowel portions of the speech. The black lines mark the formant frequencies selected by the computation process that has been described. When these data were obtained, the system was operated in a mode in which each spectrum was examined independently, without using data from the previous one as a first approximation.

A method for evaluating the performance of the system in a quantitative way has not yet been developed, but consideration is being given to this problem. Work on the general procedure is continuing in an effort to improve the accuracy of formant-following for vowels and to extend the method to consonant sounds.

C. G. Bell, J. M. Heinz, G. Rosen, K. N. Stevens

B. PERCEPTION OF SPEECH-LIKE SOUNDS*

Two experimental methods for studying the perception of speech-like sounds have been investigated recently. In one of these, the process whereby subjects acquire the

*This work was supported in part by National Science Foundation.

ability to categorize the members of multidimensional auditory displays is examined. The other method deals with the perception of stimuli that are characterized by rapidly changing formant patterns.

1. The Learning of Multidimensional Auditory Displays

In a series of experiments on the information of multidimensional auditory displays, Pollack (1) has evaluated the information transmission in an experimental situation in which subjects, after a period of learning, are required to identify the members of such displays. He has shown that the information transmission is a function of the number of dimensions and of the fineness with which these dimensions are subdivided. The present experiments utilize stimuli that are more speech-like than those of Pollack, and examine the performance of the subjects during the time they are learning to categorize the stimuli and to associate them with a set of buttons on a response box.

In our present experiments the number of stimuli in the ensemble is always eight. A typical stimulus is described by the schematic patterns shown in Fig. XIV-6. It consists of an initial one-formant vowel-like portion of fixed intensity, duration, fundamental frequency (125 cps), and frequency position (300 cps), followed by a gap of duration T, followed by a burst of noise of intensity I whose energy is concentrated at frequency F. The total length of the stimulus is fixed, and hence the duration of the noise burst decreases as T increases. The variables in the experiments are T, I, and F. Seven different stimulus ensembles are studied in seven experiments, including one-dimensional ensembles in which only one variable is changed, two-dimensional ensembles in which two of the variables are involved, and three-dimensional

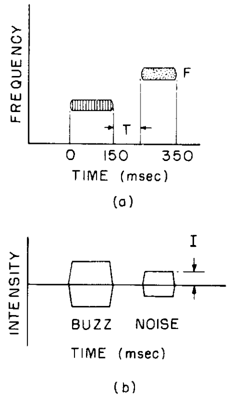


Fig. XIV-6. Description of stimuli used in learning experiment. (a) Schematic intensity-frequency-time pattern showing an initial buzz portion, a gap of duration T, and a final noise portion centered at frequency F. (b) Envelope of a typical stimulus.

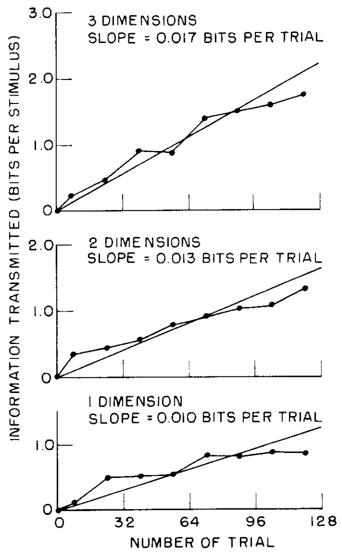


Fig. XIV-7. Learning curves associated with one-, two-, and three-dimensional auditory displays of the type shown in Fig. XIV-6. Average data for three subjects.

ensembles in which each of the three variables assumes two different values.

The stimuli are presented in quasi-random order to the subjects, who are asked to identify each stimulus by pressing one of eight buttons on a response box. After the subject makes each response and before the next stimulus is presented, an indicator light on his box correctly identifies the stimulus for him. The experiment proceeds until the 128 responses have been made. The order of stimulus presentation is adjusted so that each stimulus occurs twice in successive blocks of 16 presentations.

Typical average results for three subjects are shown in Fig. XIV-7. The information transmitted per stimulus is plotted as a function of the number of trials for the one-, two-, and three-dimensional ensembles. The learning curves plotted in this way can be fitted approximately by straight lines. As would be expected, the data show that the rate at which the ensembles are learned is highest for the three-dimensional ensemble and lowest for the one-dimensional case. Similar learning curves can be plotted for the individual dimensions, and they provide a quantitative comparison between the rates of learning for the different dimensions.

Implications of these experimental data for the study of the perception of speech will be discussed after more experimental data have been obtained.

2. Perception of Time-Variant Formant Patterns

For this experiment the stimuli, which are shown schematically in Fig. XIV-8, are generated by repetitive impulsive excitation (at 125 cps) of a tuned circuit whose resonant frequency is electronically tunable. The reason for our interest

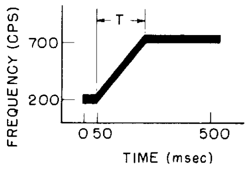


Fig. XIV-8. Schematic intensity-frequency-time pattern of typical stimulus used in ABX categorization experiment. The variable in the experiment is the duration T of the time-variant portion of the stimulus.

in stimuli of this type stems from the fact that many speech sounds are characterized by moving vocal-tract resonances or formants. Formant motions that take place relatively slowly, say during a time interval of 200 msec, are observed in spectrograms of diphthongs; faster formant motions occur for glides such as /w/ and /j/; and still more rapid changes characterize the formant transitions between consonants and vowels. The study of stimuli of this type, therefore, may yield some insight into the perceptual correlates of the consonant-vowel distinction in speech.

In the experiments described here, the resonant frequency of the tuned circuit is moved in a piecewise-linear fashion from 200 cps to 700 cps in time T, as shown in Fig. XIV-8. The stimuli are presented to the subjects in groups of three, in a manner similar to the sequences in an ABX discrimination experiment. In a given experiment the value of T in the first (or second) member of the group is fixed at, say, T_1, and the value T in the second (or first) member is fixed at T_2. For the third member of the group the value of T is intermediate between T_1 and T_2, that is, $T_1 \leq T_3 \leq T_2$. The subjects are required to categorize the third stimulus as "more like the first or more like the second sound" in the group of three. In a given experiment, a sequence of groups with different values of T_3 are presented, and a plot of the responses indicates the value of T_3 that bisects the range between T_1 and T_2. Several experiments with different values for the end points T_1 and T_2 have been performed.

Preliminary results show that the value of T for the stimulus judged to be equidistant from the stimuli with transition times T_1 and T_2 is the arithmetic mean of T_1 and T_2. The data suggest, therefore, that equal linear changes in the physical variable T are associated with equal distances along a psychological interval scale derived from the results, at least over a range of T from 16 msec to 400 msec. It is of interest to note that a plot of the psychological scale derived from these experimental data increases uniformly and monotonically as T is increased. This result would not be expected if the stimuli were close approximations to speech sounds. The psychological scale derived from experiments with such stimuli would be expected to show discontinuities at the boundaries between a diphthong and a glide and between a glide and a stop consonant.

In further studies we expect to examine different frequency ranges for the stimuli and to perform similar experiments with stimuli that are still closer approximations to the sounds encountered in speech.

J. B. Arnold, M. Halle, T. T. Sandel, K. N. Stevens

References

1. I. Pollack, J. Acoust. Soc. Am. 24, 745 (1952), 25, 765 (1953); I. Pollack and L. Ficks, J. Acoust. Soc. Am. 26, 155 (1954).

C. THE LOUDNESS OF SOUNDS IN THE PRESENCE OF A MASKING NOISE (1)

During the past thirty years many experiments have been performed to determine the "loudness" of various acoustic stimuli. There have been various calculation schemes for predicting the loudness of sounds, including pure tones, complex waveforms, and white noise. However, with the use of these calculation methods we could calculate only the loudness of a stimulus without background noise. Since this condition rarely occurs, work should be concentrated on determining the loudness of an acoustic stimulus in the presence of noise.

Three loudness-matching tests were conducted by using between 10 and 18 subjects who were individually instructed to listen (with earphones) first to a sound in quiet, then to the same sound in the presence of noise. Each subject was told to ignore the masking noise and adjust the sound in quiet until it was equal in loudness to the sound in the presence of noise. The sounds used were: (a) pure tones at 300 cps and 1000 cps; (b) narrow bands of noise centered at 300 cps and 1000 cps; and (c) complex noises with frequency components near 200 cps and 1600 cps.

An analysis of the test results indicates that the sound in the presence of noise is always less loud than the same sound in quiet by approximately a constant amount (in sones). This constant difference can be used in a computational procedure for determining the loudness of sounds in the presence of noise. However, more data are needed before a complete procedure can be evolved. This research might include investigation of the relation between the "constant difference" and the characteristics of the masking noise.

K. S. Pearsons

References

1. This report is a summary of an S.M. thesis submitted by K. S. Pearsons to the Department of Electrical Engineering, M.I.T., June 1959.