XI. COGNITIVE INFORMATION PROCESSING*

Academic and Research Staff

Prof. M. Eden
Prof. J. Allen
Prof. B. A. Blesser
Prof. T. S. Huang
Prof. F. F. Lee
Prof. S. J. Mason
Prof. W. F. Schreiber
Prof. D. E. Troxel
Prof. I. T. Young
Dr. R. R. Archer
Dr. J. E. Green
Dr. K. R. Ingham
Dr. D. M. Ozonoff
Dr. O. J. Tretiak
F. X. Carroll
D. A. Fay
Deborah A. Finkel
C. L. Fontaine
L. Hatfield
E. R. Jensen
Christine C. Nora
E. Peper
J. M. Sachs
Sandra A. Sommers
J. S. Ventura

Graduate Students

B. S. Barbay
T. P. Barnwell III
W. L. Bass
T. R. Bourk
J. E. Bowie
B. E. Boyle
R. L. Brenner
S. A. Ellias
J. R. Ellis, Jr.
A. E. Filip
A. C. Goldstein
R. E. Greenwood
W. B. Grossman
E. G. Guttmann
D. W. Hartman
L. P. A. Henckels
P. D. Henshaw
M. Hubelbank
T. Kitamura
J. W. Klovstad
H. S. Magnuski
J. I. Makhoul
G. P. Marston
G. G. Matison
G. F. Pfister
D. S. Prerau
N. Rashid
R. L. Rees
G. M. Robbins
R. D. Shapiro
R. Singer
D. G. Sitler
A. A. Smith
R. D. Solomon
W. W. Stallings
K. P. Wacks
T. R. Willemain
Y. D. Willems

A. ESTIMATION OF THE IMPULSE RESPONSE OF IMAGE-DEGRADING SYSTEMS

1. Introduction

An image-degrading process can often be modeled as passing the picture through a linear, spatially invariant system. For such cases, the received picture, $r(x, y)$, is simply the convolution of the original picture, $s(x, y)$, and the impulse response of the degrading system, $h(x, y)$.

$$r(x, y) = s(x, y) * h(x, y).$$

If the system impulse response is known, the original image can be recovered by passing the received picture through an inverse filter:

$$\underset{\sim}{s}(x, y) = s(x, y) * h(x, y) * h^{-1}(x, y),$$

*This work was supported principally by the National Institutes of Health (Grants 5 PO1 GM14940-04 and 5 PO1 GM15006-03), and in part by the Joint Services Electronics Programs (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract DA 28-043-AMC-02536(E).

where

$$h(x, y) * h^{-1}(x, y) = u_o(x, y).$$

For example, in the case of linear motion (blurring) in the x direction,

$$h(x, y) = \begin{cases} \frac{1}{vT} & 0 \leq x \leq vT \\ 0 & \text{elsewhere} \end{cases}$$

where v is velocity, and T is exposure time. The inverse filter for this system may now be computed and used to deblur the received picture. The problem that is being considered differs from the example just given, in that no a priori knowledge of the impulse response is available. In the rest of this report we shall describe a procedure that was developed to estimate the impulse response.

2. Generalized Linear Filtering

In a sense, the problem is similar to that of classical estimation theory in which we estimate a signal that is corrupted by additive noise; in this case, the signal, h(x, y), is corrupted by "convolutional noise." By using generalized linear filtering techniques,[1,2] the convolutional noise is mapped into an additive noise component so that it now becomes possible to use any of the methods of classical estimation theory to estimate h(x, y).[3] The over-all system is shown in Fig. XI-1. D_x and D_x^{-1} are called the characteristic and the inverse characteristic systems, respectively. For the deconvolution

Fig. XI-1. Generalized linear filter.

problem, they may be realized as shown in Fig. XI-2. ZT and IZT represent the z transform and the inverse z transform, respectively. The z transform was used because of the discrete nature of the signals, a consequence of processing on a digital computer. There are certain issues associated with the nonanalytic nature of the complex logarithm used in D_*.[1,2] These require that IMAG [log X(z)] = ANG [X(z)] be (i) continuous, (ii) odd, and (iii) periodic. For the example included here no effort has been made to insure the continuity of ANG [X(z)].

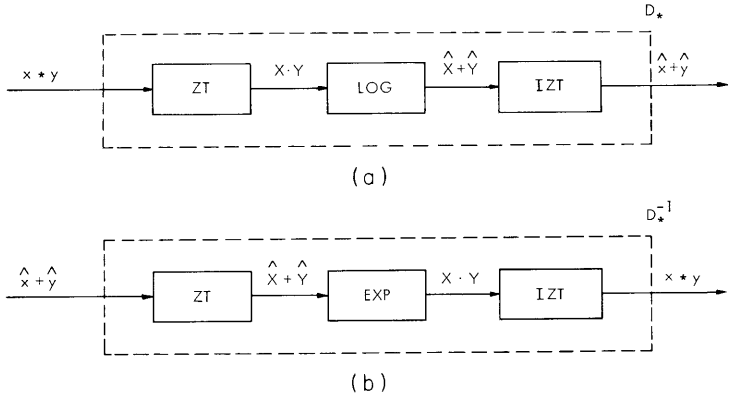


Fig. XI-2. (a) Characteristic system, D_*.
(b) Inverse characteristic system, D_*^{-1}.

Having mapped the convolved signals into additive signals, we then passed the result through a system that extracts an estimate of the desired component. The estimate is then inverse-mapped by D_*^{-1}.

3. The Linear Estimator

The estimation algorithm used in the linear processor was derived heuristically. Assume that s(n) is divided into M sections, $s_i(n)$, each of which is N points long as in Fig. XI-3. (For notational convenience, the functions are depicted as one-dimensional functions.) Thus

$$s(n) = \sum_{i=0}^{M-1} s_i(n)$$

$$s_i(n) = \begin{cases} s(n) & iN \leq n < (i+1)N \\ 0 & \text{elsewhere} \end{cases}$$

$$s(n) * h(n) = \left(\sum_{i=0}^{M-1} s_i(n) \right) * h(n)$$

$$\approx \sum_{i=0}^{M-1} (s_i(n) * h(n)).$$

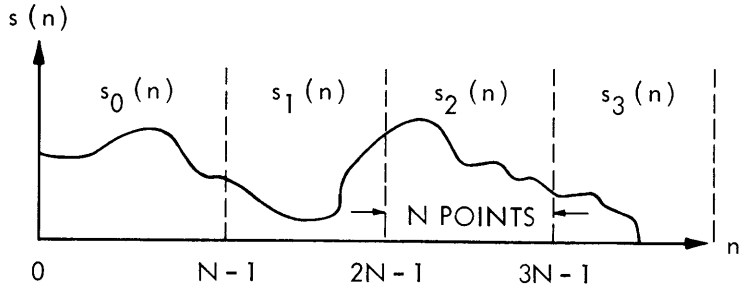


Fig. XI-3. s(n) divided into M sections, each N points long.

The approximation is valid only if the duration of the impulse response is much shorter than the section length, N. Under the assumption that this condition is met, the output of D_* is

$$\hat{s}_i(n) + \hat{h}(n) \qquad i = 0, 1, \ldots M-1.$$

The filter then calculates the average output for the M sections, reasoning that the additive noise, $\hat{s}_i(n)$, will average to zero (or a constant).

4. Experimental Results

Two examples were run to test the system. Both pictures have a raster of 512 × 512 picture elements. The number of sections used in both cases was 16 (M=16), with each section having 128 × 128 points (N = 128).

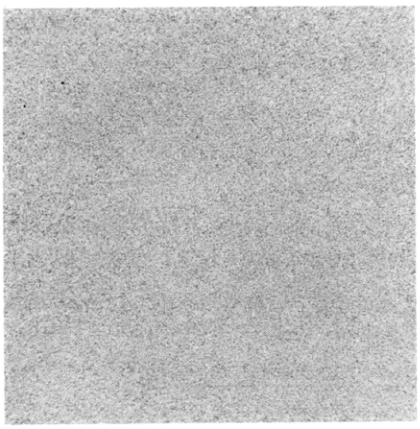
Example 1 was a computer-generated random picture having a uniform brightness distribution (Fig. XI-4a). The picture was then convolved with a 16-point long impulse response, simulating linear motion in the x direction (Fig. XI-4b). The resulting estimate of h(n, m) is shown in Fig. XI-4c.

For Example 2 we used a new crowd scene (CIPG No. 11), Fig. XI-5a, which was similarly blurred in the x direction (Fig. XI-5b). Figure XI-5c shows the result of the estimation procedure.

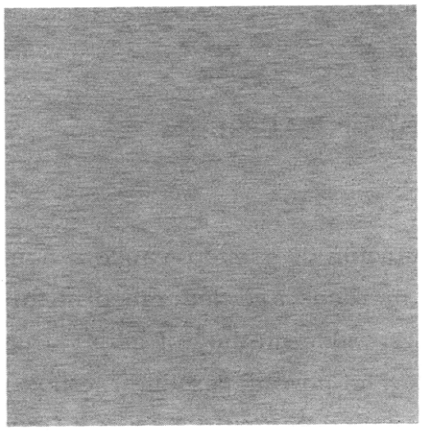
Note that in both cases, the impulse response is 1/8 of the section size. The estimate of h(n, m), using the random picture, is significantly better than that using the crowd scene. The reason for this is that the crowd scene has a relatively high correlation between sections.

5. Conclusion

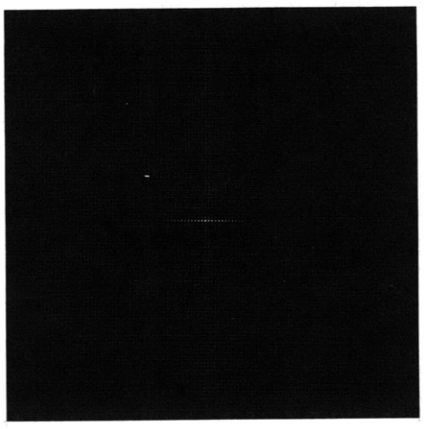
While results, thus far, look promising, more work must be done. A routine to satisfy the continuity requirement on ANG [X(z)] is now being debugged. More



(a)



(b)



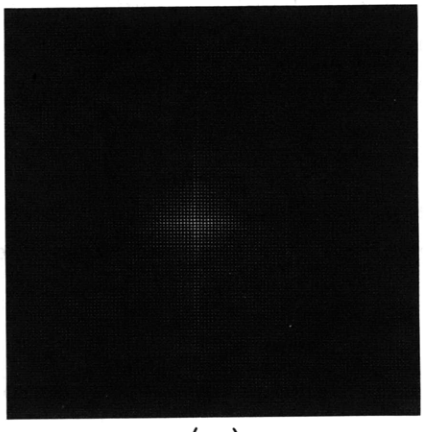
(c)

Fig. XI-4.

Random picture. (a) Uniform brightness distribution. (b) 16-point blur in the x direction. (c) Estimate of h(n, m).

(a)

(b)



(c)

Fig. XI-5.

(a) Crowd scene. (b) 16-point blur in the x direction. (c) Estimate of h(n, m).

sophisticated estimation procedures will be tried, and also the inverse filtering must be performed as a final test of the utility of the system.

A. E. Filip

References

1. A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, "Nonlinear Filtering of Multiplied and Convolved Signals," Proc. IEEE 56, 1264-1291 (1968).
2. R. W. Schafer, "Echo Removal by Discrete Generalized Linear Filtering," Ph. D. Thesis, Department of Electrical Engineering, M. I. T., February 1968.
3. H. L. Van Trees, Detection, Estimation and Modulation Theory, Vol. I (John Wiley and Sons, Inc., New York, 1968).
4. C. Rader and B. Gold, Digital Processing of Signals (McGraw-Hill Book Co., New York, 1969).

B. GENERATING STATISTICALLY INDEPENDENT GAUSSIAN PSEUDO-RANDOM NUMBERS

There are several ways of generating a population of pseudo-random numbers with a Gaussian probability density function. Three methods are described in this report, all of which start with a uniform random variable. One method of generating a uniform random variable is the power-residue method.[1] The power-residue method works in the following manner. Select a starting integer a and compute $z = a \cdot x \pmod{y}$, where x and y are appropriate integers. Choose some function of z as the random variable. To compute the next element of the population, let z be the starting number and repeat the process. If the original starting number a is an odd number, the multiplier $x = 8I \pm 3$, where I is an integer, and $y = 2^n$, then this process is periodic[1] with period 2^{n-2}.

It is possible to transform a uniform random variable into a Gaussian random variable directly. For example, the random variable

$$y = \text{erf}^{-1}(x)$$

where erf^{-1} denotes the inverse error function, is Gaussian if x is uniform over the interval (0, 1).

Second, the Central Limit theorem states that the random variable

$$y = \sum_{j=1}^{N} x_j$$

is Gaussian if the x_j are independent uniform random variables over the same range, and if N is sufficiently large. This is called the sum method.

Third, the random variables

$$y_1 = \sqrt{-\log (x_1)}\ \sin (2\pi x_2)$$

$$y_2 = \sqrt{-\log (x_1)}\ \cos (2\pi x_2)$$

are independent Gaussian random variables[2] if x_1 and x_2 are independent uniform random variables over the interval (0, 1). This is called the Chartres method.

The first method is inferior to the third, since it is computationally more involved. For example, the Taylor series for $\text{erf}^{-1}(x)$ converges very slowly if x is near 0 or 1.

$$\text{erf}^{-1}(x) = \sum_{i=0}^{\infty} \frac{f_i(2i)}{2i+1} (2\pi)^{i+1/2} \left(x - \frac{1}{2}\right)^{2i+1}$$

where

$$f_0(n) = 1$$

and

$$f_i(n) = \sum_{j=0}^{n-2i} \left[\left(\frac{n+1-2i-j}{n-j}\right) f_{i-1}(n-1-j)\right].$$

Similarly, the approximation[3]

$$\text{erf}^{-1}(x) = \eta - \frac{2.515517 + .802853\eta + .010328\eta^2}{1 + 1.432788\eta + .188269\eta^2 + .001308\eta^3},$$

where

$$\eta = \sqrt{\ln\left(\frac{1}{x^2}\right)},$$

also requires lengthy calculation.

In order to compare the second and third methods, three tests were conducted on sets of numbers produced by each of the two methods. In these tests the parameters used in the power-residue method were selected for ease of computation, and were $y = 2^n = 2^{36}$ and $x = 2^{18} + 3 = 262,147$. The multiplier x was chosen near $\sqrt{y}$ as suggested in reference 1. The period of repetition was 2^{34}; in each test only a small fraction of a period was used. The random number was selected as the eight high-order bits of z, and the random variable ranged from -128 to 127. Only y_1 was used in the Chartres method, and the two values N = 12 and N = 21 were used in the sum method.

First, 512 populations of $2^{14} = 16,384$ random numbers were generated by each method. Each population from the sum method was compared against a normal distribution with a chi-square test, and similarly against the theoretic distribution of the sum of N independent uniform random variables. Each population from the Chartres method was compared against a normal distribution with a chi-square test. The resultant value from each chi-square test was converted to a point, z, on a standardized normal curve using the approximation[4]

$$z = \sqrt{2\chi^2} - \sqrt{2(\text{d.f.}) - 1}.$$

Figure XI-6 shows a histogram of the values of z for each of the five cases. Comparison of Fig. XI-6a and XI-6b shows that the medians of the two histograms are separated by approximately .2 standard deviations. This indicates that the theoretic distribution of the sum of 12 independent uniform random variables is a much better fit to the population of this process than the Gaussian distribution. Comparison of Fig. XI-6c and XI-6d shows that the medians of the two curves are about equal. This indicates that the theoretic distribution of the sum of 21 independent uniform random variables is quite similar to that of Gaussian distribution. Finally, comparison of Fig. XI-6c, and XI-6e shows that the median of the histogram of the Chartres method is at least approximately .1 standard deviation lower than that of either of the sum methods.

These three results indicate that the Chartres method produces a more Gaussianlike population than the sum method. In order to produce a given number of populations that pass a chi-square test at the α% confidence level, more attempts (on the average) must be taken when using either of the sum methods than with the Chartres method. For α = 90-95% level, the difference is approximately 20-50% more attempts; for $\alpha \geq 99\%$ level, the difference is approximately twice as many attempts or more.

Second, in order to test the statistical independence of the samples, each method generated 512 sequences of $2^7 = 128$ random numbers. The power density spectrum of each sequence was computed with the aid of the Fast Fourier transform.[5,6] Figure XI-7 shows the sum of the spectra of each set of 512 sequences as normalized dimensionless quantities. If the random process were to generate statistically independent samples, then the process would be white noise, and the spectrum at any one frequency would be of chi-square distribution with 1023 degrees of freedom, except for DC and the 64^{th} frequency, which would have only 511 degrees of freedom. The distribution across the 65 frequencies was similar to such a chi-square distribution in every case tested according to a Kolmogorov-Smirnov test at the 10% confidence level.

The third test was performed to compare the running times that are necessary for each of the two methods. The sum method required ~440 μs to generate a Gaussian

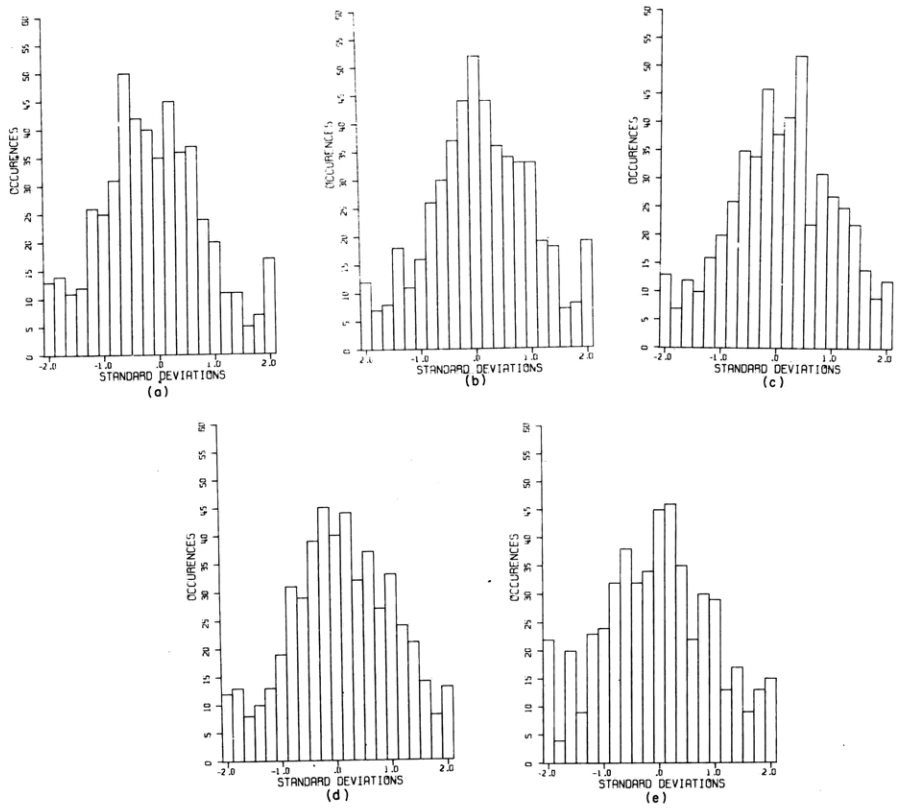


Fig. XI-6. Histograms of the results of chi-square tests converted to a point on a zero-mean, unit standard deviation normal curve. All points that are more than two standard deviations from the mean are shown as two standard deviations from the mean. (a) Sum method, N = 12, populations compared with normal distribution, median = .07. (b) Sum method, N = 12, populations compared with distribution of sum of 12 independent uniform random variables, median = -.14. (c) Sum method, N = 21, populations compared with normal distribution, median = .10. (d) Sum method, N = 21, populations compared with distribution of sum of 21 independent uniform random variables, median = .09. (e) Chartres method, populations compared with normal distribution, median = -.02.

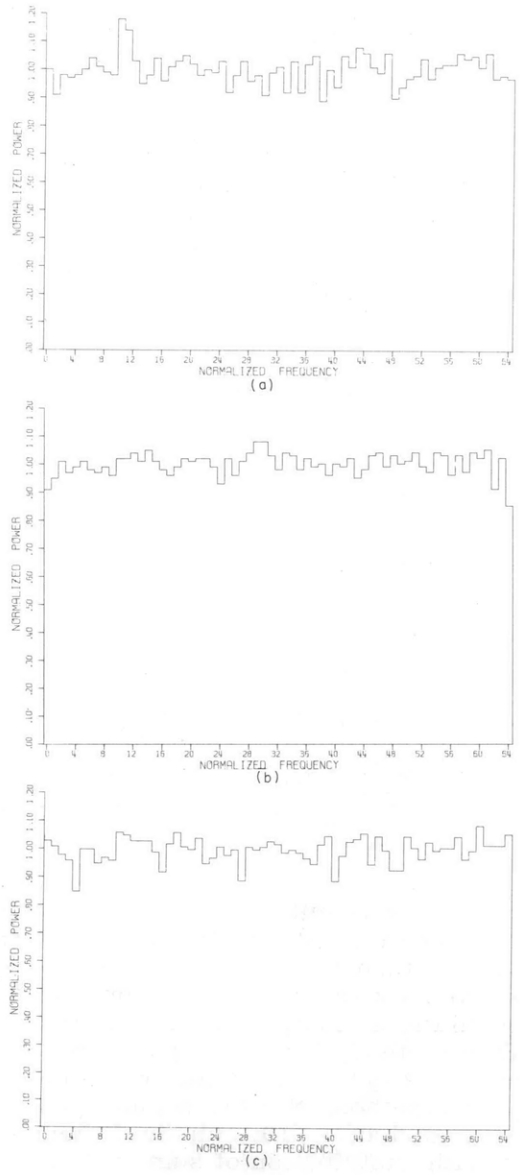


Fig. XI-7. Normalized average spectra of populations of pseudo-random numbers. (a) Sum method, N = 12. (b) Sum method, N = 21. (c) Chartres method.

number with N = 12, and 740 μs with N = 21; and the Chartres method, ~340 μs. Only the first number was used in the test with the Chartres method because there was some concern about the independence of the Gaussian numbers. This concern was unnecessary in principle, and probably in practice also. If both numbers had been used, approximately 190 μs would be required to generate each Gaussian number. The computer was PDP-9 with a 1-μs cycle time and an EAE unit (hardware multiply), but this feature was relatively unimportant.

In conclusion, it seems that the Chartres method is a relatively fast and simple procedure for generating statistically independent Gaussian pseudo-random numbers.

I would like to thank John Jagodnik for his aid with Calcomp subroutines that were used in preparing the figures. John Doles and Steven Robbins read the manuscript and made helpful comments and suggestions.

R. E. Greenwood

References

1. I. B. M. Publication C20-8011 (1959); see references within it.
2. B. A. Chartres, Technical Report 5, Brown University, 1959.
3. C. Hastings, Approximations for Digital Computers (Princeton University Press, Princeton, N. J., 1955).
4. B. W. Lindgren and G. W. McElrath, Introduction to Probability and Statistics (Mac Millan Company, New York, 1959), p. 192.
5. J. W. Cooley and J. W. Tukey, Math. Comp. 19, 297 (1965).
6. G-AE Subcommittee on Measurement Concepts, IEEE Trans. Audio and Elec., Vol. AU-15, p. 45, 1967.

C. TEXT-TO-SPEECH CONVERSION

For some time we have been working on those aspects of speech synthesis that have direct application to the development of a reading machine for the blind. While this task still provides a large measure of motivation, the emphasis has recently shifted in the direction of broader contexts. Currently, we view our efforts as directed toward the general problem of text-to-speech conversion with particular attention being paid to the consideration of engineering techniques for such conversion as a model for the cognitive aspects of reading. In this way, we are studying the fundamentals of an important transformation between language representations while also investigating human reading behavior, which provides the performance standard that we are trying to achieve algorithmically. In this report, we set out our goals and line of investigation, together with a summary of present progress. Future reports will give details of particular aspects of the research.

It is important to establish initially the goals for converting text to speech. Any (English, in this case) text should be allowable as input, although we have not emphasized conversational dialogues and poetry. This means that all of the words of a language must be recognized. Not only is there a large number of such words, but they are constantly changing with time. Thus "earthrise" has been recently coined. To be useful, speech must be generated in real time, by which we mean a comfortable reading rate of approximately 150 words per minute. Since storage is expensive, it is desirable to perform the task with a minimum amount of data, and instead, to use rules of the language to derive the speech output control parameters. Finally, the output speech must be not only intelligible but also sufficiently natural to permit long-term use.

These requirements cannot be met by the use of stored recordings of words, even in encoded form, and it appears to be necessary to first derive the phonetic specification of a sentence from its underlying linguistic form, and then convert this description into the actual acoustic waveform corresponding to the speech. These demands are not easy to meet, but we shall present the form of a solution, and report on the present level of progress.

Assuming that we are given sentences as strings of words, which in turn are composed of strings of letters, we have to somehow relate this information to a stream of speech sounds. We first note that written words are not just random strings of letters, but that they have an internal structure made up of parts, which we can think of as atomic, since they are often the minimum units of grammatical form. This structure arises in two ways. First, many words are formed from root words by adding prefixes and suffixes, such as "engulf, books, miniskirt, finalize, restarted," etc. Second, and this is very free in English, two roots may be concatenated together to form a compound word, as in the previously mentioned "earthrise" and "handcuff, bookcase, outfit," etc.

In view of this internal structure present in words, it is wasteful to store speech representations for all inflected and compound words of the language. Instead, an algorithm is used to decompose words to prefixes, suffixes, and roots by performing a longest-match-first search from the end of the word in an attempt to isolate the constituents of the word. The algorithm is also able to compensate for consonant-doubling effects, as in "bidding." Recently, we have added to the power of the algorithm by recognizing two forms of "pseudo-roots." One is the so-called functional root, such as "pel" in "impel, repel, dispel, compel, propel," where "pel" clearly does not occur alone. The other case arises with suffixes such as "-ate," and "-ation" which can form (among others) "agitate" and "agitation", but "agit" may not exist alone. To remove this apparent redundancy, only "agitate" need be stored, and detection of "-ation" in "agitation" automatically causes a search for the same residual root (here "agit") which ends with "-ate."

By means of such decomposition techniques, an order-of-magnitude reduction in the number of words that have to be stored can be realized. At present, an interactive version of the decomposition algorithm has been coded. Starting with a "base" dictionary of roots, prefixes, and suffixes, the entire Brown Corpus[1] of approximately 50,000 words will be decomposed into the nascent lexicon. This is achieved by first sorting the Brown Corpus by length of word, shortest words first. Each word of the Corpus is then presented to the decomposition algorithm and the lexicon existing at that moment. If decomposition succeeds, there is nothing to do, but if it fails, a decision must be made about whether the word should be added to the lexicon. The new lexicon is sorted, and the next Corpus word is presented for decomposition. Clearly, in the initial stages of this process, most Corpus words will be added to the lexicon, but as this growing dictionary embraces more of the short high-frequency roots, fewer words will be added. There will still be, however, a large number of roots in the lexicon, and yet some new words may not decompose into dictionary entries. Additionally, decomposition must occasionally be blocked, to prevent such mistakes as "scarcity" → "scar" + "city."

In order to cope with new words, and to minimize the required dictionary size, we have developed a set of letter(s)-to-sound rules that allow the calculation of a word's pronunciation directly from its letters. Since English spelling is notoriously inconsistent, it might seem that this method would be doomed to failure, but it turns out that most of the irregular words are of high frequency, which warrants their inclusion in the dictionary. For example, "f" is completely regular in the 20,000 most frequent words of English except for the single word "of". Thus the dictionary includes the "closed" word classes (articles, auxiliaries, prepositions, pronouns, and conjunctions), prefixes and suffixes, functional roots, high-frequency monosyllabic words, and phonically irregular words and exceptions. The philosophy underlying this choice is that the

size of the dictionary should be minimized and that the sound specification should be computed by rule when possible. We have developed a complete algorithm for consonant conversion. It should be mentioned, however, that by far the most difficult problems with letter-to-sound correspondences are the medial vowel digraphs, such as "ea" in "reach, leather, steak, reality." Decomposition removes, however, many occurrences such as in "reagent" and "changeable." The "decomposed" Brown Corpus word list will be used to study these letter groups.

Once words are either decomposed into dictionary entries or converted to speech representation by letter-to-sound rules, it remains to provide a complete phonetic transcription of the sentence. Dictionary entries contain the phonemes or basic sounds of the word plus parts-of-speech information and word-level stress information. If a word has been decomposed into dictionary entries, then an algorithm computes the parts-of-speech of the word and the word-level stress. The phoneme string for the word is also obtained from those for the constituent parts. If decomposition does not occur, then the letter-to-sound rules provide only the phoneme string and some rough stress information.

Before further progress can be made toward synthesizing the speech, it is necessary to parse the sentence to reveal its constituent structure. This information is used to disambiguate stress in noun/verb ambiguities ("refuse"), allow pauses to be inserted appropriately, compute phrase-level stress contours, and derive the intonation contour of the sentence. The present parser operates very fast, uses less than 2K of computer memory, and provides a complete bracketing of the sentence, including embedding, ellipsis, and other involved transformational effects.

By means of linguistic and phonetic rules, values for vowel duration and pitch are computed, as a function of phonemic content and syntactic structure. In this way, overall sentence intonation is provided, as well as the appropriate correlates for stress. Intonation, stress, and juncture are often referred to as the prosodic features of speech, or those that extend over several segments of speech. They are necessary to provide the listener with information about the structure of the sentence, so that he will "hear" the stress and intonation correctly. Listeners can often compensate for poor segmental or phonemic information when they know the context of the entire sentence.

Once the prosodic "frame" is known, the individual sounds or phonemes must be realized. Given the name of a phoneme, a synthesis-by-rule algorithm computes the control parameters that are needed by a terminal analog synthesizer to produce the physical speech waveform. This program must compute the spectral correlates corresponding to the vocal-tract configuration used to produce the sounds of the sentence. Recently, our efforts have focused on improved synthesis of stops and fricatives. Much work remains to be done in this area, particularly in the realization of contextual effects.

While, at present, a terminal analog synthesizer is being used, a digital synthesizer would provide increased flexibility and reliability. The synthesizer is the only special-purpose piece of hardware required by the system, since all of the other algorithms can be computed by a general-purpose computer.

The process by which we have described the conversion of text to speech has been found to be fundamentally sound. Certainly, much work must be done to reveal the linguistic and phonetic regularity of our language. The present prospects are very encouraging, however, and warrant a continued strong effort in this area.

We believe that this procedure for converting text to speech may have some validity as a partial model for reading. High-frequency, phonically irregular words are "read" as chunks, and words are decomposed into roots and affixes when possible. The less frequent words are not stored in any lexicon, but converted directly to speech by rules of the language. Hence both "phonic" and "whole-word" techniques are used. An experiment is under way which seeks to discover further information about the letter-groups within words that are perceived as units. Four-letter monosyllabic "words" such as "flon" and "Wtad" have been selected which contain consonant clusters that are either legitimate or not English. Tachistoscopic presentation will then be used and subjects will be asked to spell the "word." Error analysis of the results should reveal whether or not legal consonant clusters function as perceptual units.

J. Allen

References

1. H. Kucera and W. N. Francis, Computational Analysis of Present-Day American English (Brown University Press, Providence, R.I., 1967).

D. SOME SYLLABIC JUNCTURAL EFFECTS IN ENGLISH

The problem of defining a syllable in general is probably an impossible problem. The reason for this is that the syllable may have a definition on at least three different functional levels. First, it may be defined graphemically, so that printed words may only be broken at the end of a line at a "syllable juncture." Second, it may be defined morphemically, and the syllable boundaries may be placed at morph boundaries. Third, it might be defined acoustically, with syllable boundaries placed at locations of acoustical significance, that is, the "spoken" syllable boundary. Clearly, these three levels of definition are somewhat related, but it is not exactly clear what this relationship might be.

The problem that this report seeks to address is that of the effect, if any, of syllable junctures on the acoustic correlate of segment duration. Hence interest must lie in the third kind of syllable, the "acoustic" syllable. The hypothesis of its existence must be somewhat like this: There are many different-sized segments that are identifiable in

spoken speech. These include the phoneme (functional segment), the morph, the word, the phrase (derived constituent), the sentence, and so forth. There is no question that subjects do not have trouble identifying word boundaries in the sentences that they perceive; likewise, they can identify syllables. This does not mean that there is universal agreement among subjects about the precise location of syllable boundaries; certainly, there is not, but it may be said that, for all practical purposes, there is universal agreement about the number of syllables perceived and which vowels make up each syllable nucleus. This, of course, could be an entirely perceptual phenomenon. It could also be due to the subject's identification of sequential vowels, and assigning one syllable to each. But, it could also be due to the direct interpretation of some acoustical correlate for syllabification. Certainly, it is not unreasonable to expect, inasmuch as subjects recognize the existence of syllables in spoken speech, thay they might well have some direct acoustic cue for syllables in their speech.

It would be naive to think that we must choose only one of these three possible explanations for the perception of syllables. As in most speech phenomena, it is undoubtedly true that some combination effect leads to the perceptual result, and, in many cases, the direct acoustic cues for syllables may be optional. If, however, there are cases in which the syllabic acoustical phenomenon is not optional, clearly these cases should be studied.

The present study was concerned with three elements of the syllable problem. The first of these was the effect of a syllable boundary on vowel duration. In particular, the author has shown that vowel duration could be considered as a function of structure and phonemic context.[1] The phonemic context really is simply that the duration of a particular vowel is greatly influenced by the following consonant. The hypothesis of the present study is that if a syllable boundary exists between a vowel and the following consonant, then the effect of that consonant on the duration of that vowel would be greatly reduced. If this hypothesis should prove to be true, it would help to explain certain previous experimental results.[1]

The second element of interest lies in a general phenomenon reported in two studies.[2, 3] These authors sought to investigate syllabic junctural effects by studying words and phrases that differed only in their word or syllable junctures (night rate vs nitrate). These authors claimed that, generally, phonemes tended to be lengthened toward the beginning of a syllable and shortened toward the end. It should be noted, however, that these studies did not succeed in their attempted format, for, in general, not only did they vary syllabic junctures but also derived constituent structures (stress), and, as has been seen previously, stress can be a determining factor in duration. Hence we desired to check whether the phenomenon so reported is really general.

The last element of this study was to try to throw some light on the problem of the definition of "syllable" in an acoustic context. In particular, we desired to study how

a subject varied his acoustic correlates depending on where he thought the syllable juncture should be placed. This problem will be discussed in more detail as the experiments are described.

1. Experiment

The experiment was divided into two parts. In the first part, subjects were asked to read single words into a tape recorder microphone. The words were written on flash cards, one on each card, and held up individually for each subject to read. In the second part, subjects were given a list of the words they had just read from the flash cards. They were asked to "mark the place in each word where you think you say the syllable boundary." After they had completed this task, the subjects were asked to read the words again, this time from the list they had just marked. In both parts of the experiment, spectrograms were made of each test word on a Kay Sonograph, and the durations of each phoneme in each word were measured and tabulated.

The list of words used in the experiment is shown in Table XI-1. The main purpose of this experiment was to check whether a vowel's duration was less affected by the following consonant if that vowel were separated from that consonant by a syllable boundary. Hence an experimental environment that allowed variations in syllable boundary position and following consonant for the same vowel and the same stress configuration was necessary. Therefore we decided to use only two syllable words having stress on the first syllable and the same vowel, /i/, as the nucleus of their syllable. Likewise, these words were chosen so that the /i/ in their first syllable was always followed by a stop consonant, either voiced or unvoiced. The reason for this was that previous work had shown that as far as their durational effect on the preceding vowel is concerned, the voiced-stop consonants could be considered as a group and the unvoiced stop consonants could also be considered as a group.[1]

As we have stated, there is no general agreement about where the syllable boundaries lie in many words. Hence the word list was divided into three groups. The first group contained words for which there was general agreement that the syllable boundary lay between the first vowel and the following consonant. The second group included words for which there was general agreement that the syllable boundary lay after the first vowel's following consonant. These two groups clearly represent the main test of the hypothesis. The third group of words comprised those for which there is no general agreement about the location of the syllable boundary. The results from this last group were intended to test whether these words could be considered acoustically part of Group #1 or Group #2, or whether they fall somewhere in between.

In the second part of the experiment we intended to see what effects a subject's knowledge that he was being tested on syllable effects, and his attempts to acoustically support his own syllabic markings, had on his previous results. It is important to

Table XI-1. Test words in syllable experiment.

Group #1 Open Syllables	Group #2 Closed Syllables	Group #3 Questionable Syllables
Decoy	feetless	beaters
Despot	meetless	beater
Beebee	seatless	beetle
Cetane	neatness	heater
Detail (noun)	needful	meeting
Detour	needless	meter
Veto	seedling	peter
Cedam	speedless	needle
Cedar	meekness	beader
Cetus	cheekness	deeded
Heclion	beakness	beaker
Pekoe	cheapness	cheeky
Sego	deepness	speaker
Cecal		feeblish
Ceacum		seepage
Deacon		peeper
Beacon		keeper
Decrease (noun)		deeply
Fecund		cheaply
Sequel		seta
Secant		feeble
Sepoy		Phoebe

understand that there was no real interest about where the subjects marked their syllable boundaries, since they were untrained subjects doing an ambiguous task and hence their results would probably not be consistent. What was of interest was the effect that these markings, whatever they happened to be, would have on the previous results.

All of these tests were taken by 10 subjects. All subjects were native speakers of English, though probably not of the same dialect. Six of the subjects were male and four were female.

2. Results for Vowels

Before presenting the results of this experiment, two general points should be made. First, in setting up the experiment, we assumed that as far as their effect on the

previous vowel's duration is concerned, unvoiced, and likewise voiced, stop consonants could be considered as a group. This assumption was tested for each of the subjects and found to be acceptable. Hence, in presenting the vowel-duration data, no differences will be noted among the stop consonants, other than the group to which they belong (voiced or unvoiced).

The results of the vowel-duration tests in the first part of the experiment for one subject are shown in Fig. XI-8. Recall that the main hypothesis was that if a syllable boundary could be said to fall between the vowel and the following consonant (Group #1),

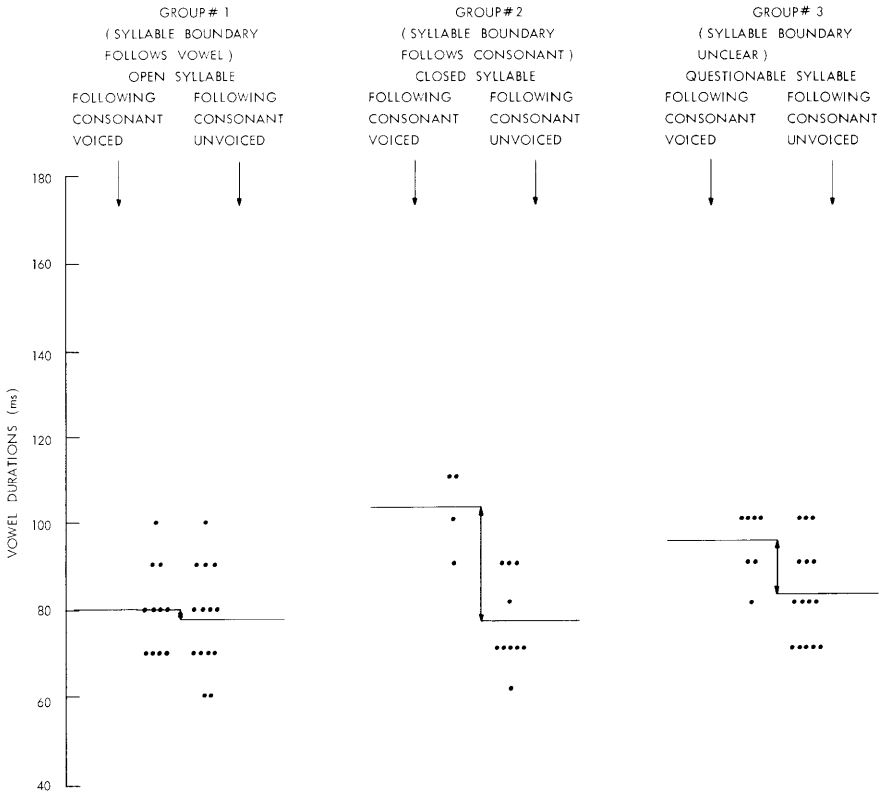


Fig. XI-8. Vowel durations for Subject 1 as a function of Group and following consonant type. Straight lines are averages.

then the effect of the following consonant would be greatly reduced; that is, the average vowel duration before voiced consonants would be almost the same as before unvoiced consonants. As can be seen from Fig. XI-8, not only is the voiced-unvoiced difference reduced in Group #1, it is almost nonexistent. But in Group #2, where the vowel and the following consonant are assumed to be in the same syllable, the voiced-unvoiced splitting effect is still very much in evidence. Hence, at least from the point of view of this subject's data, the hypothesis is strongly supported.

Now notice the results from Group #3. This is the group in which there was no general agreement about the location of the syllable boundary. One hypothesis about these data might be that each of these words, "acoustically," really belongs either to Group #1 or Group #2, and there is simply no known way to predict which. If this were the case, the points on the "voiced" side of Group #3 could be expected to break into two separate groups, one centered around the average of the voiced side of Group #1 and one around the average of the voiced side of Group #2. This is not the case, however. What is closer to being true is that there is one group of points centered somewhere in between the averages for Group #1 and Group #2. This would appear to indicate that Group #3 may represent a separate acoustic phenomenon from either Group #1 or Group #2.

Another point is that the durations of vowels before unvoiced stop consonants and the durations of vowels directly before syllable boundaries (Group #1) all seem to be about the same. Hence, for this subject at least, it would appear that following a vowel by a syllable boundary is durationally almost equivalent to following it by an unvoiced stop consonant.

Figure XI-9 shows the results for another subject. Notice that there are several

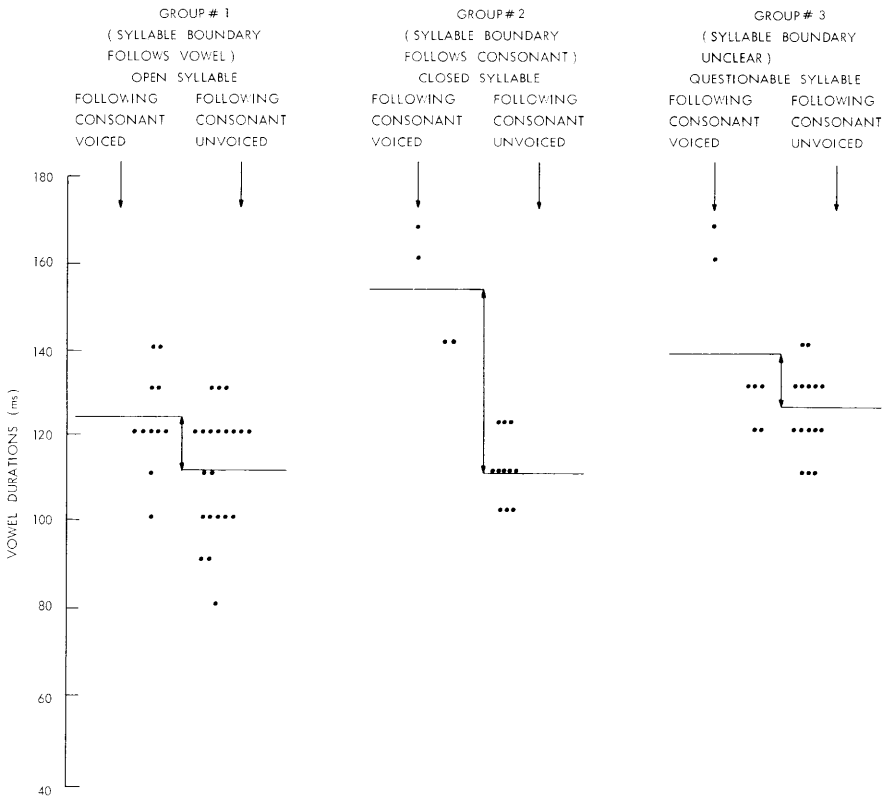


Fig. XI-9. Vowel durations for Subject 2 as a function of Group and following consonant type. Straight lines are averages.

important differences between this subject and the last subject. First, note that all of his vowel durations are longer. This is probably just a personal idiosyncrasy. Second, notice that once again the original hypothesis is strongly supported by Group #1 and Group #2, but note that, this time, there is more splitting in the Group #1 averages than before. Hence, for this subject it can be said that the syllable boundary greatly reduces the effect of the following consonant, but does not completely destroy it.

Now notice the results for Group #3. Here, as was not the case before, the Group #3 voiced section does split into two well-defined groups. Hence, for this subject it might be argued that all words really belong to either Group #1 or Group #2, but it should be noted that this was the only subject who exhibited this characteristic. It should also be noted that the average vowel duration before unvoiced stop consonants in Group #3 is longer than the average before unvoiced stop consonants in either Group #1 or Group #2. This was a characteristic exhibited by four of the ten subjects, and will be shown in the section on consonant durations to have a direct correlation with the duration of the following stop consonant.

Figure XI-10 shows the experimental result of another subject. This subject is included for three reasons. First, once again there is strong evidence in favor of the

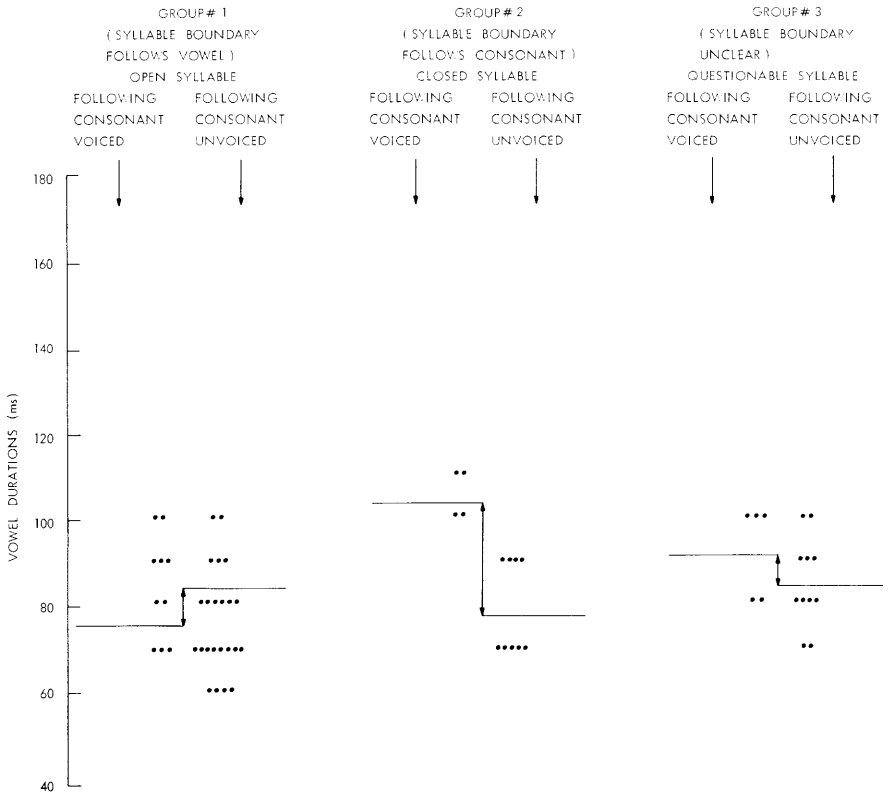


Fig. XI-10. Vowel durations for Subject 3 as a function of Group and following consonant type. Straight lines are averages.

hypothesis. Second, the general level of vowel durations is much shorter than for either of the two previous subjects, which clearly illustrates why subjects must be considered separately. Third, this subject has the same long vowel duration before unvoiced stop consonants in Group #3, as did the last subject.

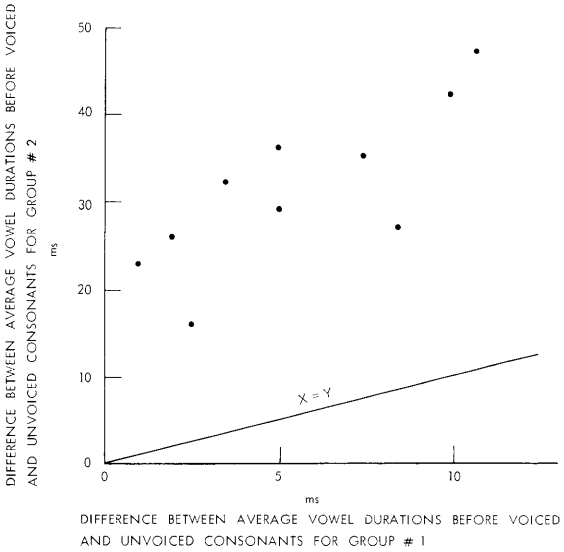


Fig. XI-11. Results for all ten subjects.

These three subjects were presented because they illustrated most of the phenomena related to a vowel duration which were observed in the first part of the experiment. Figure XI-11 is a plot of the difference in the average vowel durations between vowels followed by voiced and unvoiced stop consonants for Group #1 against the same difference for Group #2. This plot is not presented to suggest any functional relation between these two quantities, but only to show that the differences plotted for Group #2 are always greater than those for Group #1. This strongly supports the hypotheses for all ten subjects.

To summarize, it may be said that the results of this experiment all strongly support the hypothesis that the presence of a syllable boundary between a vowel and the following consonant greatly reduces the effect of that consonant on the vowel's duration. Likewise, an additional result was that, for all subjects, the average duration of a vowel before a stop consonant tended to be the same for Group #1 and Group #2.

The results for Group #3 may be said to be that in general the difference between the durations of vowels before voiced and unvoiced stop consonants fell somewhere between Group #1 and Group #2. For six of the subjects, the average duration of a vowel before an unvoiced stop consonant was very close to that for Group #1 or Group #2. For four of the subjects, this average duration was longer.

3. Results for Consonants

The general hypothesis concerning consonants, suggested by Lehiste[2] and by Hoard,[3] was that consonants near the beginning of a syllable tend to be longer than they are near the end. In this experiment, the stop consonants in Group #1 were all at the end of the first syllable, while those in Group #2 were all at the beginning of the second syllable. Hence one would expect that if the average duration for the stop consonants in Group #1 minus those in Group #2 were plotted as in Fig. XI-12 then the average results should be

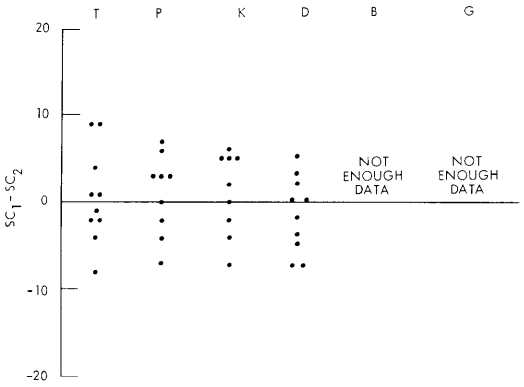


Fig. XI-12.

Average duration for stop consonant in Group #1 minus average duration for stop consonants in Group #2.

positive. It can be seen from Fig. XI-12 that this is not the case. What would appear to be true is that there is no measurable average difference between the stop consonants in Group #1 and Group #2. Hence Lehiste's and Hoard's results are not supported.

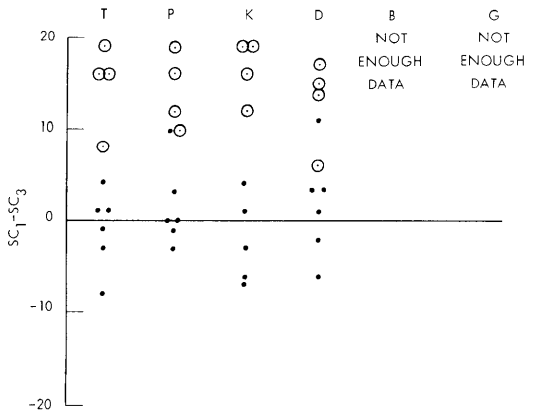


Fig. XI-13.

Average durations for stop consonants in Group #1 minus those for Group #3. Circled points are for 4 special subjects.

Figure XI-13 is a plot of the average durations for stop consonants in Group #1 minus those in Group #3. Two things are worth noting in this plot. First, for 6 of the subjects the results were essentially the same as for Group #1 and Group #2; that is, there was no measurable effect on consonant durations as a function of syllable

boundaries. Second, for four of the subjects, however, the consonant durations were quite short. It turns out that these are the same four subjects who exhibited long vowel durations for Group #3. In particular, with longer vowel durations and shorter stop-consonant durations, the durations of the combinations remain nearly constant. There is no clear explanation why this occurred, but the data are included for completeness.

4. Results for Syllable Markings

In the second part of the experiment, subjects were asked to choose where they "thought they said the syllable breaks," and were then asked to read the words again from their marked list. The idea behind this part of the experiment was not to see where subjects marked syllable breaks because, as untrained subjects with an ambiguous task, they could not be expected to perform well. For those who are interested, however, the

Table XI-2. Results of syllable placement tests.

		GROUP FOR WORD			
		GROUP #1	GROUP #2	GROUP #3	
GROUP CHOSEN BY SUBJECTS	GROUP #1	88%	8%	64%	GROUP #1 = V/C GROUP #2 = VC/ GROUP #3 = ?
	GROUP #2	12%	92%	36%	

Table XI-3. Results of syllable boundary placement tests for vowels.

		VOWEL DURATION CHOSEN		
		LONGER	SAME	SHORTER
VOWEL DURATION FOUND	LONGER	52%	36%	31%
	SAME	24%	20%	12%
	SHORTER	24%	44%	57%

SAME: Subject chose word to be a member of the same Group as experiment (within 10 ms (±5 ms)).

LONGER: Subject chose word to be a member of a Group with longer vowels than experiment; that is, member of Group #1 or Group #3 with voiced-stop consonant chosen as a member of Group #2.

SHORTER: Member of Group #2 or #3 with voiced-stop consonant chosen as a member of Group #1.

results of the syllable markings are shown in Table XI-2. The real point of this experiment was to see if subjects could be forced to vary their outputs in order to support their syllabic markings. The hypothesis for vowels was the following: If a subject chose an open or a questionable syllable as closed, we would expect his vowel duration before a voiced consonant to be longer to support this claim; likewise, if a subject chose a closed or a questionable syllable as open, then we would expect the vowel duration before a voiced consonant to be shorter. The composite results for all subjects are shown in Table XI-3.

Table XI-4. Results of syllable boundary test for stop consonants.

		STOP CONSONANTS CHOSEN		
		LONGER	SAME	SHORTER
STOP CONSONANTS FOUND	LONGER	27%	7%	12%
	SAME	56%	71%	47%
	SHORTER	17%	22%	41%

SAME: Subject chose same Group as experiment (within 10 ms (±5 ms)).

LONGER: Subject chose a member of Group #2 or #3 to be part of Group #1.

SHORTER: Subject chose member of Group #1 or #3 to be a member of Group #2.

Two points should be made concerning the results of this experiment. First, statistically, there is a noticeable tendency to uphold the hypothesis. Second the results are comparatively weak, however, and there is also a definite tendency not to change the way the word was said. This is particularly true concerning the consonants as shown in Table XI-4, where the majority of the consonant durations remain the same regardless of the syllable marking chosen by the subject. Once again, the second part of the experiment seems to support the hypothesis for vowel durations, but does not support any hypothesis for consonants.

T. P. Barnwell III

References

1. T. P. Barnwell III, "An Algorithm for Segment Durations in a Reading Machine Context," Ph. D. Thesis, M. I. T., August 1970.

2. Ilse Lehiste, "Juncture," Proc. of the Fifth International Congress of Phonetic Sciences, 1964.

3. J. E. Hoard, "Juncture and Syllable Structure in English," Phonetica, Vol. 15, pp. 96-109, 1966.

E. TACTILE PITCH FEEDBACK FOR DEAF SPEAKERS

1. Introduction

Several investigators have reported the problems of profoundly deaf speakers with pitch control. The characteristic difficulties include abnormally high average pitch (Angelocci, Kopp, and Holbrook[1]), and unnatural intonation patterns (Martony[2]). These anomalies are sufficient in themselves to make deaf speech sound unnatural and even unintelligible (Børrild[3]).

To help deaf speakers acquire better pitch control, various researchers (Dolansky, et al.,[4] Risberg,[5] Martony[2]) have devised and tested visual pitch displays. While visual displays have certain inherent advantages, they are limited to use in training sessions and necessarily interfere with lipreading by their users. No work has been done with tactile pitch displays, although this modality has a potential for continuous feedback from wearable displays and could supplement lipreading without handicapping it. Our research[6] was a pilot study, intended to explore the utility of simple pitch detectors and simple tactile displays.

2. Apparatus

The prototype pitch detector is straightforward. A throat microphone detects voiced speech; its output is amplified, lowpass-filtered, and converted to a square-wave pulse train by a Schmitt trigger. The pulse train is gated for a fixed time, and pitch frequency is determined from a zero-crossing count on the gated pulse train. The pitch measurement is quantized into one of 8 channels. The first seven channels are adjusted to correspond to the range 100-240 Hz in bandwidths of approximately 20 Hz each. The eighth channel corresponds to all pitch frequencies above 240 Hz. Counts in each channel are recorded for analysis of a speaker's pitch distribution. The speech input is sampled periodically for durations of 50, 100, 200 or 400 ms. Immediately after sampling, the pitch is displayed to the speaker for 50 ms. Thus the total cycle time from the beginning of one display to the next is variable in steps of 100, 150, 250, and 450 ms, and feedback performance as a function of display rate can be investigated. Note that the feedback is in quantized, sampled-data form.

The display is also quite simple. Solenoids poke the fingers of the speaker to provide the tactile feedback. The eventual goal of research in tactile pitch feedback is the design of a wearable speech aid, and an important criterion for such an aid is that it be simple and consequently inconspicuous to the user. Therefore, the displays that we used employed only two and three solenoid pokers. Switching circuits allow the experimenter to assign counts in any channel to any tactor. Thus, the eight channels can be grouped for display into "high," "low," and "ok" bands when three pokers are used, or into "high" and "low" bands when two are used.

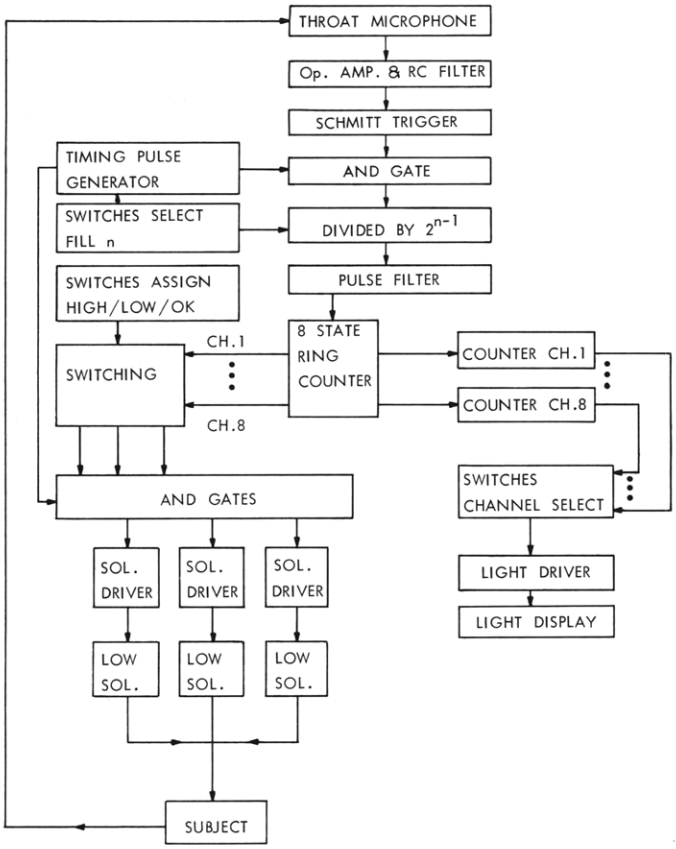


Fig. XI-14. System diagram.

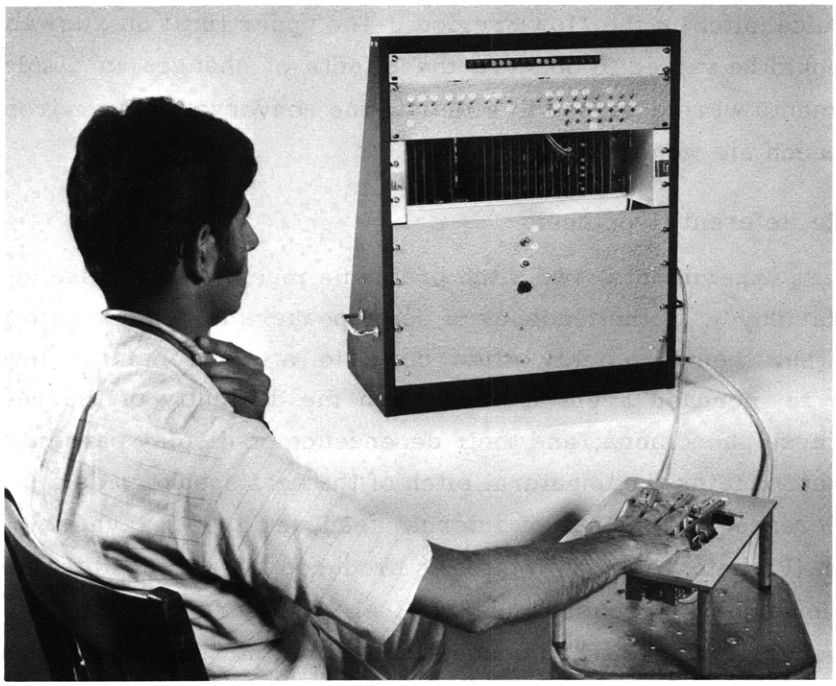


Fig. XI-15. Experimental arrangement.

Figure XI-14 is a block diagram of the system. Figure XI-15 shows the experimental arrangement: a subject holds the throat microphone with his left hand and touches the tactile display with his right.

3. Experiments

Two sets of experiments were conducted with 26 profoundly deaf boys and girls, 13-17 years old, participating at the Boston School for the Deaf. All subjects had suffered sensorineural hearing losses for all or most of their lives. Each spent one-half hour per week with the device over a period of up to 4 weeks. In the first series of experiments, a three-tactor display was used. The experimenter would assign one of the eight channels to the "ok" tactor and the remaining channels to the "high" and "low" tactors, as appropriate. The subject would place his hand on the display and attempt to sustain a hum in the selected target channel. These experiments used humming as a way to avoid linguistic influences of speech on voice pitch. By varying the target channel and the display rate, the experimenter could investigate the pitch control characteristics of the closed-loop system composed of subject, pitch detector, and display. The subjects' natural pitch distributions could be recorded by having the subject hum or speak without using the display.

In the second series of experiments a two-tactor display was used. In these cases, the 8 channels were grouped into "high" and "low" bands. The subjects were asked to repeat their names or read certain text passages or word lists while simultaneously maintaining voice pitch in the "low" region. The upper limit on allowable pitch and the display rate could be varied to observe the effects of changes in display parameters. These experiments were designed to simulate the conversational environment in which a wearable speech aid would function.

4. Kinesthetic Referent Hypothesis

During these experiments, two pitch problems manifested themselves. One, noted previously (Martony[2]), is the tendency of deaf speakers to begin breath groups at abnormally high pitch, then to quickly slide down to a more natural level. The other is a tendency to increase average pitch when the difficulty of the required utterance increases. These phenomena, and their dependence on display parameters, prompted a viewpoint that explains the unnatural pitch of the deaf speakers as a by-product of their attempts to increase the amount of internal feedback that is available during voicing.

It is known (Pickett[7]) that high pitch is produced by increased tension in the cricothyroid muscle and by increased subglottal air pressure. The extra vocal effort that is needed to generate high-pitched sounds leads to an increased kinesthetic awareness of voicing beyond that possibly available from residual hearing. We suggest that deaf speakers generate high-pitched tones as a way of better marking the onset of voicing

and the progress of voicing. The behavior of the subjects during experimental sessions also suggests that they generate high-pitched tones to serve as a reference or calibration which they use to "tune" their voices. Because the internal feedback is provided by the kinesthetic sense and appears to be used for frequency and voicing reference, the explanation for the use of high pitch was dubbed the "kinesthetic referent hypothesis." This hypothesis provides an appealing framework for understanding the successes achieved with the tactile display.

5. Experimental Corroboration

The use of high pitch as a kinesthetic referent appeared in both the humming and speaking experiments. In the former, the use of high pitch at the start of breath groups manifested itself in the data as a large number of counts in the highest channel, Channel 8. It was the dependence of this high Channel 8 count on display parameters that first suggested the hypothesis. We found that the subjects generally had a "natural" channel or group of channels in which they could hum fairly consistently. But when the target channel was chosen to be a more "unnatural" channel, confusion arose and performance became much more dependent on the monitoring information supplied by the tactile display. Performance at slower display rates was generally worse than that at high rates – at the slower rates, the subjects were observed to resort more to use of high pitch at the start of breath groups, presumably to supplement the frequency calibration data that was appearing in insufficient amounts on the tactile display. Figure XI-16 indicates the increasing reliance on the high-pitched referent as the display rate was slowed. Figure XI-17a and XI-17b illustrates the increasing reliance on the high-pitched referent as the target channel was moved farther from the "preferred" channel (in this case, Channel 5). Figure XI-17c and XI-17d demonstrates that for an easy pitch control task, that is, one in which the target channel is the most comfortable channel, good performance can be achieved even with slower feedback rates. These results are typical of the evidence from the humming experiments, which suggests the kinesthetic referent hypothesis.

Further evidence that deaf subjects generate their own pitch references comes from the need for referents on the part of normal hearing subjects in discrimination experiments. Stewart[8] noted:

> The trained ear can cope with five or six degrees of length (Jones, 1956) and can manage at least four pitch levels without difficulty, but only if these length and pitch distinctions are manifested in some sort of system. Given a largely random sequence of pitch or length factors to sort out, analysis by ear breaks down: deprived of meaningful relativity the ear fails to measure in absolute terms.

Analogously, we suggest that deaf speakers, when confused about the interpretation

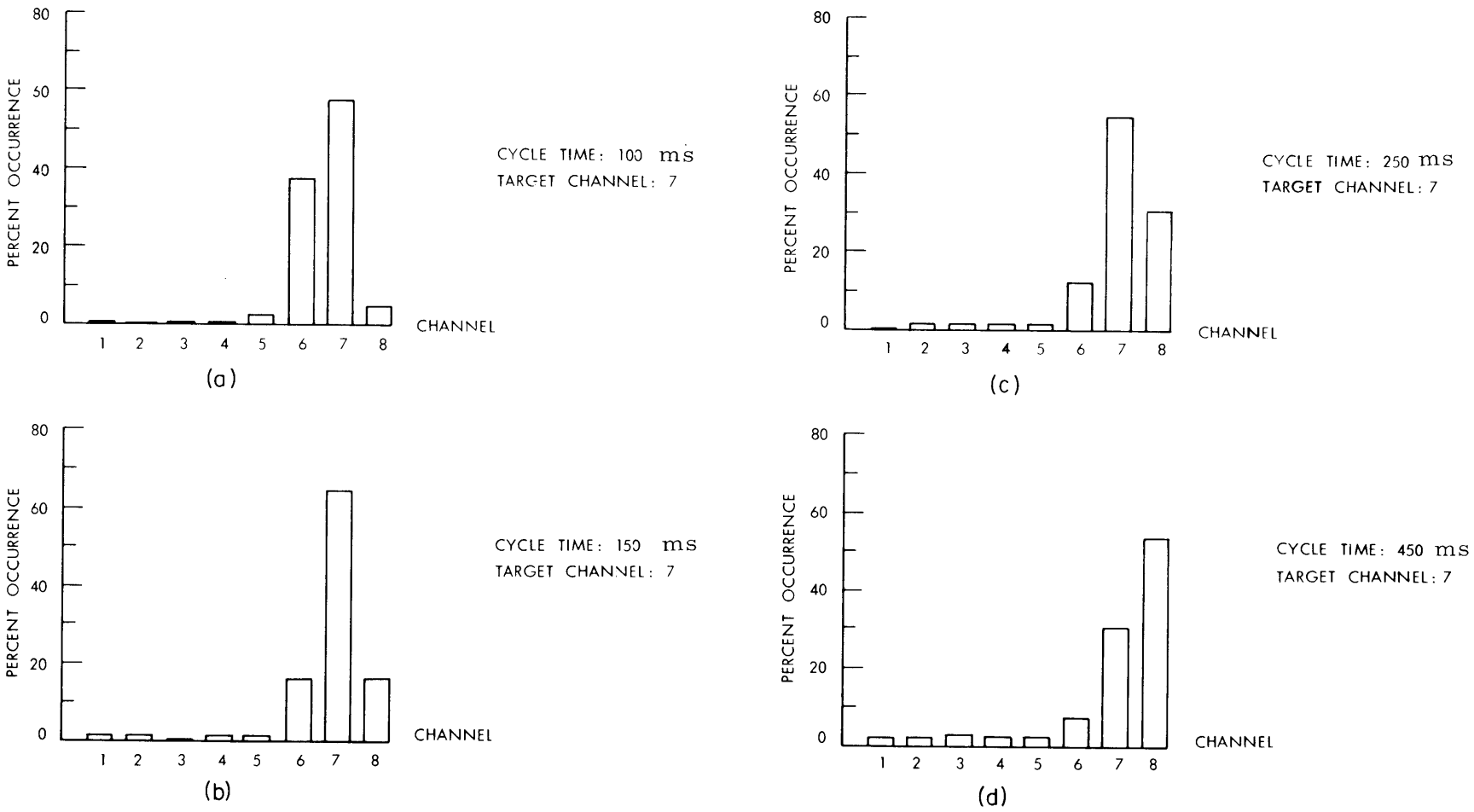


Fig. XI-16. Relation between performance and tactile display rate in a humming experiment. (Subject: KP (female).)

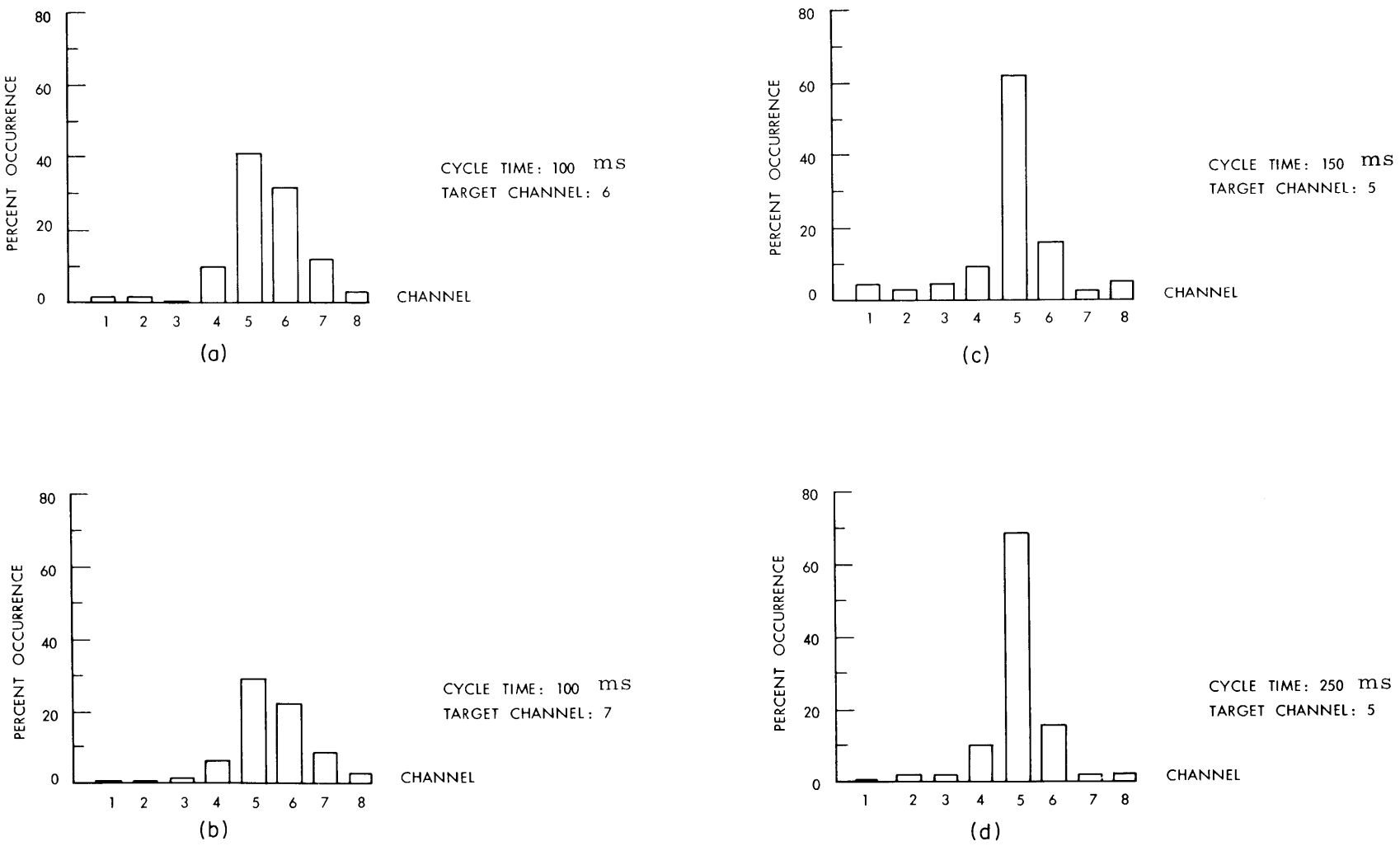


Fig. XI-17. Relation between performance and target channel in humming experiment. (Subject: SF (male).)

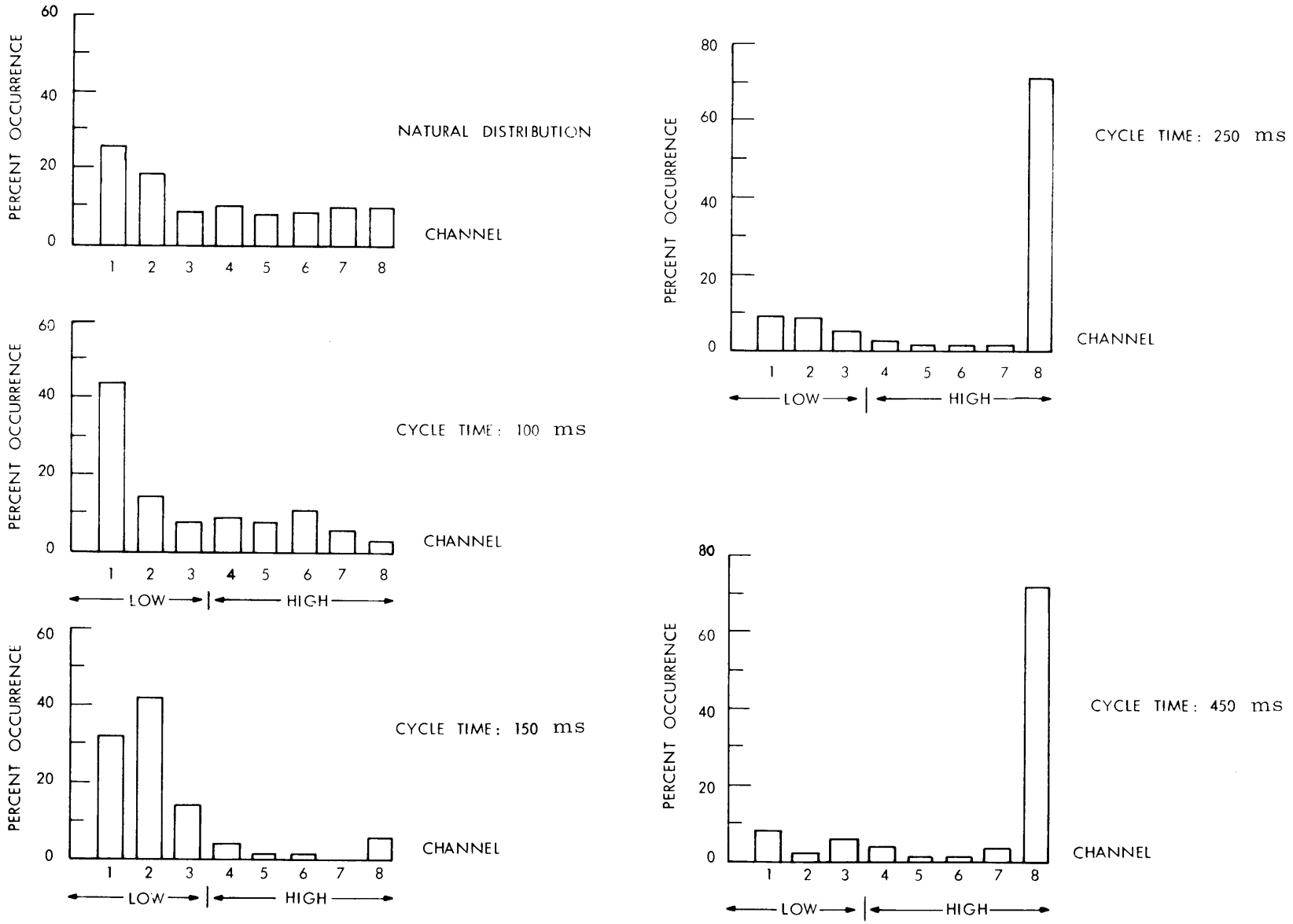


Fig. XI-18. Dependence of performance on display rate in a speaking experiment. (Subject: RP (male).)

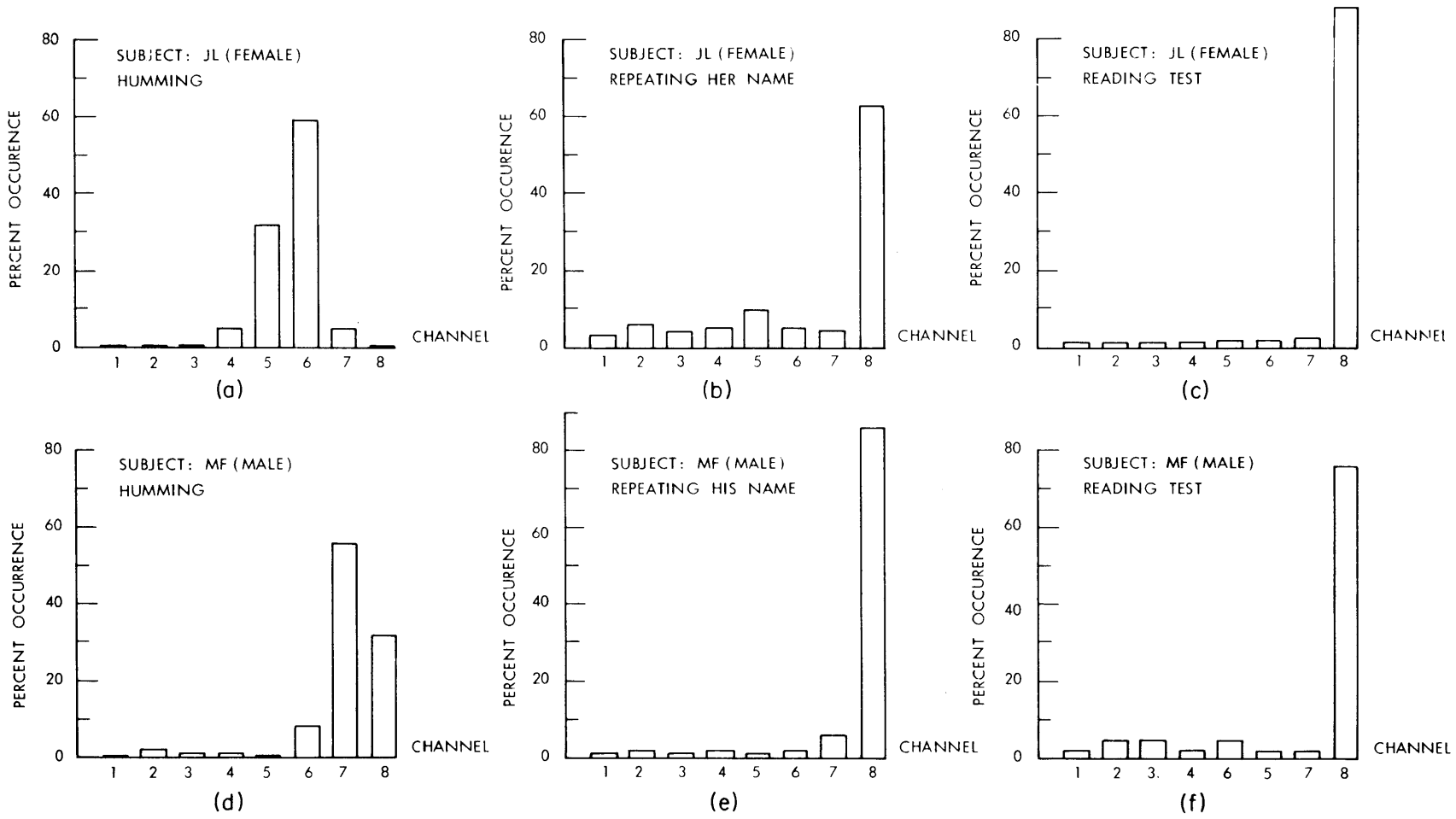


Fig. XI-19. Dependence of high pitch on type of utternace.

of the display or when receiving insufficient data from it, used a high-pitched tone as a fixed frequency reference, from which they could adjust their pitch downward toward the target channel.

Similar pitch behavior was noted in the second series of experiments, in which the subjects spoke rather than hummed. Figure XI-18 presents 7 successive experiments that indicate the dynamics of pitch modification and the tendency to resort to high pitch at low display rates. The first histogram shows the natural pitch distribution for the utterance, which was the subject's name; the rest were performed with tactile feedback. Note the effects of adjustment to the feedback and to changes in the display rate.

Also observed in the experiments involving speaking was a dependence of the average pitch on the type of utterance required. In general, the average pitch when the subject hummed was lower than that when he repeated his name, and this in turn was lower than the average pitch when reading text. The primary cause for the increases in average channel were increases in the Channel 8 component. In terms of the hypothesis, we would describe the subjects as increasing their monitoring data rates to match the increasing information content of their utterances. Figure XI-19 illustrates this dependence of pitch on complexity of utterance for two subjects.

We speculate that the closer monitoring of voicing achieved by use of high pitch might be caused by tension. In the setting of a school for the deaf, all occasions for reading aloud might readily be transformed into rather difficult tests of the students' abilities to achieve correct speech. The observed behavior of the subjects led the authors to conclude that tne deaf students had lost – or never acquired – the ability to read aloud casually. The obvious undercurrent of strain, even when the subjects read aloud without using the system, brought to mind the subjects' first encounters with the authors and the system. In many cases, the first session nervousness gave rise to pitch distributions that were abnormally high even for the subjects. Figure XI-20 shows some marked changes in pitch distribution for the same experiments performed first in the initial session and repeated one week later. We suspect that a similar type of tension might become associated with reading aloud and, after a period of years, become habitual. The results of the experiments using the tactile display indicate that the tendency to use high pitch can be controlled.

6. Results with Tactile Feedback

If the kinesthetic referent hypothesis is a good model for the pitch behavior of the profoundly deaf, it also provides a basis for hope that the pitch problems can be corrected. The sensitivity of the results of the humming experiments to the display rate suggests the answer: If sufficient feedback is provided via the tactile channel, this information can substitute for that otherwise provided by the kinesthetic sense.

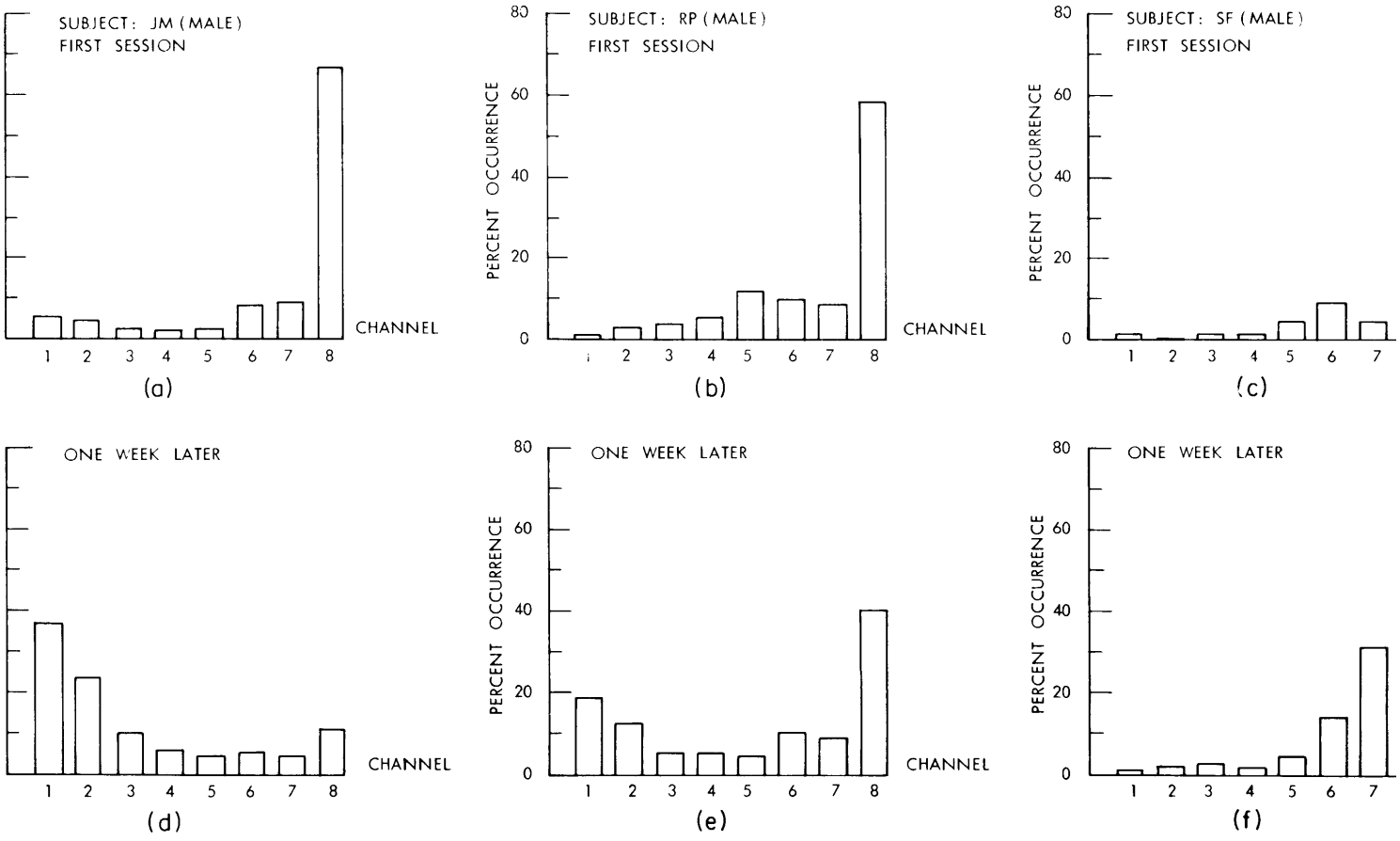


Fig. XI-20. Effects of first-session nervousness on pitch distributions.

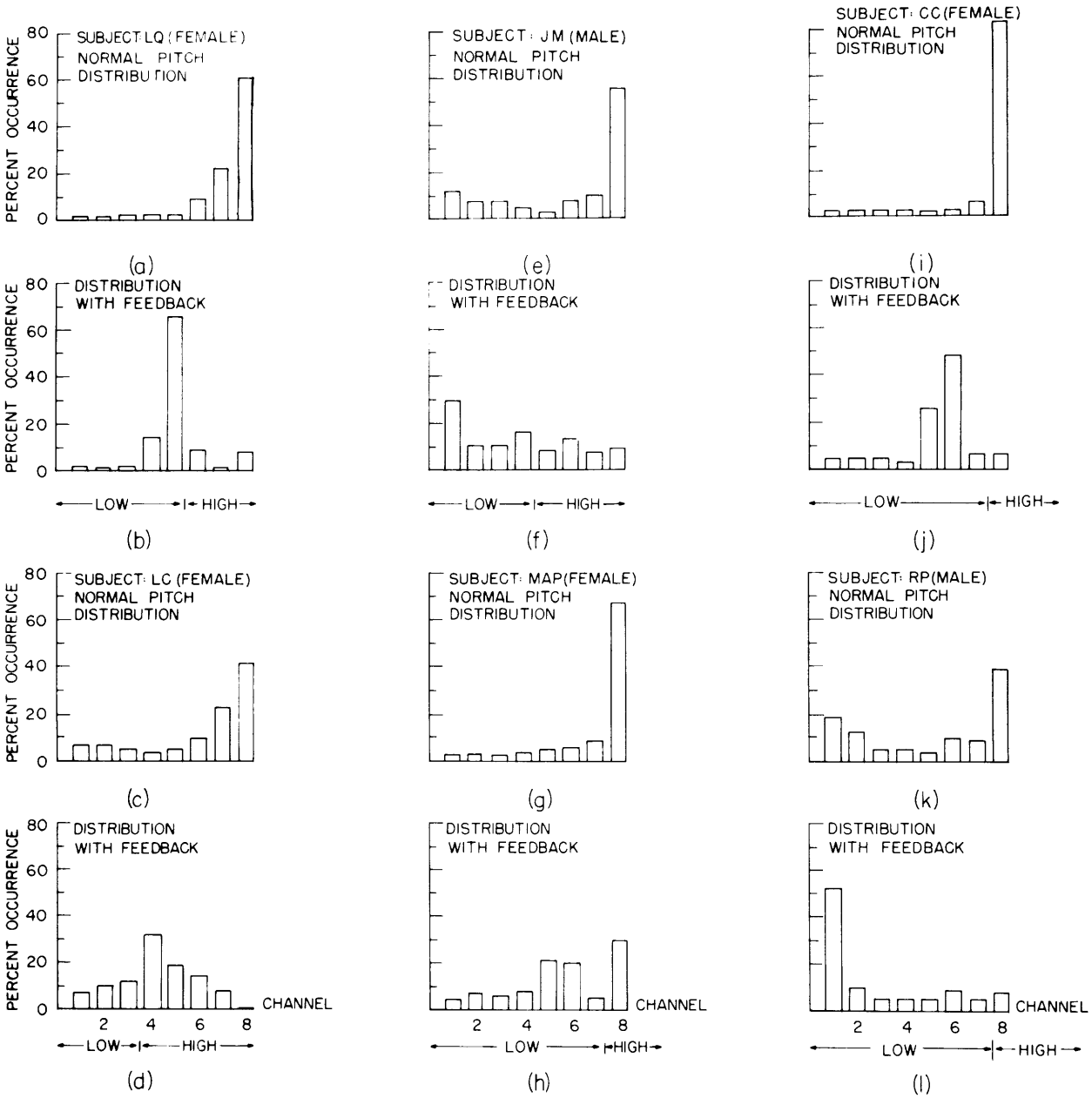


Fig. XI-21. Modification of pitch distributions for text reading by tactile feedback.

The results of tests using the tactile aid confirm this concept of sensory substitution. Not every subject learned – or had enough time to learn – to use the binary tactile display to eliminate the use of abnormally high pitch at the start of breath groups. But many did come to understand the concept of pitch and to acquire sufficient motor control to limit their pitch to lower frequencies. In Fig. XI-21 some experiments performed with and without binary tactile feedback are compared, and the extent of the changes that are possible are indicated.

Experiments were also conducted in which the subjects were given no feedback but urged to speak with as low a pitch as possible. The extent of the shifts in pitch distribution produced in this way were often of the same order as those achieved with the display. Some subjects could not lower their pitch, however, by any means other than use of the display. For others, use of the display helped shift the pitch distribution downward beyond what could be accomplished unaided. These results suggest that continuous use of a wearable aid would provide a sufficient reminder to use proper intonation, and might be abandoned as soon as the user internalizes the motor controls for acceptable pitch. Figure XI-22 illustrates the comparison between unaided efforts at pitch control and efforts assisted by tactile feedback; it also shows another instance of the increasing dependence on high pitch as the complexity of the required utterance increases, and of the ability of the deaf speaker to counter this dependence if special attention is called to it. It should be noted that concentration on use of the display for pitch control sometimes led to deterioration in other aspects of voice quality: Durations of utterances sometimes increased, loudness occasionally dropped, and articulation sometimes suffered. These problems were neither universal nor severe, and one might suppose that more extensive familiarity with the device (no child had more than 1 hour of actual feedback experience) would correct these difficulties.

7. Conclusions

A simple tactile display driven by an uncomplicated pitch detector has been successfully used to correct a common defect in the intonation patterns of deaf speakers. An interesting dependence of unusually high pitch at the beginning of breath groups on the complexity of an utterance was observed in the speech of profoundly deaf teenagers. The dependence of this anomalous intonation pattern on complexity of utterance (and therefore, in some sense, on tension) and on display parameters led to a hypothesis that sees high-pitched tones as kinesthetically monitored referents for voicing and frequency information. This "kinesthetic referent hypothesis" provides a framework in which to interpret the pitch problem and the usefulness of the tactile speech aid in dealing with it. The success of this aid in infrequent test sessions justifies more confidence in the usefulness of a wearable aid, which could realize the presumed advantages of continuous feedback. Further work toward a wearable aid should be encouraged. The particular

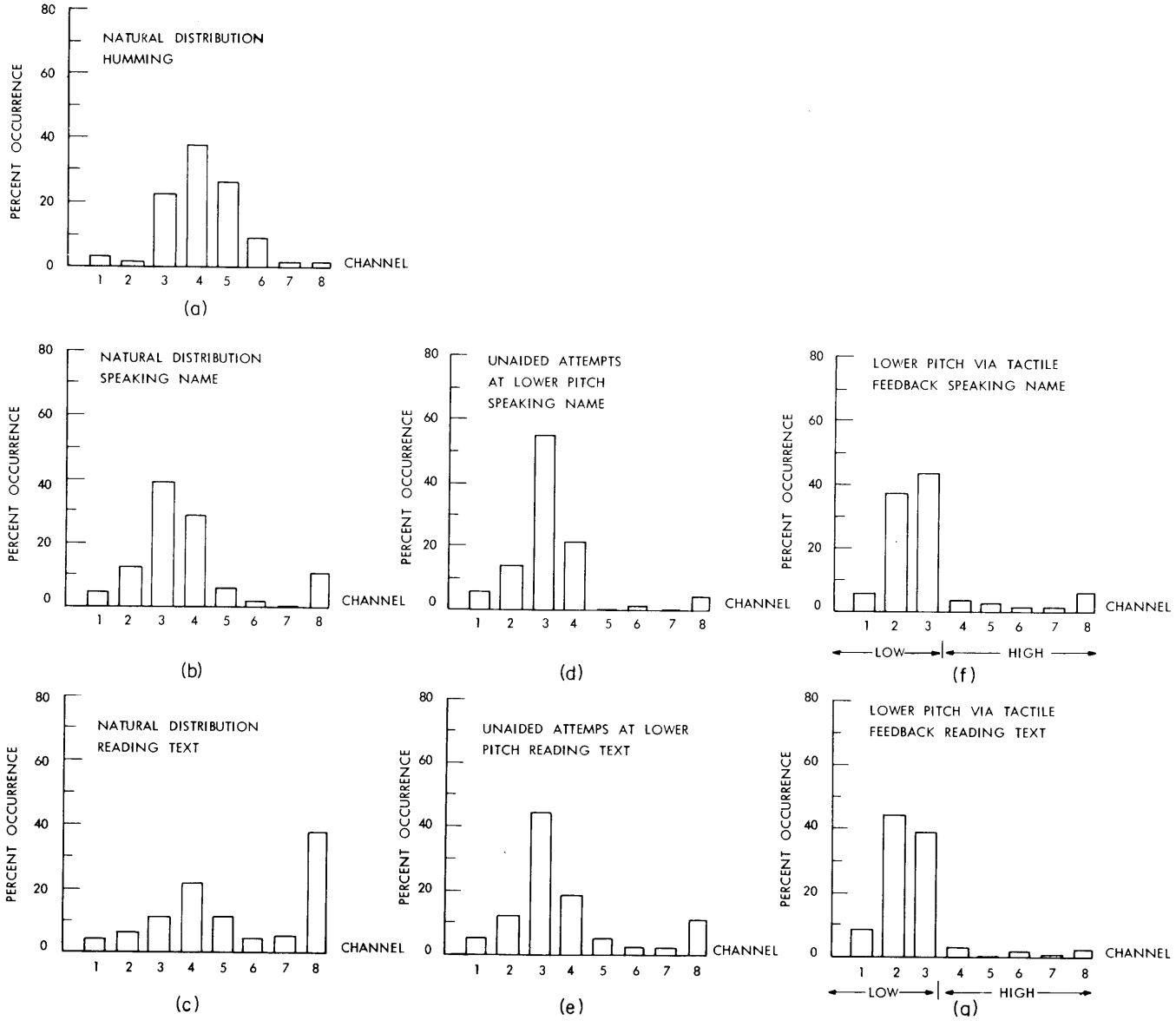


Fig. XI-22. Pitch distribution as a function of utterance and the effectiveness of pitch modification with and without tactile feedback. (Subject: MF (male).)

pitch detector used here can be easily miniaturized; the outstanding technical problems remaining are the choice of tactile display mechanism and the human factors aspects of the design.

Thanks are due to Sister Kieran and to the staff and students of the Boston School for the Deaf, Randolph, Massachusetts, for their generous cooperation.

This report is based on a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science, in the Department of Electrical Engineering, M. I. T., August 20, 1970.

T. R. Willemain, F. F. Lee

References

1. A. Angelocci, G. Kopp, and A. Holbrook, "The Vowel Formants of Deaf and Normal Hearing Eleven-to-Fourteen-Year-Old Boys," J. Speech Hearing Disorders 29, 2 (May 1964).
2. J. Martony, "On the Correction of the Voice Pitch Level for Severely Hard of Hearing Subjects," Am. Ann. Deaf 113, 2 (March 1968).
3. K. Børrild, "Experience with the Design and Use of Technical Aids for the Training of Deaf and Hard of Hearing Children," Am. Ann. Deaf 113, 2 (March 1968).
4. L. Dolansky, W. Pronovost, D. Anderson, S. Bass, and N. Phillips, "Teaching of Intonation Patterns to the Deaf Using the Instantaneous Pitch-Period Indicator," VRA Project No. 2360-S, Northeastern University, Boston, Mass., February 1969.
5. A. Risberg, "Visual Aids for Speech Correction," Am. Ann. Deaf 113, 2 (March 1968).
6. T. P. Willemain III, "Tactile Pitch Feedback for the Profoundly Deaf," S. M. Thesis, Department of Electrical Engineering, M. I. T., August 1970.
7. J. Pickett, "Sound Patterns of Speech: An Introductory Sketch," Am. Ann. Deaf 113, 2 (March 1968).
8. R. Stewart, "By Ear Alone," Am. Ann. Deaf 113, 2 (March 1968).

F. PROPOSED AUTOMATIC LEUKOCYTE DIFFERENTIAL ANALYZER

The leukocyte differential count is an extremely common medical test that thus far has escaped automation. The differential involves the visual classification of white blood cell types. These types are normally distinguished by size, color, and the presence or absence of granulation in the cytoplasm of the cells. It has been found[1] that the spectral extinction and spatial frequency properties of leukocyte images can alone be used as discriminators in an automatic recognition procedure. Details of these two properties and their measurement are discussed in this report, and a proposal is made for an optical processing system to perform the leukocyte differential.

In previous studies the gross spectral differences between white and red blood cells[2] and the particular spectral properties of the white-cell cytoplasm have been analyzed.[3] Spectral extinction of whole white cells can itself be used, however, to separate the individual leukocyte classes. Spectral measurements were made on Wright's stained blood smears using a photometer attached directly to a high-power microscope. The photometer sampling aperture, when projected into the object plane, was approximately 17 μm in diameter. Fifty white cells were analyzed on this system at wavelengths of 535, 570, 605 and 640 nm. Six features based on these measurements were considered. A signature analysis of the data showed that 4 leukocyte categories could be distinguished:

(i) Lymphocytes, (ii) Neutrophils, (iii) Eosinophils, and (iv) Basophils and Monocytes. Figure XI-23 shows the confusion matrix for the spectral classifications. There did not seem to be any way to separate category 4 which would be based on spectral differences alone.

Three of the basic leukocyte classes, neutrophils, eosinophils and basophils, have distinctive granulation in their cytoplasm. Knowledge of the size, color, and number of

CELL TYPE \ IDENTIFICATION	L	N	E	MB
L	9	0	0	1
N	1	9	1	0
E	0	1	9	0
MB	1	2	0	17

L = LYMPHOCYTE E = EOSINOPHIL

N = NEUTROPHIL MB = MONOCYTE-BASOPHIL

Fig. XI-23. Confusion matrix for spectral data.

these granules is sufficient to determine the type of cell under observation. A possible technique for recognizing granulation is two-dimensional Fourier analysis. Objects of constant size and shape have characteristic Fourier patterns. Applying Fourier analysis to blood cells, then, might yield a useful classification prodcedure for the granulated classes. Another important reason for consideration of Fourier methods is that certain optical configurations can perform a Fourier transformation of an input light distribution.[4] A simulation of this technique was performed on a PDP-9 digital computer. Photographs of Wright's stained leukocytes were taken at wavelengths of 535 and 605 nm. These colors were chosen to accentuate differences in granular color among cell types. The processed transparencies were then scanned on a 256 × 256 raster by a flying-spot scanner and stored in the computed memory. The stored picture represented a 25-μm square field. A Fast Fourier transform algorithm was applied to the picture and the spatial frequency intensities in two annular rings were computed. These two values were then used as features in a recognition procedure. Thirty cells, 10 each of neutrophils, eosinophils, and basophils, were analyzed by this method. Figure XI-24 shows the results of this analysis.

CELL TYPE \ IDENTIFICATION	N	E	B
N	8	1	1
E	0	10	0
B	1	0	9

Fig. XI-24. Confusion matrix for Fourier data.

Ten monocytes were also analyzed in an attempt to separate the monocyte-basophil category found in the spectral analysis. The monocytes and basophils are normally easy to distinguish. Basophils are small cells with prominent dark granules. Monocytes are larger cells with little or no granulation. Fourier analysis should be able to separate these two cell types. All of the monocytes tested were classified as either neutrophil or eosinophil. These results suggest that the two methods, when used in conjunction, could successfully discriminate the 5 basic leukocyte categories.

It has been mentioned that both analysis methods can be realized by an optical processing system. Such a system has, together with the near instantaneous speed of all optical devices, two valuable characteristics. It can be shown[1] that the appropriate configuration is invariant with respect to lateral motion of the object and small vertical (focus) motion. This means that the specimen blood cells could be in continuous motion during processing. The medium itself could be either a prepared

slide or a suspension of cells such as that described by Kametsky and Melamed.[5] In either case, such a system would seem to offer truly high-speed analysis of white blood cells.

J. E. Bowie

References

1. J. E. Bowie, "Differential Leukocyte Classification Using an Optical Processing System," S. M. Thesis, Department of Electrical Engineering, M. I. T., May 1970.
2. O. B. Lurie, R. E. Bykov, and E. P. Popetchitelev, "Automatic Recognition of Leukocyte Cells in Blood Preparations," Proc. 8th International Conference on Medical and Biological Engineering, Chicago, 1969.
3. I. T. Young, "Automated Leukocyte Recognition," Ph. D. Thesis, Department of Electrical Engineering, M. I. T., June 1969.
4. J. S. Goodman, "Introduction to Fourier Optics (McGraw-Hill Book Company, New York, 1968), p. 9.
5. L. A. Kametsky and M. R. Melamed, "Spectrophotometric Cell Sorter, "Science 156, 1364 (1967).