

ALTERNATIVE STRATEGIES IN THE EVALUATION OF SPEECH SYSTEMS

by

David I. Mostofsky

Educational Research Corporation

10 Craigie Street

Cambridge, Massachusetts 02138

Contract No. F19628-68-C-0155

Project No. 4610

Task No. 461002

Unit No. 46100201

FINAL REPORT

Period Covered: 15 February 1968 - 14 July 1969

August 1969

Contract Monitor: Caldwell P. Smith

Data Sciences Laboratory

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

Prepared for

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS 01730

AD 694100

ABSTRACT

This report summarizes the results of a program of research concerned with the perception of degraded speech in normal and pathological listeners. The comparisons were derived from performance in a two-category judgment task where the anchor values remain accessible to the subject and under the control of the experimenter. Response decisions and frequency of anchor testings are recorded together with decision times in computer compatible tape format. In addition, studies were designed to examine selected features of utility effects and applications of unidimensional and multidimensional scaling procedures. The respective sections of this report describe (1) the development and applications of the modified, non-verbal psychophysical technique; (2) the design of a data collection system; (3) variations in response utility; and (4) selected scaling procedures. These investigations were shown to be particularly appropriate for assessments of communication with vocoded materials in addition to suggesting alternative experimental paradigms for the study of audiological and perceptual problems.

TABLE OF CONTENTS

Introduction.....	1
Section I--Studies in the Perception of Vocoded Speech Using a Modified Psychophysical Technique.....	7
Appendix A--Test Stimulus Materials.....	64
Appendix B--Anchor Material: Connected Text.....	70
Appendix C--A Program for DADER: <u>D</u> ata <u>A</u> nalysis and <u>D</u> ump of <u>E</u> vent <u>R</u> ecords.....	72
Bibliography.....	76
Section II--Digital Data Acquisition Facility.....	82
Section III--The Effect of Utility on Intelligibility Testing.....	100
Reference:	108
Section IV--Selected Scaling Procedures.....	109
References.....	123

LIST OF TABLES

Section II

Table 1	Sample Tape Format.....	87
Table 2	Sample Event Code Configuration for 10 Events.....	91
Table 3	Sample Guide for Entering Manual Data.....	92

Section III

Table 1	Analysis of Variance For DRT at Different Quantizing Levels.....	106
Table 2	Means (\bar{X}) and Standard Deviations (σ) For DRT Performance.....	107

LIST OF FIGURES

Section I

Figure 1	Sample Trial Record.....	36
Figure 2	Diagram of Room Arrangement.....	41
Figure 3	Mean Decision Times at Each Level and For Standard and Reversed Speech (Normal)..	44
Figure 4	Mean Decision Times Level (Ears Pooled) for Standard and Reversed Speech (Normal)..	45
Figure 5	Mean Decision Times at Each Level and Each Ear for Standard and Reversed Speech (Hemiplegic).....	46
Figure 6	Mean Response Times Averaged Over Three Levels for Each Ear for Standard and Reversed Speech (Hemiplegics).....	47
Figure 7	Mean Frequency of Anchor Tests at Each Level and Each Ear for Standard and Reversed Speech.....	49
Figure 8	Mean Frequency of Anchor Tests at Each Level and Each Ear for Standard and Reversed Speech (Hemiplegic).....	50
Figure 9	Mean Frequency Decision Category I at Each Level for Each Ear for Standard and Reversed Speech (Normal).....	51
Figure 10	Mean Frequency Decision Category II, Each Level and Each Ear, for Standard and Reversed Speech (Normal).....	52
Figure 11	Mean Frequency Response Category I at Each Level for Each Ear for Standard and Reversed Speech (Hemiplegics).....	54
Figure 12	Response Category II for All Levels Each Ear, Hemiplegic, Standard and Reversed Speech.....	56

Section II

Figure 1	Data Acquisition System: Time Clock Section	94
Figure 2	Data Acquisition System: Binary Encoder Section.....	95
Figure 3	Data Acquisition System: Alarm and Write Section.....	96
Figure 4	Data Acquisition System: Interface with Experimental Apparatus used in the Study...	97
Figure 5	Details of Subject Response Station Circuitry.....	98
Figure 6	Photograph of Completed System. Console houses digital tape deck together with logic modules and power supply.....	99

BLANK PAGE

INTRODUCTION

The following four sections of this report, respectively, describe a major accomplishment of the research effort performed by the Education Research Corporation under contract F19628-68-C-0155 from 15 February 1968 through 14 July 1969.

Publications

While publications based on the research conducted under this contract have not yet appeared in the relevant professional literature, six reports are in various stages of preparation for submission to the appropriate technical journals. They are tentatively titled as follows:

"Modified category judgment technique in the assessment of complex stimuli" D. Mostofsky and Marianne Noyes

"Audiological implications of responses to synthetic speech by brain-damaged subjects" Marianne Noyes

"Magnetic, digital tape data collection system for multiple event recording" D. Mostofsky

"The effect of utility on speech intelligibility scores" D. Mostofsky

"Estimates of scaled similarity following ANOVA" D. Mostofsky and J. Alman

"Multidimensional scale contours of complex auditory stimuli" D. Mostofsky and R. VandenBossche

A dissertation "Perception of degraded speech by normal and brain-damaged subjects" by Marianne Noyes has been submitted to Boston University and derives, in part, from the activities of this contract.

Technical Accomplishments

The objectives of the contract activity were to explore alternative strategies which might be applied to the study of speech systems. Most previously documented efforts have relied upon a limited number of techniques and have, in many instances, drawn mainly upon the verbal response system. The complexity of the speech signal and its perception suggested that investigations using non-verbal response repertoires from normal as well as brain injured subjects would profitably contribute to an understanding of extant data, and provide direction for continued research in the fundamentals of the perception of vocoded and normal speech. Such an investigation was completed following the design of a technique which fulfilled crucial requirements and which afforded additional insights to the judgmental process. The technique begins with a standard category judgment method, which required the subject to assign each of n different stimuli to either response category j or k . No other categories are made available, nor are any additional referents or dimensions provided or described. Traditionally, such psychophysical determinations employing category assignment techniques do not allow for systematically observing subject's utilization of the anchor stimuli. In many situations subject's referral to the anchors is assumed to take place (if at all) by memory or by recourse to instructions or the customary brief exposure given at the beginning of the experiment. In the present experiment subjects were required to decide on the category for each of nine different vocoded speech samples. Prior to

making his decision the subject was able to produce and listen to either anchor value as often as wished. The number of such "testings" provided data in addition to the conventional ogive obtained in the conventional procedures, and suggested an approach to the study of attention and judgmental confidence. It seems tenable to assume that whenever the subject elects to test an anchor this behavior gives evidence of uncertainty concerning his decision. The uncertainty may result from the ambiguity of the test stimulus with respect to its position between the extremes of the relevant dimension, and thus describes a measure of stimulus difficulty. The uncertainty can then be regarded to have been somewhat diminished as a result of cue or discriminative properties that were made available to the subject by added exposure to the anchor. The frequency of testing would therefore constitute legitimate evidence for both attention and judgment confidence, although decision time has been more popularly offered for the latter. Furthermore, to the extent to which performance characteristics for several stimuli are alike, one might expect other correlates of the stimuli to be judged and to exert comparable similarity. For example, if two speech samples are both judged assignable to the same anchor category with comparable decision times and testing frequencies, one might regard the stimuli as comparable in producing like effects in the performance of the communication system, despite any other differences which may describe their physical compositions.

This technique and the use of vocoded materials has particularly powerful implications for audiological research and practice. The

problems, issues, and the results of the study in which the technique was applied, are discussed in Section I of this report.

The nature of the data required for this and related tasks can be feasibly collected and processed if an appropriate recording system can be implemented in the overall program. Not only must the hardware be reliable and cost effective, but the data collection facility should be modular and expandable. In addition, media conversion or pre-processing for computer analyses should not be necessary. Details of the development of such a data collection system are described in Section II.

The central role of the performing subject in defining stimulus differences suggested a re-examination of factors which have been shown to be relevant in other, unrelated, psychological studies. Section III of this report describes the results of such an investigation of the utility or cost factors of accurate responding.

The subject also gives evidence of his discriminatory capabilities not only by correctly identifying a test value from a set of alternatives, but also by the extent to which he is likely to confuse, or regard as comparable, different values of test stimuli. These judgmental activities do not in themselves require an extensive verbal response repertory and may therefore be adapted for handicapped respondents. Such techniques may be derived post hoc - as from psychophysical procedures - and estimates of interscale distances determined from other analyses. One such novel procedure was formulated and is described in Section IV. This section also examines a direct

multidimensional scaling technique together with appropriate analytic routines which are already in the public domain. The applicability of these techniques to speech scaling seems unusually promising and deserves the serious attention of concerned scientists.

A portion of the contract period was directed to exploring the extent of scientific activity by behavioral scientists in the above problem areas and the possibilities for stimulating such energies in the future. The exploration of invitational meetings was considered, and was in part satisfied by a symposium sponsored by the American Speech and Hearing Association, where Dr. Mostofsky was invited to chair the session "Theoretical and applied aspects of time compressed and transposed speech."

The general conduct of this research program was fraught with the usual setbacks, delays, and inconveniences. Its success, however, is due in large measure to the collective efforts of many people. Some have rendered invaluable service, and their contribution has been essential and inescapable of mention. I am particularly grateful to be able to thank Caldwell Smith and his associates who have cooperated in every instance, and without whom I hesitate to contemplate how the program might have been effectively pursued. From his patience and cooperation I was sustained, and from his help I learned much. Marianne Noyes was particularly valuable in the conduct of the experiment. John Alman developed the extrapolation of scale estimates from ANOVA, and contributed greatly to many phases of the study. The administrative support from Rita Muldowney and Sarah James made the work humanly possible. For all that is worthwhile I share with them. For all else the author assumes sole responsibility.

SECTION I

STUDIES IN THE PERCEPTION OF
VOCODED SPEECH USING A
MODIFIED PSYCHOPHYSICAL TECHNIQUE

PERCEPTION OF VOCODED SPEECH USING A MODIFIED PSYCHOPHYSICAL TECHNIQUE

Introduction

Whenever a human observer or participant forms a vital link in a complicated system, the system itself, of necessity, must be additionally described from the referent of the human serving as the transducer of that system. It is no longer adequate under such circumstances to conceive of descriptive indices derived exclusively from the standpoint of the inanimate engineered components. Indeed such has been the past history of many technologically engineered systems. Each such system (if it is truly a system) reaches a critical point where the issues of consequence are almost exclusively defined by the performance of the human within the system. In the case of speech and communication systems, the specifications of physical correlates, attributes, or quanta may be of secondary importance. The problem (at some point) is no longer soluble except in terms of the psychological characteristics of the performing subject. These may be analyzed in the framework of classical psychophysical orientations, or even as independent manifestations of problems associated with judgment and perception. The relevant "behavioral space" may be defined by at least the following criteria:

- (a) stimulus conditions
- (b) subject characteristics
- (c) task demands
- (d) response consequences and contingencies.

Previous experimental determinations involving vocoded passages have been particularly focused on variations in the physical characteristics of the speech passage. Parametric variations have been applied to study

sundry aspects important to the implementation of vocoding, among them speech intelligibility, speech quality, and speaker recognizability. The documented application of accepted techniques and concepts have been effective in furthering the available knowledge in these areas.

Similarly, these studies have in general relied upon trained teams of listeners in the experiments thus conducted. There is much to be said for restricting participation to non-naive subjects, although such a restriction by its very nature potentially restricts generalizability of the findings. In addition, one might legitimately hope for parametric manipulation of familiarity as relevant to practical considerations for operational employment of vocoding systems. One might also hope that for such interests, as well as for the larger interest of speech science, subject traits other than familiarity (audiological as well as non-biological organismic factors) might be vigorously explored.

The usual response demands require the subject to attend and somehow render his rating, judgment, or response. In the usual case, the subject is instructed or informed of the relevant dimension(s) to which he is expected to attend. Also, because of feasibility, large scale testing programs require multiple choice verbal responses which utilize the recognition technique, to the exclusion of recall or alternative procedures.

Conventionally, little if any feedback is presented to the subject on a trial by trial basis. While there is little evidence to suggest the importance of feedback and cost considerations on such performance, the paucity of such information only strengthens the requirement for investigations along such lines. More positively, the operative nature

of such variables in related learning and performance areas strongly argues for the examination of these variables as likely contributors to the output of such a system.

The program of activity described in this report in no way suggests flaws of assumption, technique, or inference in the many studies which have been conducted in the past. It is rather directed to an examination of other techniques which may be properly applied to answer similar, if not identical, issues and thereby extend previous findings. Indeed, it assumes that the contribution made by previous studies demands corroboration to establish that those findings were not technique-specific or limited to the specific designs incorporated. Furthermore, the implications for psychoacoustics, audiology, and psychological theory are penetrating, and argue for the application of diverse tactics in any assessment protocol.

More particularly, a major objective of the contract was to examine perceived differences in vocoded materials while placing minimal reliance on verbal instructions or on verbal skills. To this end, the development of an appropriate technique was undertaken together with the necessary instrumentation design for eventual automated larger scale testing. Such a technique would be especially desirable when a priori characteristics (dimensions) of the signal are not intended for announcement at the beginning of a test session. This may be desired if (a) a priori characteristics are unknown or only tentatively identifiable, (b) it would prove too cumbersome to outline this information to an experimental subject, and (c) it is preferred to determine the number and extent of such characteristics from the subjects' eventual response protocol. Naturally,

it is expected that such a technique would not violate accepted psychophysical procedure, but would rather build upon and extend the capabilities inherent in extant models. Not only would such a technique lend itself to a study of synthetic speech in systems using intact subjects, but it would be of particular value when applied to an examination and assessment of vocoded (and other) materials by subjects having known and specifiable organic damage. Effectively applied, one might hope that further insight to the very fundamentals of speech perception and attentional hierarchies itself might be enhanced.

While directed to the development and refinement of such a technique, the contract activity was also concerned with a number of related considerations which lie on the periphery of psychophysical testing programs. Initially, the design of an accompanying data acquisition facility for this and comparable studies was undertaken and completed. This phase also included basic computer programming needs for efficient maximization of the advantages offered by such hardware configurations. The system enables an off line computer analysis for simultaneous testing of subjects where events and time of occurrence of events are of interest. In addition, the contract activity examined the contribution of response utility to estimates of intelligibility scores. As will be seen in the report, there appears to be sufficient grounds to justify a carefully detailed series of examinations of this issue. This encouragement derives not only from an impressive history where utility has been known to alter comparable behavior, but may be seen even from the investigation conducted during this contract period which used naive subjects, variable in performance under moderately utile conditions with stimulus

presentations where intelligibility indices were previously shown to be well beyond minimal levels of acceptability and having a limited ceiling for noting marked improvement. Finally, the contract activity devoted part of its energy to testing the feasibility and merit of applying scaling procedures in subsequent evaluative programs. One attempt was pursued which might enable a scale derivation following conventional analysis of variance. Of greater interest was the promise held in the multidimensional method of incomplete triads; a technique which is workable with such materials and which might enable a radically different avenue of perceptual interpretations. These efforts all argue for future support of intensive parametric studies. Not discussed in this report are still other issues which suggested themselves during the course of this program of research. Among such concerns are the relationship of data derived from trained vs untrained subjects; the implications of such findings for the training of personnel who must respond to such speech materials, and the specific task objectives, risks, and quality of performance which such personnel must face. Finally, one cannot fail to be impressed with the contributions of such a research program to audiological research. In many instances, small, pilot sized studies were in fact initiated, but, in the main, such findings formed the propedeutic base for the conduct of those investigations which were completed and which are documented on the following pages.

Theoretical Considerations

From the standpoint of communication, auditory perception might be divided into two aspects. A distinction can be made between classical psychophysics and the recognition of acoustic signals presented within a linguistic framework. The first has to do with the abilities and limitations of the auditory system for all signals while the second is concerned with the identification and classification of auditory patterns which have significance associated with speech.

A further difference derives from the methodology employed. In classical psychophysics for example, usually only one dimension of a signal is examined at a time: frequency, intensity or duration. The task is normally one of close comparisons, and the objective is to discover differential thresholds.

Speech, however, is a complex signal which varies simultaneously in all dimensions and which has linguistic significance. The sounds of speech constitute a code which in order to be an effective transmitter of meaning, must be capable of being recognized absolutely. That is, it must be possible to break the code into elements which can be categorized absolutely. The task then becomes one of recognition and categorization rather than one of discriminative differences. One of the major problems in the study of speech perception is the identification of the significant perceptual elements (40, 41, 42). It is well known that in spite of wide acoustic differences in sound production by different speakers, a correct categorization will nevertheless be made by the listener. For a recognition to be made, not all of the

speech signal is required. Many studies have demonstrated that removing part of the speech signal by frequency filtering, clipping, time or frequency compression results in a signal which is readily understood by the normal ear (3, 4, 5, 8, 12, 43, 49, 58). Our present state of knowledge does not permit an understanding of how the ear and brain translate the code of speech.

Although a complete understanding of speech perception is not available, there is much that can be said about areas related to the perception of speech. Some psychophysical studies can be related to speech even though the signals are not linguistic. In addition, much investigation has centered on the acoustic cues required for recognition. Beyond this, studies concerning intelligibility of transmission systems and studies of speech discrimination in normal and pathological subjects help to identify factors upon which perception depends.

Psychophysical Studies

Psychophysical measurements of sensation are typically concerned with measurements of close comparison. In the study of speech, however, a measure of recognition is more commonly used. It is known that in measures of close comparison the resolving power of the ear is extremely fine. For example, in tests for determining difference limens for frequency, the difference may be as small as one part in a thousand (53).

Compared to differential sensitivity, absolute identification of frequency is quite poor. In a differential task, as many as three hundred thousand judgments may be made, while in an absolute judgment

task, only about five judgments are accurate (50). However, if more dimensions are added to the stimulus (by frequency, intensity or duration) accuracy of absolute discrimination increases with the number of dimensions (20). It is clear that tests of differential and absolute discrimination will yield different estimates of the capacity of the ear to perceive ongoing stimuli. Differential discrimination results suggest that many more discriminations can be made than are actually used in the identification-recognition task. Thus, recognition may be accurate in spite of acoustic differences in the speech signal.

With no reference to linguistic context, some psychophysical studies have been concerned with such topics as difference limens for formant frequency, formant amplitude, formant bandwidth and fundamental frequency. Most of such studies utilize synthetic speech in which the various dimensions may be readily varied. These studies demonstrate that the difference limen, when applied to speech is a highly complex measure and varies with such factors as frequency, proximity of formants, amplitude and place of articulation (16, 25, 28, 29, 41). It can be shown that in some cases a particular acoustic event will be of perceptual significance, while in other cases it is not. Apparently many of the small acoustic variations in the speech wave are meaningless and do not contribute to perception (20).

Recognition of Speech Signals

Many studies of speech sound perception have included the use of synthesized versions of speech sounds as a method of eliminating speaker

variations in the interest of quantification and objectivization. Such a technique permits variation of the produced sound in all of its dimensions in order to determine the important acoustic cues associated with a perception. For example, a speech synthesizer may be made to generate vowels with any number of formants desired which may then be used for recognition experiments. Such studies have demonstrated that the first three formants are more than sufficient for the identification of vowels (24). This does not mean that the same combination of resonances will always be identified as the same sound. Formants for the same vowel may vary over a wide range. This is in keeping with the fact that appreciable scatter is found when different speakers produce the same vowel. There is also some evidence that when the lower formants are absent, vowels may be identified from only the higher formants (16). This suggests that there may be several acoustic cues available for recognition.

A considerable amount of study has been concerned with the perception of isolated syllables. Much of this effort has been directed toward the identification of acoustic cues necessary for phoneme identification, and the effect of neighboring sounds upon the final percept.

Haskins Laboratory has been an original and important investigator of acoustic cues. Using the Pattern Playback which generates sounds from spectrographic patterns painted on a moving belt, different patterns may be varied in number and shape of the resonances and the resultant sounds presented for recognition (16). While prominent

acoustic features have been found to be powerful determinants of recognition, it was found that not all features were required, and that the strongest one appeared to be the transition from consonant to vowel and particularly the second formant transition (41, 42). This finding suggests that perception of speech may occur in syllable length units.

Heinz and Stevens (25), using filtered noise tuned to a narrow frequency, conducted a multiple choice experiment using five fricative sounds. While changes in bandwidth did not change the percept, a change in the resonant frequency gave rise to different percepts. When these synthesized fricatives are combined with a vowel, other acoustic cues appear to be important: consonant to vowel intensities and formant transitions. Other studies show that the second formant frequency is associated with the place of articulation (29).

Duration has also been identified as an important correlate of perceived stress (16). Experiments using synthetic vowels in which the vowel length was varied led to the perception of stress for long vowels and also to the perception of different vowels even though the formants remained the same. It may be concluded that although there may be several acoustic cues for the recognition of speech, the acoustics of the speech wave alone does not totally account for recognition.

Linguistic Effects in Recognition

The effect of a listener's linguistic experience is known to be a strong influence in recognition. House et al (30) in a study using naturally produced syllables and synthesized syllables of different degrees of approximation to real speech, demonstrated that learning

was most rapid for real speech and for the synthesized versions least like real speech. Apparently, if the sounds are speech they are identified phonemically. If they do not elicit a linguistic association, they are recognized along the natural parameters of sound.

Linguistic effects appear in differential discrimination as well. In a synthetic consonant-vowel experiment (40) fourteen two-formant syllables were systematically varied to span three different consonants, and subjects assigned the syllables to one of the three consonants. A sharp drop was found in the discriminable differences within the boundaries of each consonant, demonstrating that sounds are classified rather discreetly as phonemes.

A further linguistic effect derives from the meaning involved. We know which sound combinations make meaningful words and we know the grammatical rules which determine which words may follow each other. When recognition of words is compared to nonsense syllable recognition, word recognition is superior (46). Similarly, sentence recognition is better than word recognition. Sentences have grammatical as well as phonemic constraints which serves to reduce the number of alternative choices. That the number of alternatives available is an important influence can be demonstrated by a recognition experiment using vocabularies of different sizes. The results show that as vocabulary size increases, correct recognition decreases (46).

The linguistic effect of meaning upon the perception of speech is clearly one of facilitation (39). It may be that in the perception of running speech, many of the acoustic cues may be ignored and meaning derived from such other cues as intonation and stress. It has been

thoroughly demonstrated that when monosyllabic word recognition has been depressed to fifty percent by filtering, running speech is readily understandable (24). It is possible to distort the speech wave, interrupt it, clip it or filter out portions of the spectrum and still have high intelligibility. Perception of speech apparently uses not only acoustic cues, but linguistic and contextual cues.

Factors in Speech Recognition Measures

Much of the study of speech recognition has taken place within the framework of evaluating communication systems. Within this context, speech recognition measures are known as intelligibility tests. The other area concerned with such measures is clinical audiology, where the goal is the examination of the defective auditory system. In this area, speech tests are termed discrimination tests. However, the same principles apply in both cases.

Six variables may be identified in the use of speech tests: speaker, verbal materials, presentation mode, response mode, transmission system and listener characteristics. Depending upon the nature of the investigation, any one of these may be varied while the rest are held constant. In intelligibility testing, the transmission system is varied, while in discrimination testing it is the listener which is of interest.

Typically in this kind of procedure, a speaker presents a list of syllables, words or sentences to a listener or group of listeners who respond in a written or oral form. The percentage of items correctly identified constitutes the score. The higher the score, the better the

transmission or discrimination. If the distribution of speech sounds within a language is accurately represented, the results should be a realistic test of the system or of the auditory system. This is the model originally developed by Egan (18) and modified for clinical usage. The scores derived must be regarded as relative scores only, since a change in any one of the variables may be expected to produce a change in the results.

The effect of different materials upon intelligibility was investigated by Kryter (38). The intelligibility of words is higher than that of syllables and that of sentences is still higher. This is an expected effect as was seen in the facilitation of perception provided by phonemic and linguistic constraints. Other investigations of materials show that familiarity of words is a variable (48) as is frequency of occurrence (54) and context (46, 59).

The heavy reliance of investigators upon word lists as materials is an attempt to represent speech sounds by frequency as they occur in conversation while eliminating cognitive factors to the greatest extent possible. Monosyllabic words, however, are insufficiently long to include such aspects of speech as stress, inflection, and melody which are normally present in speech.

In an attempt to develop materials of adequate complexity while avoiding the pitfalls of context, Speaks and Jerger (57) used synthetic syntax sentences in which the amount of linguistic constraint and therefore the amount of information conveyed can be controlled. These sentences are constructed by drawing from a pool of the one thousand most common words, words which are independent of each other and

arranging them sequentially. This constitutes a first order approximation to a real sentence. If the words are drawn so that each one might logically follow the one before, grammatical constraints occur. Speaks and Jerger constructed four different levels of approximation to real sentences and tested the intelligibility of these materials. Intelligibility was found to increase systematically with increasing grammatical constraints.

The task presented to the listener in these procedures and the response required are further variables. In the traditional "listen, repeat" technique, the vocabulary is theoretically unlimited for the subject does not know which words will be presented. Possible responses of the listener are therefore unlimited and should provide a measure of what he actually hears. The subject's set is for repeating words however, and legitimate words are usually given as responses. Little information is obtained as to the erroneous perception. The use of highly familiar words also give rise to spuriously correct responses (48).

Multiple choice intelligibility tests, in which the subject has available a closed set of possible responses were developed for the purpose of evaluating both speakers and listeners (1). The advantage of this method is the limitation of possible responses, which renders the linguistic sophistication of the listener of little consequence. It also specifies responses which the listener has rejected and eliminates the ambiguity of response which is found in the "repeat" technique when the scorer is required to interpret responses (27, 38).

More recently, this technique was modified by Voiers (61) as a method of comparing systems for the transmission of certain speech

sound characteristics. In this version, word pairs are used in which only the initial consonant differs. In this way a fricative, for example, can be compared with a plosive, nasal, or sibilant. Voiers analyzed responses by distinctive feature and by place of articulation.

Speaks and Jerger (57) used a matching technique in testing the intelligibility of synthetic syntax sentences in order to obtain a simplified behavioral response. In this case the subject had available all possible responses so the task was one of identification. This procedure permits control of such variables as vocabulary, probability of occurrence, familiarity, length, amount of information and ambiguity of subject response. On the other hand, it does not describe any difficulty of perception which the individual subject may experience.

Clarke (14) investigated the relative value of two response modes in an attempt to derive more information from traditional procedures. In one response mode, subjects were permitted a second guess at the presented word. In the second mode subjects rated the accuracy of their responses from triple plus (absolutely certain) to triple minus (absolutely wrong). In normal listeners, second choice procedures provide no more information than traditional techniques. Confidence ratings, however, significantly increased the information obtained. Confidence ratings have also been shown to be directly related to the accuracy of message reception (51).

In general, matching procedures appear to be a better method of controlling the variables associated with intelligibility tests while confidence ratings increase the sensitivity of such tests and provide some indication of subjective difficulty.

Recently, response time has been investigated as a measure of relative difficulty. Response time has been shown to have a monotonic relationship with accuracy as measured by monosyllabic word lists (24).

Signal detection theorists have pointed out that the listener's criterion is also a major factor in testing the reception of speech (15). In terms of signal detection, a mathematical theory has been worked out involving the setting of criteria, the decision as to detection of a signal and the mathematical probabilities of "hits," "misses," and "false alarms." The theory has been extended to cover the detection of speech although in this area, the theory is not complete. This extension concerns the behavior not only of the receiver (listener) but also that of the source (speaker-scorer) (18). The source transmits a message to the receiver who interprets the message. The receiver may then do one of two things: He may make a judgment as to whether or not his interpretation was correct, or he may send the message back to the source for confirmation. The source may then accept the message as correct, or may reject it. Comparative studies of source and receiver in such an exchange demonstrate that the source, with previous knowledge of the message is more accurate than the receiver without such knowledge. It should be observed, however, that while such relationships hold true in real life situations as between pilot and Air Traffic Control, in the investigation of speech perception, better controls are afforded by obtaining an unambiguous response from the listener.

Clinical Applications

The techniques developed in intelligibility tests have been applied clinically in the examination of the auditory system. If speaker, materials, and transmission system are held constant, it should be possible to evaluate the listener, and compare the auditory capacities of different listeners. Within this context, the ability to repeat back a certain percentage of correct responses has been related to hearing acuity at certain points across the frequency spectrum. For example, it is possible to specify a relatively narrow range of frequency within which a particular speech sound will fall. If the individual has a selective loss of hearing for that range it should be predictable that he will not perceive that sound at certain intensities. If the materials represent the language in the frequency of occurrence of speech sounds, a reasonable estimate of communication difficulty should be derived from such tests.

The usual technique in speech discrimination tests consists of presenting phonemically balanced word lists to the subject at a subjective level representing the level of soft conversation. The recorded lists are presented via earphones to the subject seated in an acoustic chamber. Lists are composed of fifty words equated for difficulty and frequency of usage as well as being representative of the phonemic structure of the language.

Such tests are subject to the same problems of measurement as are intelligibility tests. In normal listeners, intelligence, vocabulary level, age, and ability to predict are some of the factors influencing

test scores (19). Beyond this, the assumption that selective hearing loss will influence scores appears to be erroneous except in a very broad sense. Linguistic factors tend to override the effect of selective hearing loss. The response of the ear to simple tonal stimuli does not always reflect the capacity of the ear to transmit complex stimuli. For this reason, speech tests should provide a more complete estimate of auditory functioning.

In subjects with auditory disorders, discrimination varies although somewhat inconsistently with the site of the lesion (32, 33). Disorders of the outer and middle ear, provided the inner ear is normal, have little or no effect upon discrimination as measured clinically. Damage to the inner ear, if severe enough, usually results in reduced discrimination for speech as measured by monosyllabic word lists. However, if essentially normal hearing is retained up to 2KHz by pure tones, discrimination scores are usually within normal limits even though subjects may report extreme difficulty in communication under some conditions. These results suggest that present discrimination tests lack the sensitivity to reflect accurately the degree of handicap in communication.

Other materials and techniques have been used experimentally as a refinement of discrimination testing, most notably multiple choice tests and synthetic syntax sentences (1, 27, 38, 57). While these techniques are promising, they have not been standardized for clinical use. The application of these tests to pathological subjects indicates that the required contralateral evidence may fail to appear for various reasons.

Expanding lesions which produce increased intracranial pressure and bilateral lesions will result in reduced discrimination bilaterally (52). Aged subjects with no known pathology have reduced discrimination with these tests bilaterally as well (44, 52, 60). Lesions of the temporal lobe which result in aphasia will again show bilaterally reduced discrimination.

Bocca (7) notes that in our present state of knowledge, there is no clearcut differentiation between defective hearing and sensory aphasia in temporal lobe pathology. The difficulty of distinguishing form from experience, that is, discrimination from recognition is a major problem in the development of these tests. Tests which utilize materials with linguistic value have proved to be of greater significance than tonal tests, yet the materials themselves present other problems. Linguistic significance provides perceptual enhancement which may override defects of functioning, while language disorders prevent adequate responses or perception. Beyond this, the findings on ear preference in certain conditions suggest a difference in interaural perception which may exaggerate or minimize differences which may exist between ears as a result of pathology (9).

The development of diagnostic auditory tests to determine the site of a lesion in the auditory pathways has become of increasing interest to the field of audiology in recent years. In several instances, modifications of conventional pure tone audiometry have successfully demonstrated the presence of both cochlear (32, 34) and retrocochlear lesions (13, 33, 34). However, the findings in lesions of the eighth

nerve are not always clear (71), and as the locus of the lesion ascends in the auditory pathway, it becomes increasingly difficult to identify.

Subjects with central lesions (i.e., with insult to brain stem or cortex) frequently demonstrate no significant hearing loss in response to pure tone stimuli. Theoretically, responses to pure tones reflect the type of activity which takes place within the cochlea rather than that of more centrally located structures. In an effort to sample the type of processing performed centrally, attention was turned to speech audiometry. The understanding of speech was viewed as the highest type of activity performed by the auditory system (7). Here too, the results were disappointing. Subjects with central lesions often display normal responses to conventional speech audiometry.

It is a well known fact that speech is highly redundant, and that even with a great deal of information removed, a message may still be understood. This excess of information may be the factor which allows for normal responses by a defective system. Therefore, it may be possible to test the patency of the central system with a speech signal which has minimal information. Theoretically, such a task requires maximum efficiency of the system and should identify subtle defects which are not otherwise detected by conventional techniques. That this is an effective technique has been demonstrated by a number of investigators (2, 3, 5, 6, 7, 10, 11, 43, 45, 52).

Several methods are available for removing the information from the speech signal. Frequency information may be removed by filtering, or phonemic information may be removed by interrupting the signal or

by compressing the signal in time. Speech which has been processed in one of these ways is termed "degraded" speech. The method of degrading does not appear to be of importance, as long as the message is made difficult to understand (43).

The method of presentation, however, does appear to have a differential effect depending upon the level of the lesion. Binaural summation tests split the speech signal and present half to each ear separately to test for fusion of the signal. Such tests appear to be specific for lesions of the brain stem (3, 43, 45). Monaural tests with degraded speech, however, appear to be more effective with cortical lesions (7).

Degraded speech tests for central auditory dysfunction are modifications of tests of discrimination. That is, the conventional technique is based upon presentation of a sample of speech and the requirement that the subject repeat back that sample. The material may be one or two syllable words or sentences. As such, they are subject to the same limitations which apply to traditional tests and techniques. Results obtained on discrimination tests are known to vary with such factors as type of materials, length of the stimulus, the familiarity of materials, linguistic value, presentation style, and response mode. Some of these may have particular significance when the subject has central damage.

Meaningful speech is known to facilitate perception. The use of this type of material may have a differential effect depending on which hemisphere is damaged. Some recent evidence suggests that a difference exists between right and left ears in the ability to perceive verbal stimuli (36, 37), which emphasizes the importance of meaning in audition.

In tests which rely on differences between ears for significance, the interaction between meaningful material and the hemisphere damaged may serve to exaggerate or minimize such differences.

In subjects with aphasia, whether of the expressive or receptive type, an auditory factor has been identified. The extent of the effect of this factor to degraded speech tests is unknown. Certainly, the expressive aphasic cannot participate in the conventional test because of the requirement of repetition. The results of such tests in subjects who may have some degree of auditory aphasia might be expected to be equal bilaterally because of the necessity of interpreting meaningful materials, thus failing to demonstrate the interaural differences required.

In the development of tests for identifying the site of lesion within the auditory pathway, it is necessary to identify the contribution of perceptual factors to interaural test results, and to determine the extent to which meaningful speech affects discrimination of degraded speech.

The theoretical basis for the use of speech tests in the evaluation of central auditory disorders is the complexity of the task performed by the auditory system. Since the reception of speech requires the temporal processing of all the parameters of sound in varied relationships, this might be thought of as the ultimate in auditory performance. Consequently, it might be expected that such material will provoke the type of activity characteristic of auditory processes at the central level. Degrading speech materials has the effect of further sensitizing the auditory system to the parameters of sound which are essential for the identification of meaning in audition.

speech tests to central damage by requiring that the system perform optimally in order to make the discrimination required.

Bocca and Calero (7) state that "the integration of a complex sound pattern begins at the level of the first and second neurons before cognitive processes come into play." However, it can be readily demonstrated that linguistic value has a facilitating effect upon perception. Under some circumstances, auditory tests with meaningful speech may have more in common with language function than with the patency of the auditory system. It may be more appropriate to utilize a non-meaningful complex stimulus and thereby eliminate some of the variables associated with tests of speech discrimination, locus of damage, and interaural differences in speech perception.

The use of a non-meaningful complex stimulus demands a technique which permits a non-verbal response as opposed to the traditional method of repeating the stimulus. A matching task satisfies this requirement and can be used with meaningful speech and non-meaningful stimuli. With this type of technique, comparisons are possible between both types of materials without requiring a verbal response. A matching technique also serves to eliminate some of the variables associated with discrimination testing, such as the linguistic sophistication of the listener.

When a behavioral response other than a verbal response is obtained, other refined indices of discrimination testing become available. Such measures include reaction time and estimates of judgment confidence which have particular relevance and appeal to the development of a complete profile of discriminatory capacity. A two-category judgment task, while enabling the capture of such data, additionally permits comparisons of

various degradation levels. Among these levels and indices one may be optimal for distinguishing between the responses of the normal and pathological subjects.

This study examined (a) the effect of "meaning" upon (b) the perception of degraded speech for (c) right and left ears respectively in (d) normal subjects and in subjects with unilateral pathology.

Problems in Central Auditory Tests

The problems which arise in an effort to devise suitable diagnostic tests for central pathology concern primarily the linguistic nature of the stimulus, the effect of dominance upon perception, the nature of the task, which may preclude response, and the sensitivity of the test to central versus peripheral factors. The facilitating effect of language upon recognition has been discussed elsewhere. Nevertheless, a complex signal is found necessary to elicit the desired information. In addition, a meaningful speech signal has been found resistant to peripheral factors and is desirable for this reason as well. The finding of interaural differences in the perception of this kind of material further contaminates test scores obtained with this kind of signal. It may be possible to use a non-linguistic complex signal in such tests as an alternative. This of course, requires a different response than the usual "repeat" technique. A non-verbal response has the further advantage of being independent of verbal ability, and is explored in this study.

In the present investigation, two types of materials are compared for their effectiveness in eliciting evidence of central damage. Meaningful speech and reversed speech were selected as analogous complex signals, the latter being devoid of meaning.

These two signals have been electronically degraded to nine different levels of quality. Processing was accomplished by an analysis-synthesis technique using a Polymodal vocoder developed by the Air Force Cambridge Research Laboratory (26, 56, 62). In this technique, an acoustic signal is led to a filter bank in which each filter is tuned to a different frequency, spanning the frequency spectrum from 0 to 5000 Hz. The intensity of the signal in each filter is quantized and synchronized. In addition, a pitch indicator identifies the fundamental frequency and a decision is made as to whether or not voice is present. The measurements are made at the rate of forty times a second and converted to digital data which completes the analysis. The data so obtained are applied to the synthesizer which reproduces the original signal. The synthesizer consists of an identical filter bank, a buzz source and a hiss source. The buzz, automatically tuned to an appropriate fundamental frequency, produces the equivalent of voice, while the hiss is used for unvoiced sounds. These are applied to the filters at the measured intensities in the analyzer and summed for the output.

The quality of the synthesized signal in this system depends upon the quantizing step in the analyzer. When many quantizing steps are used, the measurements are more accurate than with few steps and the resulting output resembles the input more precisely. This requires a higher bit rate for transmitting the data and a wider bandwidth in communications. With fewer quantizing steps, a lower bit rate and a narrower bandwidth may be used at the sacrifice of quality. The nine levels of quality represent the systematic reduction in the number of

quantizing steps. An exception is the "best quality" which is measured in analog fashion and does not suffer from the degrading effects of quantizing.

The intelligibility of these materials has been measured by Voiers (61) using the Modified Rhyme Test. Intelligibility of monosyllabic words varied systematically from approximately ninety-six percent for the best quality to eighty percent for the poorest.

Since the usual verbal response is precluded by the nature of the non-meaningful signal, an alternative technique is required. In this study, the best and poorest levels of quality serve as anchors for the quality continuum to which the intervening levels are assigned. Since it has been shown that the ear is capable of many more differential discriminations than are used in a discrimination, this technique should yield a more sensitive measure of discrimination differences between individuals. The technique has the further advantage of permitting a non-verbal response in assigning the signal to one of the anchors.

Three measures are of interest: the choice of response category for each of the levels, the number of references to anchors, and reaction time. Referring to the anchors for purposes of comparison may be viewed as a measure of confidence, with low confidence associated with high frequency of reference. The use of confidence ratings has been shown to provide more information in intelligibility tests (51) and may be expected to increase the sensitivity of the technique. Reaction time has also been shown to be directly related to message intelligibility (24) and is a further refinement of test scores.

In demonstrating discriminatory failures in central auditory defects, the clearest results are obtained with unilateral involvements, where one ear is normal and the ear contralateral to the damage shows reduced discrimination. In such cases, reduced discrimination cannot be accounted for on the basis of such variables as attention, memory, or intelligence or language defect which may be present but indistinguishable from true higher level involvement. Hemiparesis offers objective evidence of unilateral central pathology and provides an ideal situation for evaluation of central auditory dysfunction tests.

In this investigation, some of the principles developed in intelligibility testing, discrimination testing, and auditory perception are applied to the problem of identifying central auditory dysfunction. A matching technique, which serves to control the variables associated with discrimination/intelligibility testing is used together with measures of response time and confidence ratings to obtain maximum information.

The Psychophysical Technique

The usual technique in discrimination testing is one which requires verbal repetition. The use of backward speech precluded the traditional "repeat" technique. A two-category judgment, or forced-choice technique is an attractive alternative which allows for obtaining similar data without imposing a verbal response requirement. The psychophysical method was modified by providing the subject access to two extremes of the quality dimension, or anchors, for purposes of comparison prior to his final response.

The task not only reflects recognition ability, but also contains a means of assessing judgmental similarities. Of particular value, the technique permits the judgment of a complex multidimensional signal by a relatively straightforward and simple procedure, since it does not require E to detail for the subject the dimension(s) along which the judgment will be made. Stimuli, in general, may be considered similar to the extent that they may be confused. A confusion matrix of resultant judgments will, then, permit an estimate of perceived similarities. This technique has been previously described by Mostofsky and Green (47).

The differential use of the anchors may be viewed as a measure of confidence, which, together with decision time may more conclusively illuminate the understanding of speech perception for both the normal and pathologic auditory system (35).

Four response measures were obtained for each trial: total response time, corrected decision time, decision category, and frequency of anchor use. Response time has been shown to be directly related to accuracy of message reception (23) and to confidence in the judgment (35). Anchor testing may be viewed as an indication of uncertainty, or even organismic characteristics (22). Such indices of uncertainty in intelligibility studies have provided more information regarding the individual judgment than does the decision alone (17). All events were identified and recorded together with their time of occurrence on magnetic tape. A schematic of a single test trial as would be represented by an operations recorder is presented in Figure 1. This figure also illustrates the event and inter-event relationships which were available for analysis.

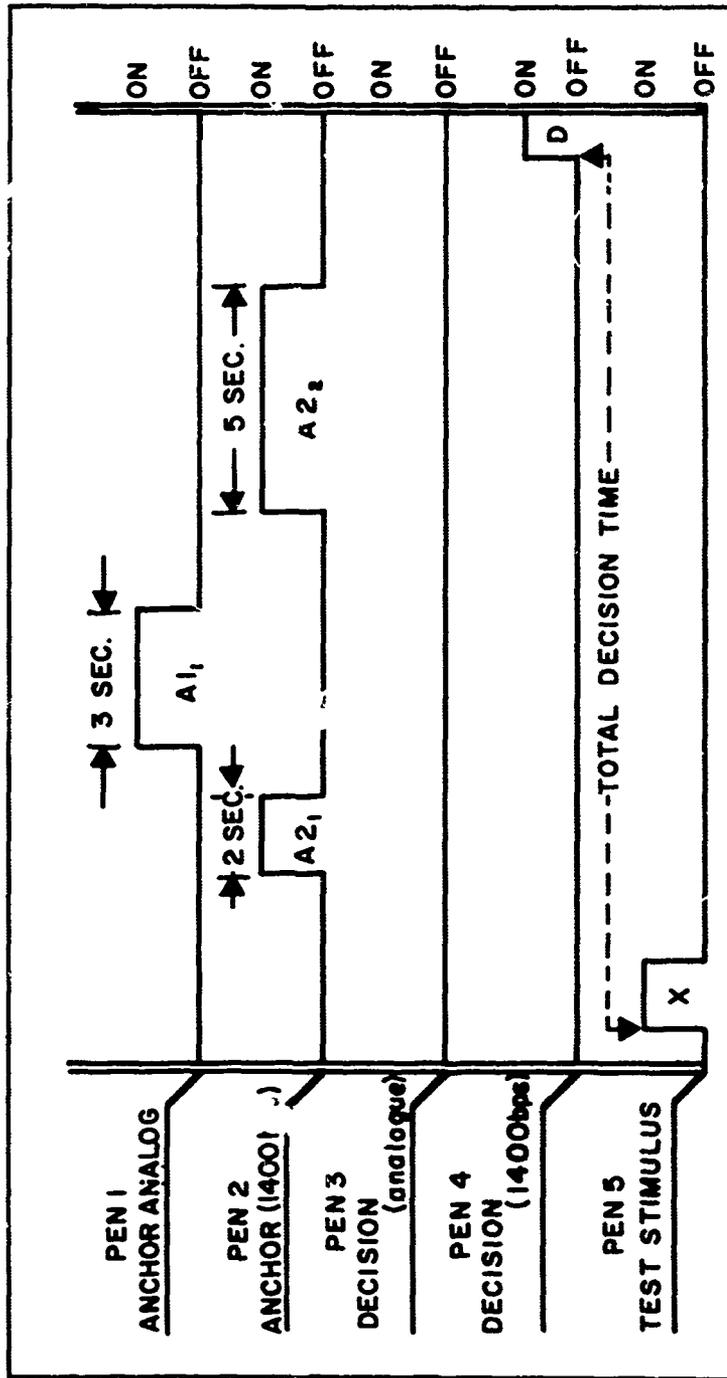


FIGURE 1

SCHEMATIC EVENT RECORDING OF A SAMPLE TRIAL. CORRECTED DECISION TIME IS OBTAINED BY SUBTRACTING THE DURATIONS OF ANCHOR TESTS A₁, A₂, .. etc. FROM THE TOTAL DECISION TIME

Method of Procedure

Two groups of subjects participated in the study: normals (n=33) and left hemiplegics (n=10). All subjects were recruited from the student body at Boston University and from the Outpatient Clinic of the Veterans Administration Hospital. Normal hearing was required of all groups which was defined as hearing levels of not more than 20 Db for pure tones at 250, 500, 1K, 2" and 4 KHz. All normals were right-handed, while handedness in the pathological groups was dictated by pathology. There were no age, sex or I.Q. requirements. Data are presented for those subjects for whom an entire error free test protocol was available. These subjects represent less than half of the original sample.

Each subject individually participated in a task using a modification of the method of constant stimuli. Two response categories were available to S for judging the subsequent stimuli. These categories were defined by the extreme values of the test stimuli. A modification of the standard technique was effected, which consisted of additionally providing S access to the end values (anchors) of the judgment continuum. S could not, however, listen to (a) both anchors at once, nor (b) to either anchor during an intertrial period, nor (c) to either anchor during a stimulus presentation.

Two types of materials were used in the procedure: speech and reversed speech. All materials had been processed by the AFCRL Polymodal Vocoder to achieve nine different levels of quality as described in the preceding section.

The stimulus materials for forward speech consisted of recorded sentences of approximately 1.5 sec. duration. The anchor tape for this

series carried continuous text recorded by the same speaker and processed at quantizing levels representing the best quality (analog) and the poorest (1400 bits per sec). Those also represented the decision categories. For reversed speech, the identical sentences and text were used for stimulus and anchor tapes respectively.¹ Recordings were made with sound peaks occurring at -3 Db relative to a 1KHz tone. The test sentences and connected speech appear in Appendix A of this Section.

In preparing the stimulus tape, ten samples of each vocoding level were recorded for each type of material. Each sample was then re-recorded in random order, subject to the single constraint that each level be represented once in every block of ten. The samples were then re-recorded in this random order, and appropriate tapes made for each type of material.

Test procedures were carried out at the Veterans Administration Outpatient Clinic in Boston. All procedures took place in an acoustic suite (IAC Model 1401). The subject was seated in one room of the suite containing earphones and the response panel. All other equipment was situated in an adjoining room.

The experimental tape was presented via Magnacord 1201 through logic and control network to a Grason-Stadler speech audiometer (162) with telephonic earphones (TDH 39) mounted in MX/AR-41 cushions.

¹The consideration of music as a third stimulus dimension was rejected because of a differential effect of vocoding upon music as opposed to speech material.

The anchor tape was carried by a two channel tape recorder (Wollensak 1580) via logic and control systems through the same speech audiometer. On the subject's response console two toggle switches were mounted ten inches apart and labelled "Test" and "Decide" respectively.

Switches are normally at rest in the neutral position. Switch positions for "Test" were labelled "Test 1" and "Test 2" for level IX and level I respectively. The listener would then hear the poorest quality or the best quality respectively if the switch were thrown during the period following termination of the stimulus and prior to a decision. The "Decision" switch was used to indicate the decision category by corresponding placement to the "1" or "2" position.

Recording Dependent Variables

Time, anchor activation, and decision were automatically recorded on an incremental digital tape recorder (Kennedy) which was part of a general data acquisition system. A more complete description of this system will be found elsewhere in this report. Essentially this system can accept up to ten independent events and record them by identification code and time of onset in a nine track IBM compatible format. The system incorporates a time clock with a time resolution ranging from one millisecond to one second. This study was conducted with a ten millisecond resolution.

A voice relay within the system detected the onset of the stimulus presentation. The output of this trigger was recorded as an event beginning a trial and which effectively initiated part of the logic preventing the use of the anchor values during this time. The beginning and termination of each use of the respective test (anchor referral)

switch was subsequently recorded. Choice of response category, signalled by the position of the "Decision" switch also signalled the end of a trial. A block design illustrating the arrangement of equipment and rooms is schematized in Figure 2.

Experimental Procedures

All subjects individually participated in a two category task with available anchor referents. The subject was seated in the test room and generally acquainted with the nature of the experiment. He was then instructed in the use of the response panel. The anchor switch was demonstrated and S was told that these were reference values of the sound signal under study. Next, intermediate values were presented and S was instructed to decide for each presentation the best assignment, given only these two anchors. The decision switch was demonstrated for indicating the decision. S was told that the anchors could be used as often as required to help him make a decision. With minor variations, the following instructions were given to each S.

"You are going to take part in an experiment which has to do with listening to different kinds of sounds and judging whether they are alike or different. You have two switches here. This one (indicate anchor switch) allows you to listen to two distinctly different examples of the same kind of sound. (Demonstrate with phones in place.) I am going to present to you, one at a time, other examples of the same kind of sound. You must compare each example with these two (indicate) and decide whether it sounds more like Test 1 or like Test 2. You may use this switch as often or as little as you need to make a decision. Let me know how you decided by using this Decision switch. If you think it sounds like Test 1, press Decision 1. If it sounds more like Test 2, press Decision 2. Any questions? I am going to have you practice a few times before we actually begin." (after practice trials) "Are you ready?"

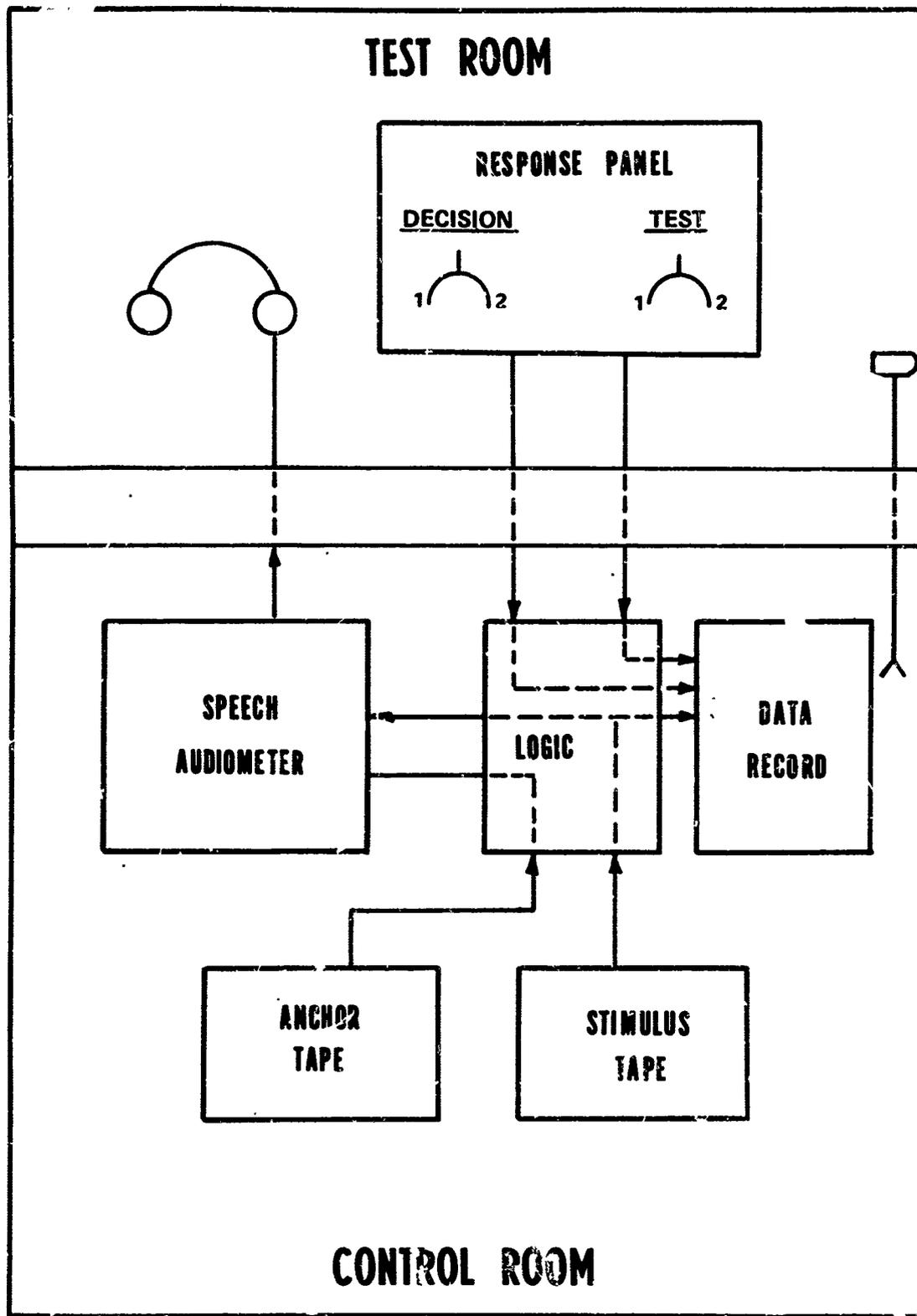


FIG. 2 BLOCK DIAGRAM

Two classes of data were analyzed: decision time and frequency of anchor use. The subscript for each measurement (X_{abcdi}) denotes the following:

<u>Factor</u>	<u>Name</u>	<u>Level</u>	
a	Material type	2	forward speech reversed speech
b	Ear tested	2	right ear left ear
c	Pathology	2	normal hemiplegic
d	Level of degradation	9	1. analog 2. 3200 bps 3. 3000 4. 2600 5. 2400 6. 2300 7. 2100 8. 1700 9. 1400
i	Replications	5	_____

Results

The three measures of anchor referral frequency, corrected decision time, and decision category obtained were summed over replications and the means calculated for each group, each ear, each type of material, and each level of quality. In the case of decision time, this represents the time from the onset of the trial to the time of decision, less the time spent in testing the anchors. Frequency of anchor use represents the mean number of times either anchor switch was used at each quality level. Response category decision represents the mean frequency with

which the respective stimulus was assigned to either category. Response category I is the poor quality anchor and response category II is the high quality anchor.

Figure 3 illustrates mean corrected decision times for the normal group in the reversed and the standard speech tasks calculated separately for each ear. While neither differences in quality levels nor in ear responsivity appear, an obvious and consistent increase in decision time is seen for the series involving the reversed (meaningless) speech. Figure 4 shows the same data with scores for the right and left ears pooled where the differences in materials are now pronounced and systematic. The resultant curves are relatively flat, with only a slight tendency toward a rise in decision time toward the extremes of the continuum. In both cases the curve for reversed materials is well above that for the standard speech task. However, there is no particular effect of increasing degradation upon the response time for either type of material.

Decision times for the pathological group for both types of materials are summarized separately for each ear in Figure 5. It is readily apparent from this figure that response times are longer than for the normals in both ears and for both types of material. The relative flatness that was characteristic of the normal subjects' curve is not apparent. In addition, a rather wide variation and overlap in decision time over the quality continuum will be noted. However, for both types of material, longer response times were obtained for the right ear. This is better seen in Figure 6 which illustrates the same data averaged over three adjacent points on the quality continuum.

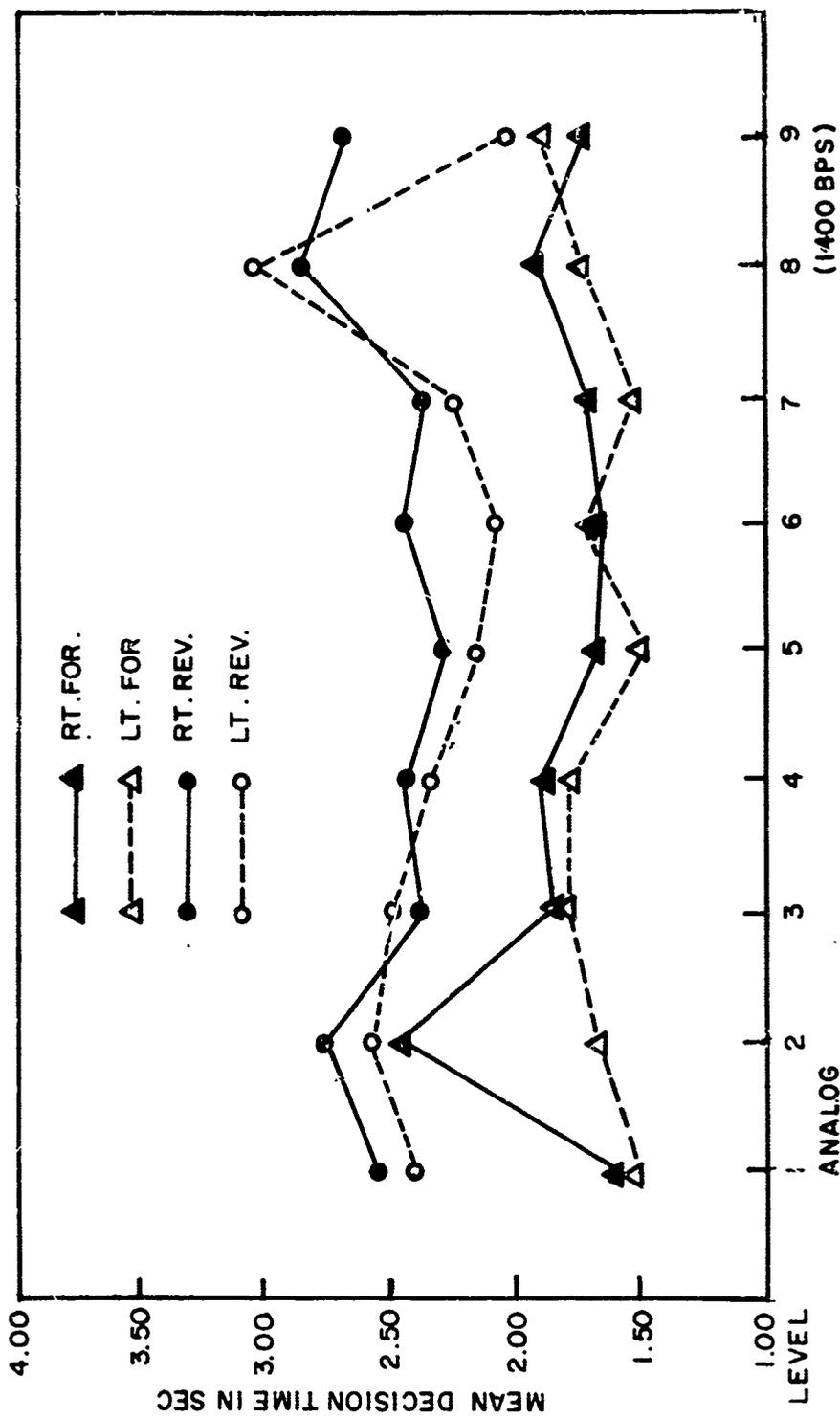


FIGURE 3
 MEAN DECISION TIMES AT EACH LEVEL AND
 FOR STANDARD AND REVERSED SPEECH (NORMAL)

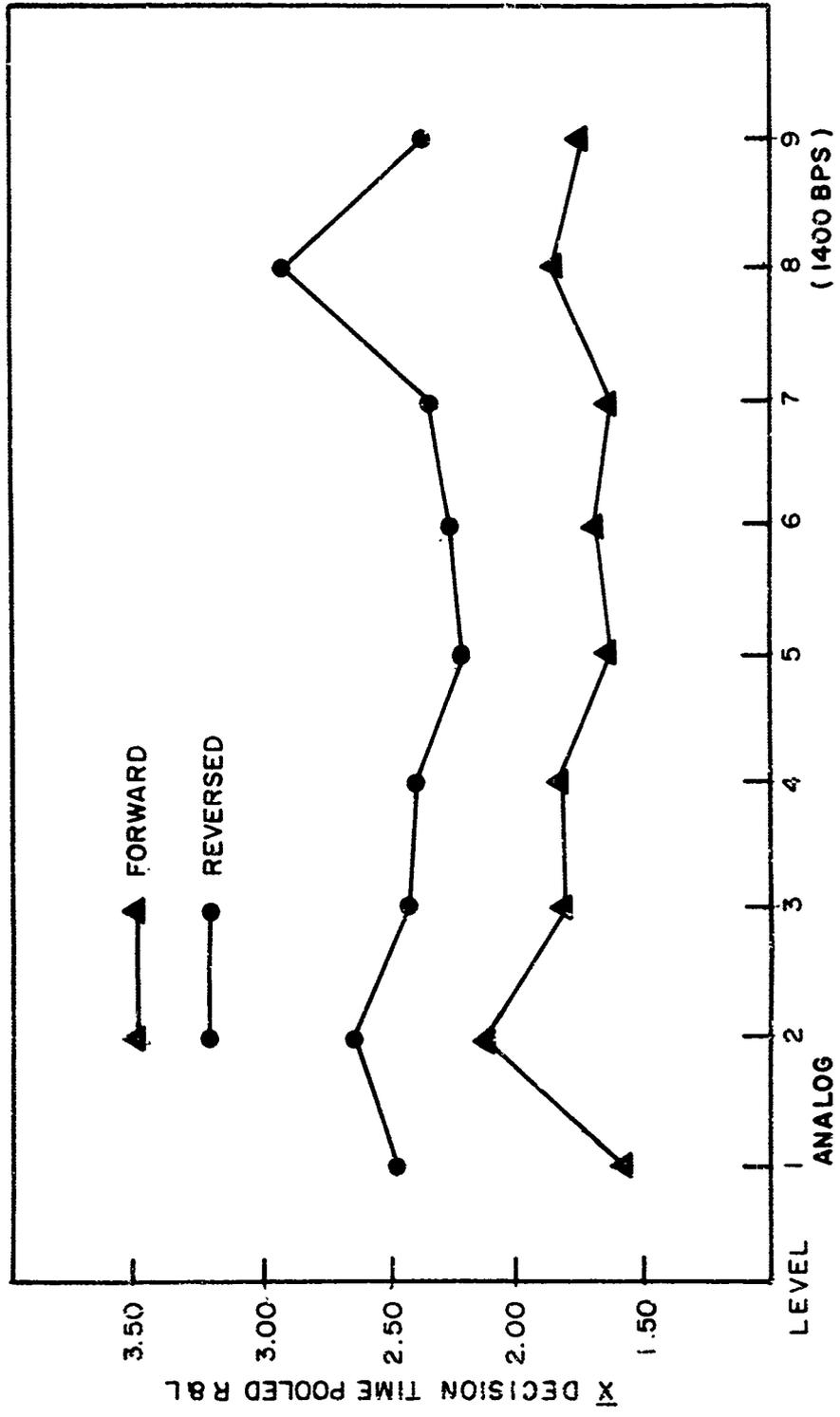


FIGURE 4.
 MEAN DECISION, TIMES LEVEL (EARS POOLED)
 FOR STANDARD AND REVERSED SPEECH (NORMAL)

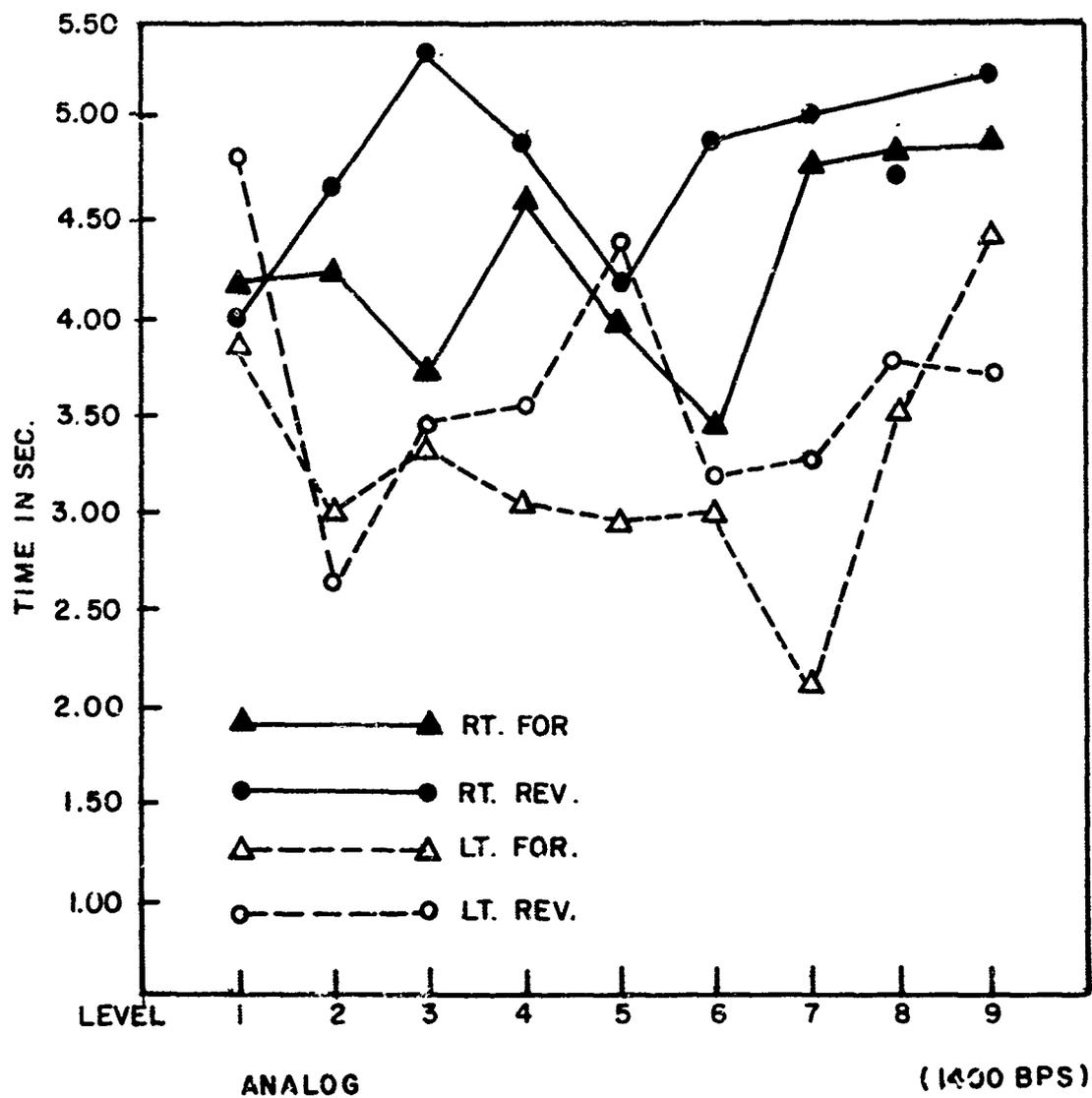


FIGURE 5
 MEAN DECISION TIMES AT EACH LEVEL AND EACH EAR
 FOR STANDARD AND REVERSED SPEECH (HEMIPLEGIC)

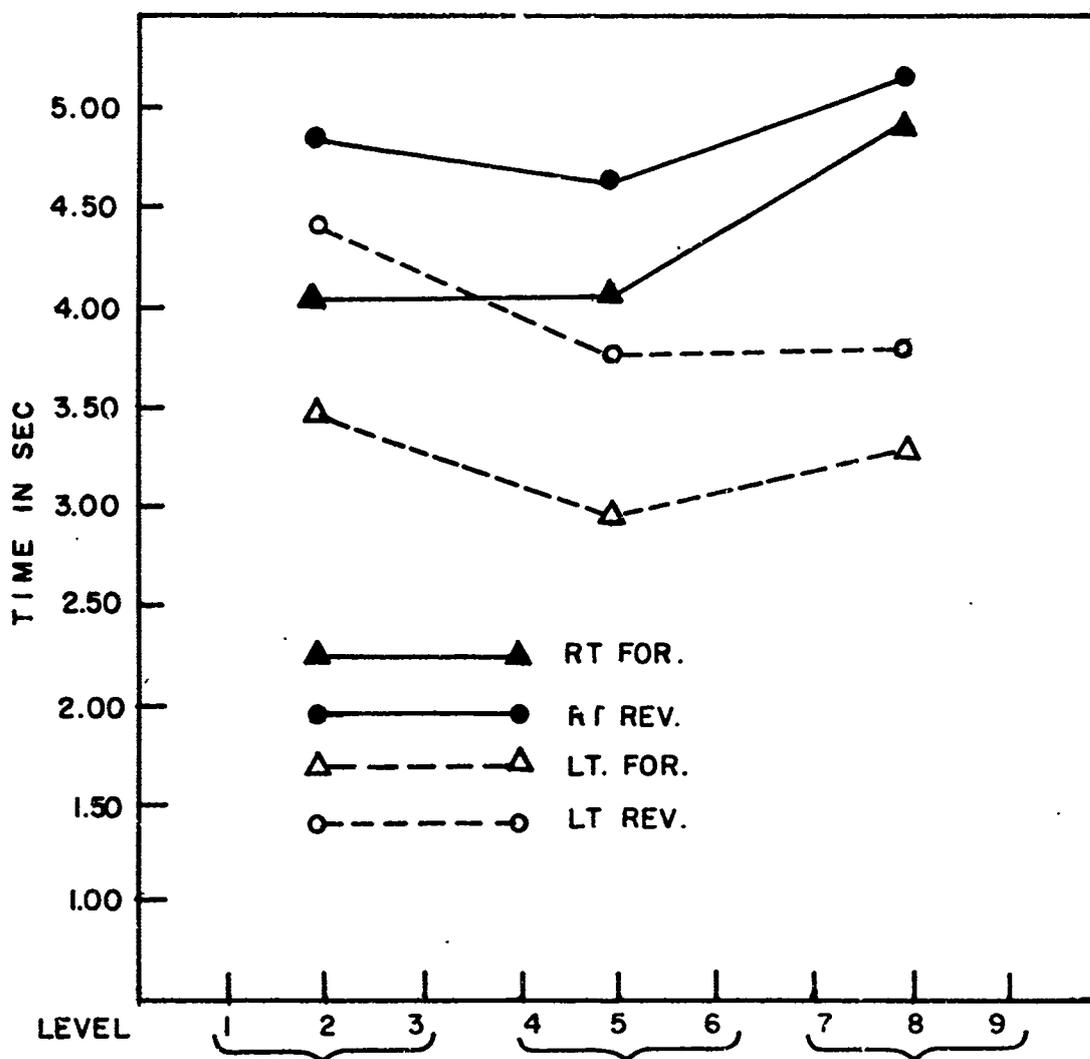


FIGURE 6
 MEAN RESPONSE TIMES AVERAGED OVER THREE LEVELS
 FOR EACH EAR FOR STANDARD AND REVERSED SPEECH (HEMIPLEGICS)

Decision times are longer for the reversed speech materials, as was the case with non-handicapped subjects. For each type of material, however, the right ear has consistently higher decision times. In addition to response latency, the frequency of anchor usage was recorded. It may be viewed in this study as a measurement of attention, uncertainty, or confidence judgment. Figure 7 illustrates the mean frequency of anchor usage. The figure is plotted against each quality level for each ear and type of material for the normal group.

Again, the effect of increasing degradation of the signal appears to have no systematic effect upon this measure. It seems clear, however, that the nature of the material has a differential effect upon anchor usage as well. For each of the ears, anchor usage increased with reversed speech, although the difference is somewhat more pronounced in the right ear. Again, the curves are relatively flat and do not offer any evidence of systematic judgment difficulty as related to the continuum.

In the pathological group, Figure 8, a generally higher rate of testing is apparent except in the condition of standard speech in the left ear. A higher rate of anchor use is associated with the reversed speech task, and a clear separation of right and left ear performance can be seen. A noticeably higher rate of testing occurred for the right ear in both tasks.

The third response measure involves the assignment of each of the stimulus values to one of the test anchors: i.e., the category decision. Figures 9 and 10 show the frequency of assignment to

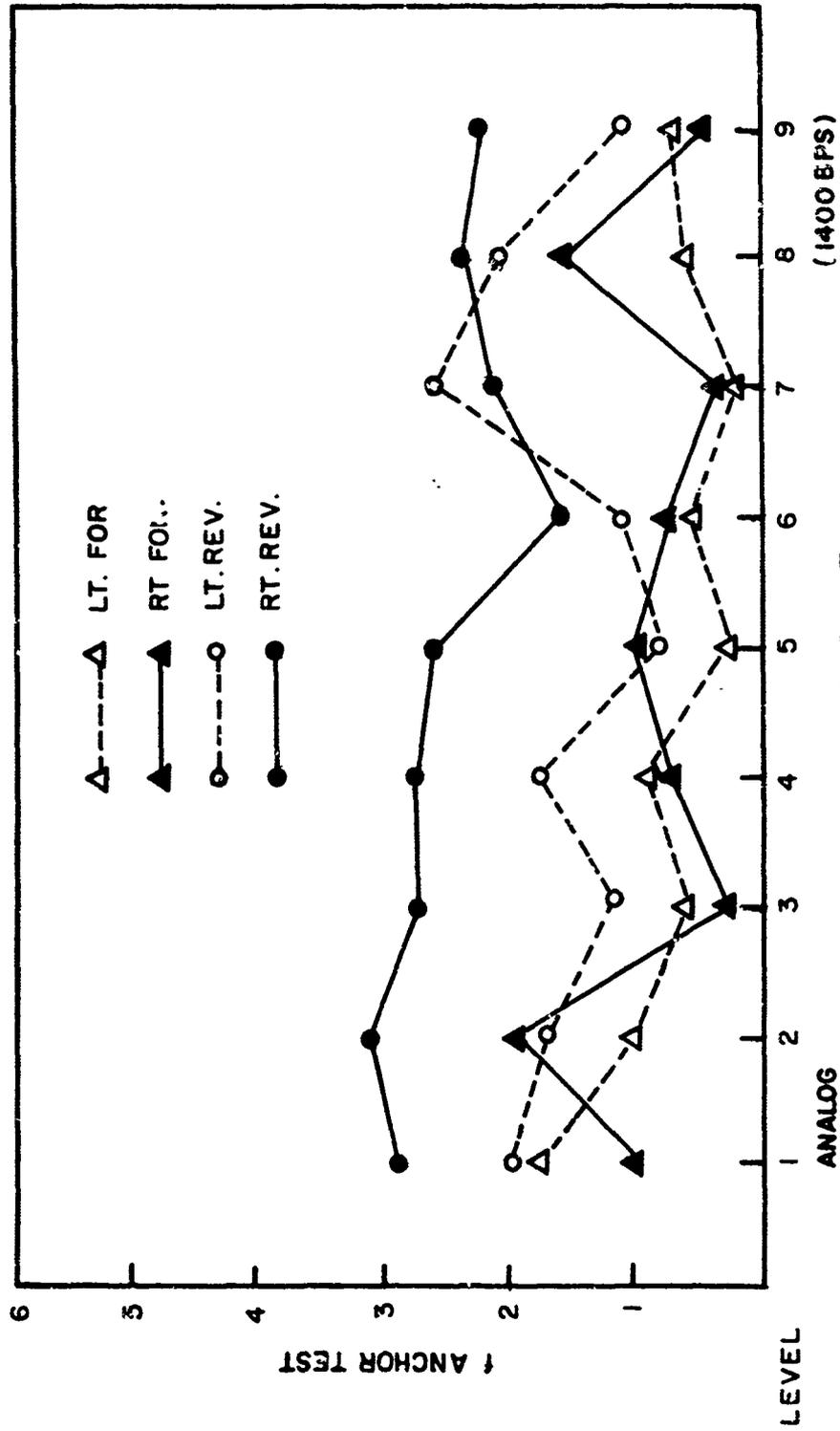


FIGURE 7
 MEAN FREQUENCY OF ANCHOR TESTS AT EACH LEVEL,
 AND EACH EAR FOR STANDARD AND REVERSED SPEECH

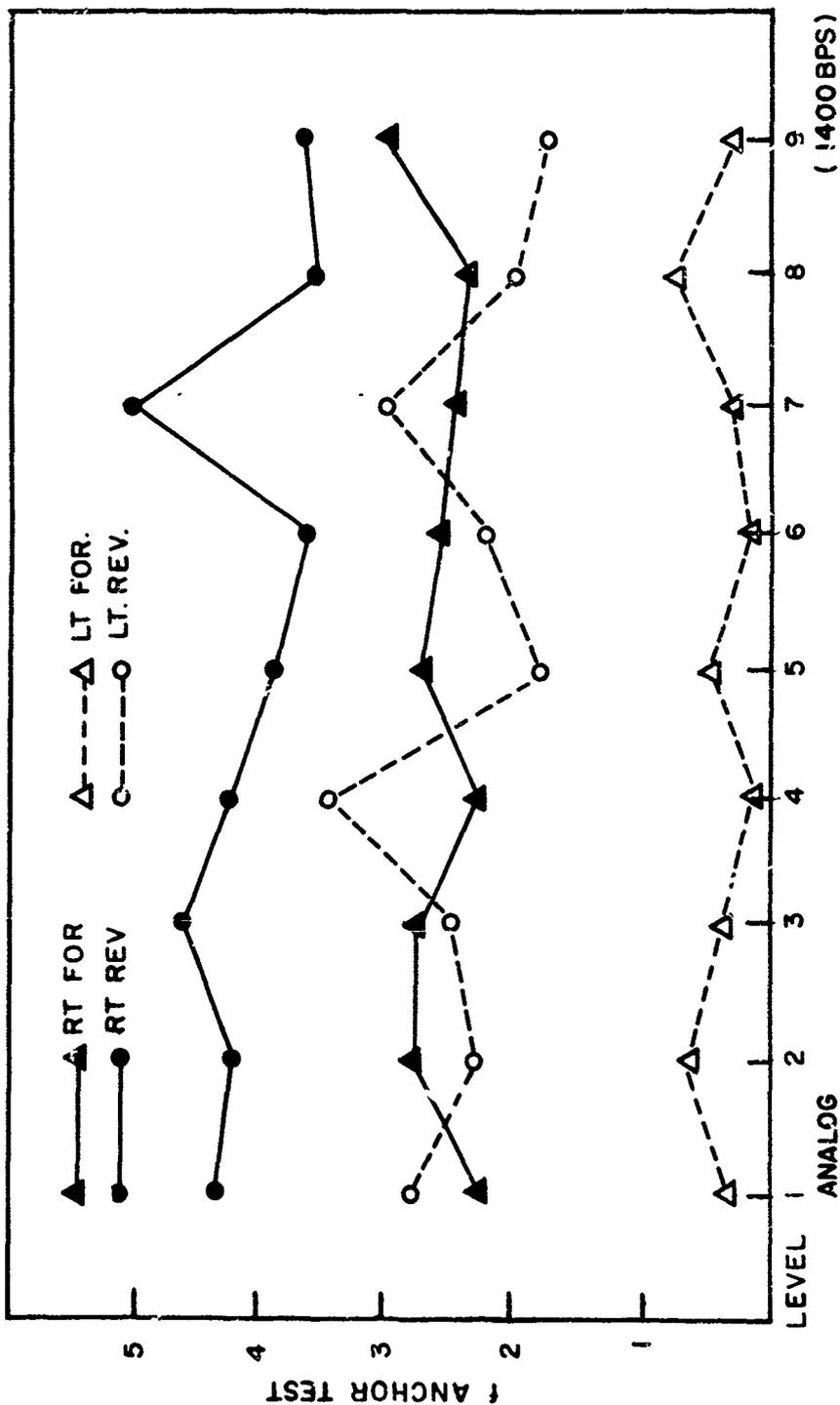


FIGURE 8
 MEAN FREQUENCY OF ANCHOR TESTS, AT EACH LEVEL
 AND EACH EAR FOR STANDARD AND REVERSED SPEECH (HEMIPLEGIC)

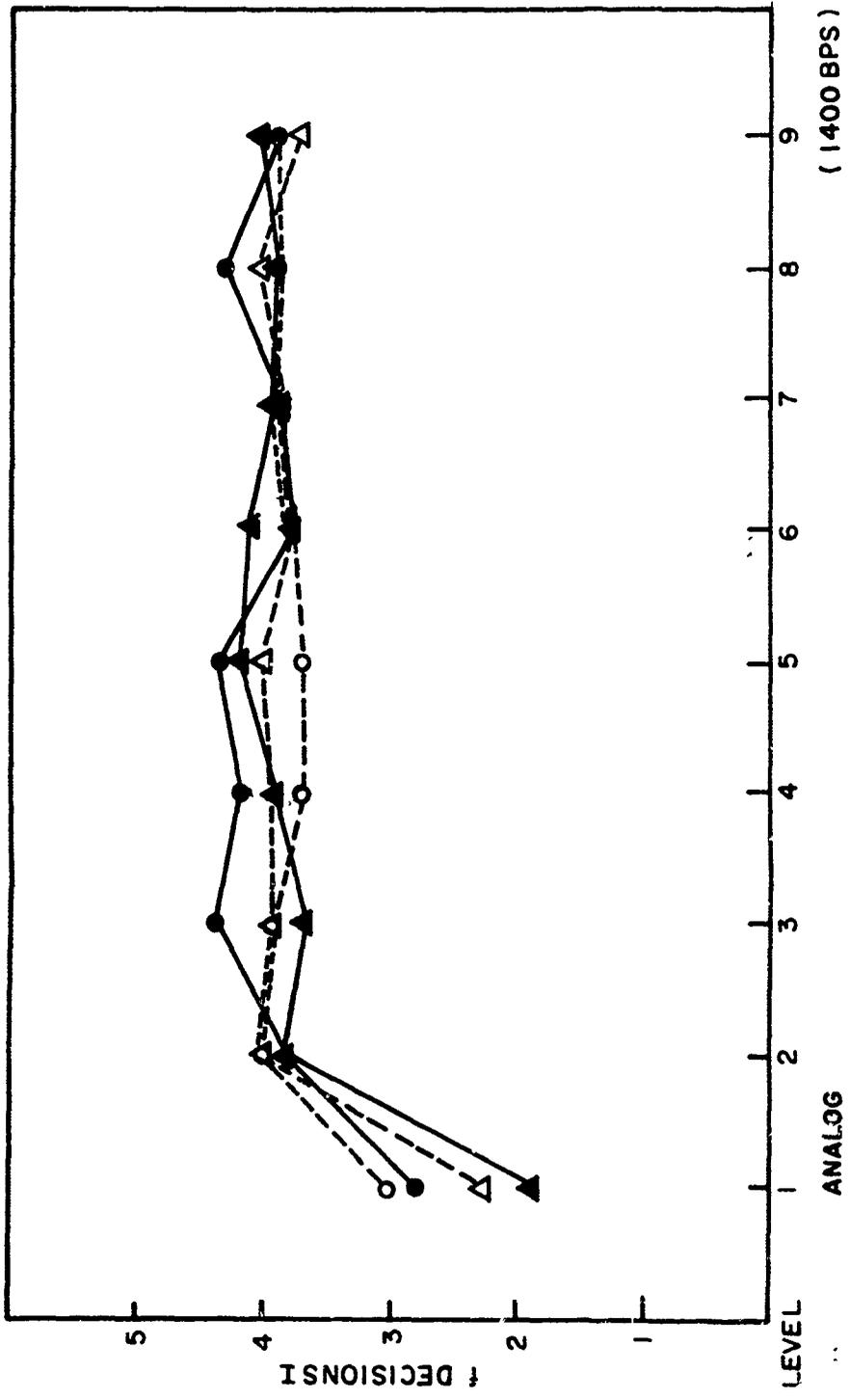


FIGURE 9
MEAN FREQUENCY DECISION CATEGORY I
AT EACH LEVEL FOR EACH EAR FOR STANDARD AND REVERSED SPEECH (NORMAL)

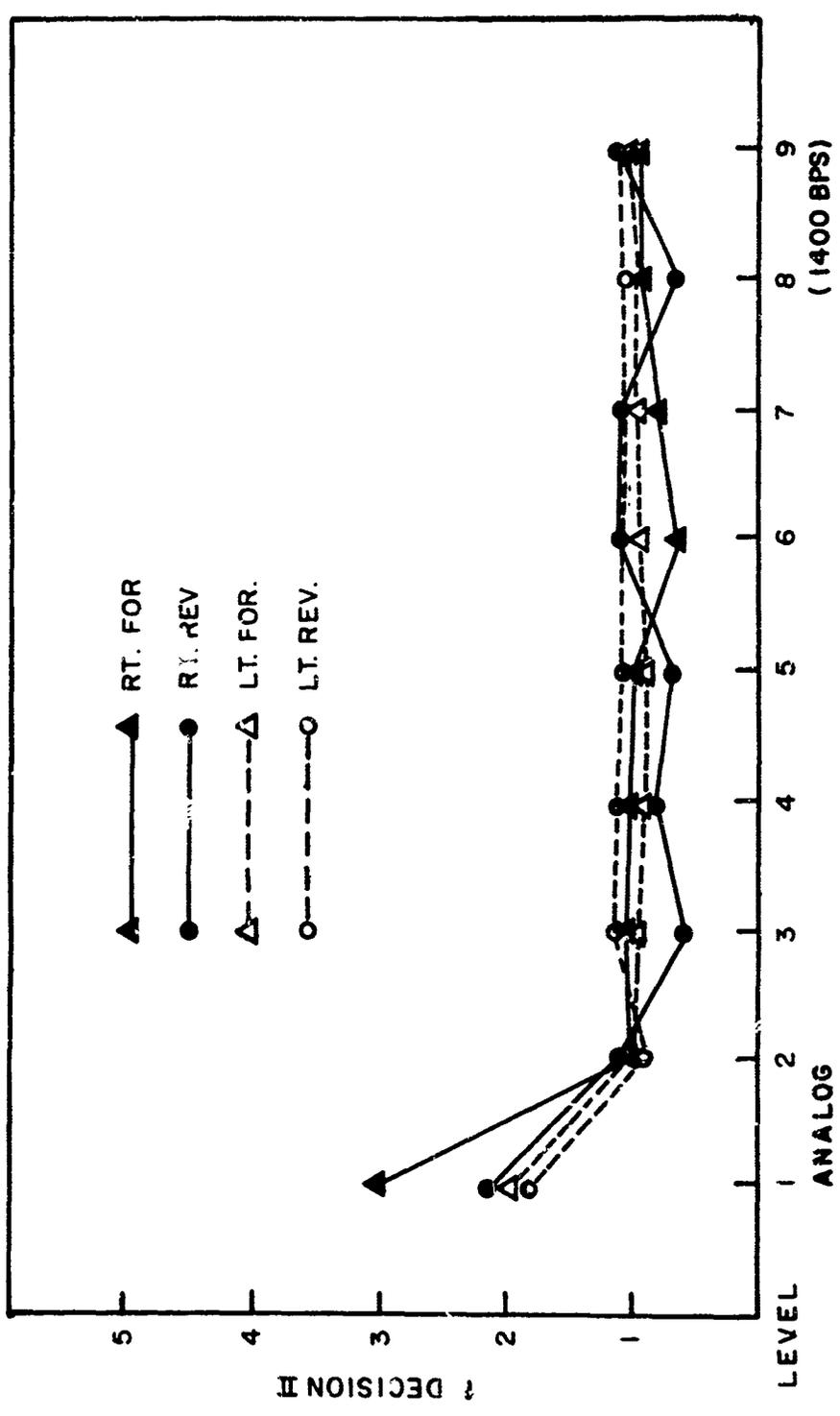


FIGURE 10
MEAN FREQUENCY DECISION CATEGORY II,
EACH LEVEL AND EACH EAR, FOR STANDARD AND REVERSED SPEECH (NORMAL)

category I (poor quality) and category II (best quality) respectively, for the normal group.

While no real difference appears between ears or tasks, the choice of response category is strongly skewed in the direction of category I. Both figures clearly show the absence of the typical ogive distribution of responding. Neither does there appear to be any evidence of systematic control exerted by quality differences. It appears that only the analog samples were categorized as "Decision I" and all others were indiscriminately judged to be best labeled as the poorer, "Decision II."

The assignments of stimuli to response categories I and II by hemiplegic subjects are shown in figures 11 and 12. In this case, the curves are relatively flat, and the increasing quality of the stimuli along the continuum is again not reflected in an increasing number of assignments to category II. Little difference appears between the tasks, but a difference between ears is again apparent. More category II responses occur when the stimulus is presented to the right ear. Presumably, auditory signals in the left ear are perceived as poorer quality than in the right ear. However, the frequency of assignments to category II is higher in both ears for the pathological subjects. This, along with the generally higher rate of anchor use, may simply reflect the difficulty of the task for the pathological subjects.

Decision time and frequency of anchor testing tend to confirm each other when the nature of the material is considered. For all subjects, the meaningless reversed speech has the effect of increasing

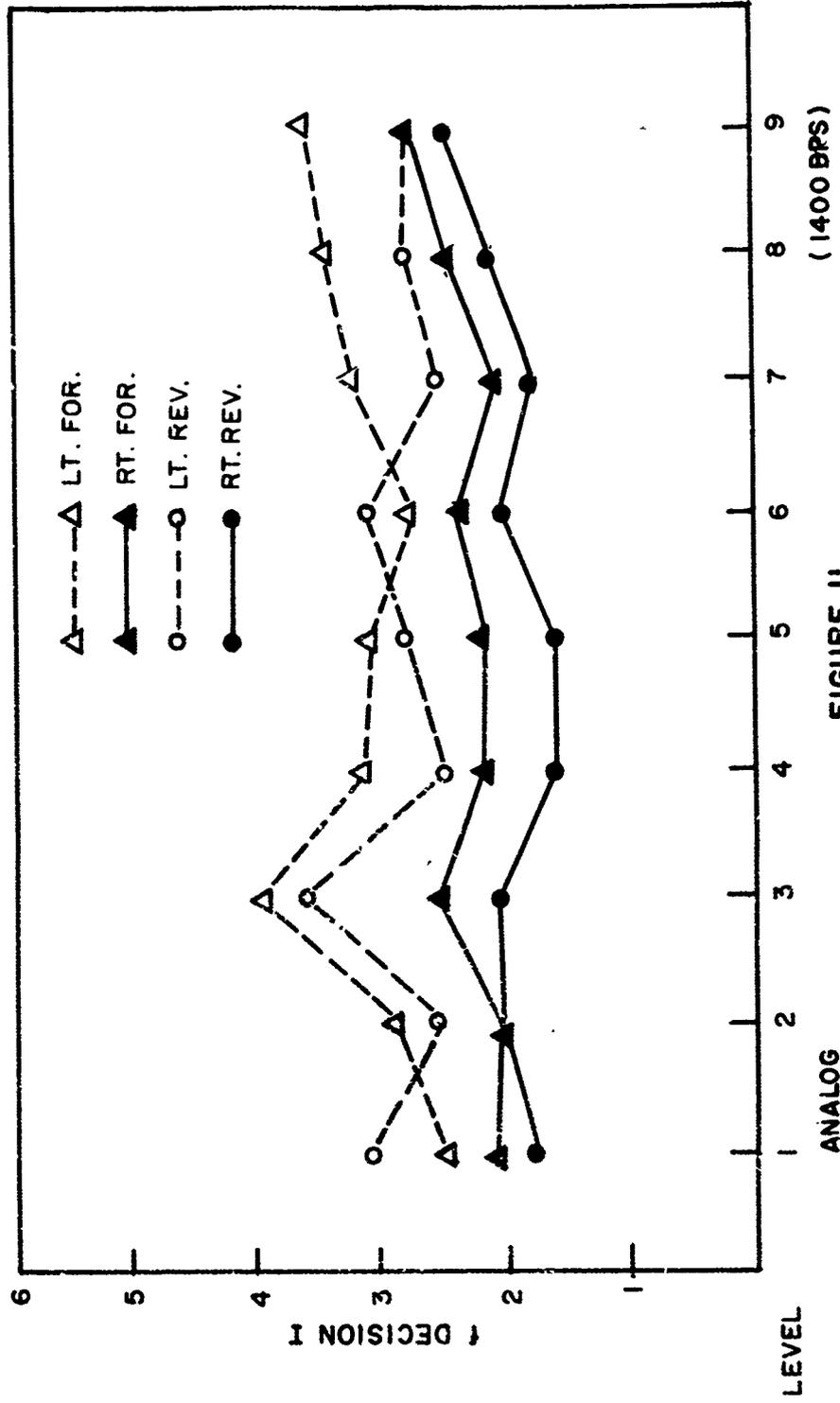


FIGURE II
MEAN FREQUENCY RESPONSE CATEGORY I AT EACH LEVEL
FOR EACH EAR FOR STANDARD AND REVERSED SPEECH (HEMIPLEGICS)

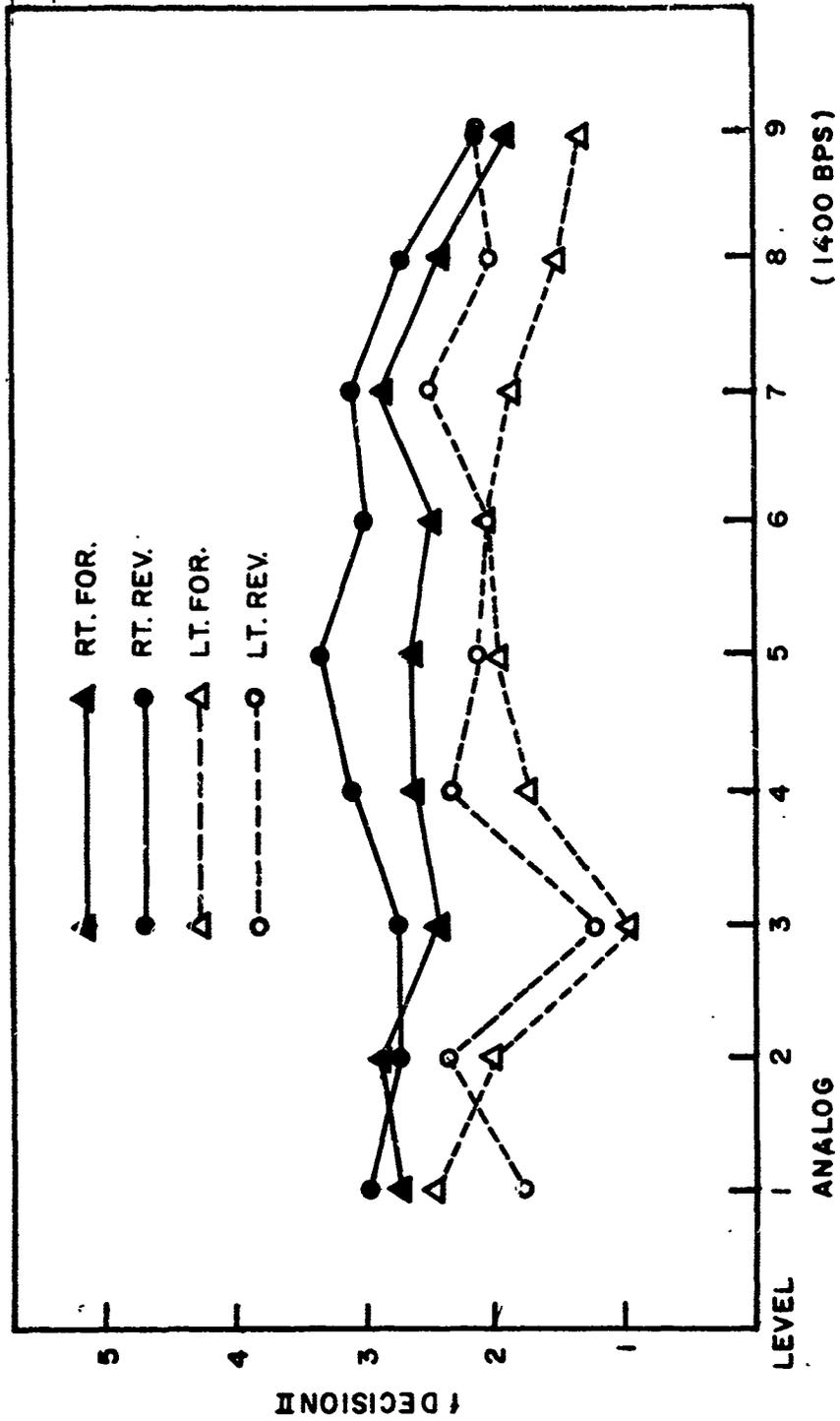


FIGURE 12
 RESPONSE CATEGORY II FOR ALL LEVELS
 EACH EAR, HEMIPLEGIC, STANDARD AND REVERSED SPEECH

decision times and the frequency of reference to the anchors. Assignments to the response category were not affected by reversing speech. This may reflect more of the attention-getting qualities of the stimuli than preferential or intelligibility attributes.

In normals, responses for all measures were essentially the same for the right and left ears respectively. A distinct difference between ears is, however, demonstrated by increased decision times and anchor references in the right ear. More assignments to response category II are also seen in the right ear for the pathological subjects. The curves for response category, however, do not reflect the decreasing quality level from level 1 to level 9.

Discussion

A priori hypotheses advanced earlier would have expected:

(1) differences between hemiplegics and normals in the perception of degraded speech, (2) that such differences would be related to the meaning value of the materials, and (3) that a significant interaural difference would be demonstrated in the hemiplegic subjects.

In the analysis of responses to degraded speech, a clear difference was indeed found between normal and pathological subjects, but this difference was not limited to performance under conditions of the reversed material. Apparently, the degrading effect of vocoding is sufficient to disturb the perception of meaningful as well as non-meaningful materials. It might have been thought that the effect of meaning would be sufficient to nullify the effects of degrading even in

the defective auditory system. In fact, the effect of meaning seems to permit judgments to be made with shorter latencies and with fewer references to the anchors, at the serious expense of accuracy. For all subjects, the presentation of non-meaningful material resulted in longer response times and in increased reference to anchors.

For all measures with pathological subjects, a distinct interaural difference was in fact obtained, with longer response times and more anchor tests when the material was presented to the right ear. From the measure of response category for the right ear, where more stimuli were assigned to the better quality anchor, there is an implied difference in the perceived quality of auditory signals between the two ears. Anecdotally it may be noted that some of the subjects did, in fact, report a subjective preference for the right ear. When this is considered, the longer response times and increased use of anchor references would seem to be contradictory since this suggests a corresponding increased judgmental difficulty. The more parsimonious explanation, however, is that the shorter response times and fewer anchor tests reflect the inability to make the discrimination. Some subjects complained that all the stimuli sounded the same when heard only in the left ear, which would tend to corroborate these findings.

Some other aspects of responses to the non-meaningful material are not represented in the data. For example, most of the subjects thought that the reversed speech was a foreign language, or speech that had been processed electronically. The quality of this material was subjectively felt to be of poorer quality than the standard speech, although this is not apparent in the choice of response category. They also reported

more difficulty in making the judgment. Again, this is clearly apparent in the increased response times and the increased frequency of anchor tests. There was also a tendency for some individuals to refer to the anchors throughout this part of the session. For the standard speech task, the tendency was to refer to the anchors only at the beginning of a session, and to rely on memory thereafter. Memory for non-meaningful materials is, of course, considerably poorer than for meaningful material. In the pathological group there was some evidence of memory deficits which might account for the increased use of anchors over both ears. This may be confirmed by the comparison of anchor tests by normal and pathological subjects.

Somewhat unexpected were the results obtained in the data which describe the selection of the decision or response category. While it was expected that the stimuli at the extremes of the continuum would be uniformly assigned to the appropriate test category, contrasted with those in the middle which would be distributed quasi-gaussian, this did not occur. There was a strong tendency for the normals to assign all but the highest quality stimulus to the "poor quality" category. Subjects tended to set a high criterion for good quality speech. It is possible that subjects used their own memory of speech as a criterion. This is not likely, however, with the reversed material for which choice of response category does not differ from the forward speech samples. The finding is more probably related to yet other unexplored stimulus characteristics.

In the processing of the materials, all of the quality levels except the "best" one were derived by quantizing measurements of the

speech wave. The best quality was derived by analog measurement. It may be that the perceptual difference between the analog and digital results is large compared with differences between the digitalized levels of quality. Therefore, subjects may simply have been detecting this difference, and assigned relatively good quality digitalized speech to the digitalized anchor.

Yet, another unexpected finding was the distribution of response times over the quality continuum. It was expected that the stimuli at the extremes of the continuum would be easy judgments compared with those in the center, and that this would be reflected by shorter decision times at those points. However, mean response times are rather evenly distributed across the quality levels. This again may be an indication of the kind of discrimination subjects actually made in spite of the instructions. The same is true for frequency of anchor use, although in this case the distribution is more widely scattered. Nevertheless, it is not ordered in relation to the quality level of the stimuli.

The measures of response time and frequency of anchor use proved to be valuable additions to the response description. Both measures clearly differentiated between the two types of materials for all subjects. In addition, both provide some estimate of the judgmental difficulty of the judgmental task. This was particularly true for the pathological subjects where a perceptual difference between ears was clearly defined, and one suspects that vocoding appears to be an effective means of degrading speech for the identification of central auditory lesions.

While an attempt was made in this study to identify an optimal level of distortion which might be useful for the study of central auditory defects, no specific level appears to be related to a failure of discrimination. This may be due in part to the nature of the technique. It is possible that the use of fewer and more widely spaced quality levels would have been preferable. As it was, the response of the defective ear was typified by an increased number of "poor quality" identifications. It was as though the range of quality had been compressed, with little discriminable difference between the levels. This is also reflected in correspondingly shorter response times accompanied by fewer references to the anchors.

By reversing speech and converting it to a non-meaningful signal, it was thought that the subject would be forced to make judgments on the basis of quality alone. Reversed speech is not, however, totally devoid of meaning. It is recognized as speech and subjects reported discerning syllables and occasional words. Some subjects also reported detecting differences in inflection, so that dimensions other than quality were present and presumably influenced subsequent responses. In any case, the responses of both groups to the reversed speech were the same, and therefore did not increase the sensitivity of the test for the identification of central auditory defect over the use of standard speech.

This study explored the usefulness of vocoded speech as a degraded speech signal in the identification of central auditory lesions. As previously noted (7), it is believed to be the removal of some information

in the speech signal which makes it specific for central lesions, and the elimination by vocoding appears to be an effective one for this purpose. The technique also appears to be useful for studying differential discrimination and speech perception, since it requires neither recognition nor context for a discrimination to be made.

A previous study (51) demonstrated that the use of confidence ratings provided more information in tests of intelligibility than conventional procedures. In this study, the use of test referents was viewed as a measure of confidence and certainty. It may be equally regarded as reflecting attentional features of the situation. This measure added considerably to the description of the subject's response, and provided an estimate of subjective difficulty between materials and between normal and pathological groups.

Response time, suggested as an intelligibility measure by Hecker (24), also provided an estimate of relative difficulty, and these two measures tend to confirm each other in the results obtained.

By using a technique which does not require a verbal response, it was possible to apply some of the refinements of intelligibility testing to a differential discrimination task utilizing speech, and to obtain additional information of the subject's response to the subjective difficulty of the judgment.

Conclusions

From the analysis of the data, the conclusion may be drawn that the meaningfulness of a stimulus contributes to the perception

of complex stimuli, but that the facilitating effect is not sufficient to invalidate tests of central auditory disorders. Apparently the distortion introduced to the speech wave by the vocoding process is sufficient to disturb the perception of such signals in the ear contralateral to a damaged hemisphere. No attempt has been made to assess the extent of damage or the locus, other than placing it in the right hemisphere

It may be further stated that in certain types of auditory tests a differential discrimination task utilizing the refinements of response time and confidence judgments will provide more information than the present recognition techniques.

There is some difficulty in drawing conclusions relative to the processing of non-meaningful, complex stimuli by normal or brain-damaged subjects on the basis of the stimulus used. Presumably, it would be possible to construct a truly non-meaningful stimulus which varies in all the dimensions which might be compared to forward speech. Early attempts with vocoded music illustrate the difficulty of this problem and further study is clearly needed.

The finding of the apparently wide perceptual difference between analog and digitalized synthetic speech has some relevance for those speech perception studies carried out with synthetic speech. While synthetic speech guarantees freedom from variability, the judgments of a subject may take place on a dimension which is not found in unprocessed speech. A study involving only one type of processed

speech at different quality levels may show quite different results. The possibility remains that some non-apparent dimension, such as noise introduced during the processing, was responsible for the high number of assignments to the poorer anchor. The technique appears to be unquestionably promising for testing the differential discrimination of speech on any dimension. It is free from dependence upon linguistic sophistication, recognition, and verbal response. It represents a narrower measure than does the absolute discrimination technique and is compatible with measures other than percentage correct. With automatic data collection, such data can be used to demonstrate the more subtle aspects of auditory discrimination.

With a non-verbal response, such as was used in this study, it is also possible to study perception in those individuals for whom a verbal response is precluded (as in aphasia). It would be useful to know whether such subjects could give valid test results and whether this type of material is resistant to peripheral distortion (so that it could be used with subjects who have peripheral hearing deficits).

The experimental paradigm has few constraints, and allows for the conduct of learning/reward studies, infra-human experimentation, and the examination of numerous environmental and organismic variables.

APPENDIX A
TEST STIMULUS MATERIALS

STIMULUS MATERIALS

I. SENTENCES

Level I Analog

1. His remarks are too dense.
2. Please return that bar stool.
3. Set Debbie's car for speed.
4. Palm trees grow very tall.
5. Prudent thieves look harmless.
6. His partner took a jar.
7. She gets to watch parties.
8. Chicken farms are for eggs.
9. The old red barn burned down.
10. Let's sit in the cool bar.

Level II 3200BPS

1. The man tore his dark suit.
2. Jack took part of the fish.
3. The cook has a red shawl.
4. The tall man took your seat.
5. The next show starts in March.
6. Hand me the blue teapot.
7. He foiled the plot this time.
8. The key is stuck in the lock.
9. If you try now, it should work.
10. That is certainly worth a thought.

Level III 3000 BPS

1. Two will be about right.
2. It was a worthless sham.
3. Take the net down and move it.
4. They fought the fire all night.
5. Show your left foot to the doc.
6. The kernel dropped to the ground.
7. It was damp and humid after dark.
8. He shot down the steep slope.
9. The pear tree was blighted and brown.
10. A huge hawk soared overhead.

Level IV 2600 BPS

1. You can judge that for yourself.
2. A few were made out of pine.
3. He bought their best skis.
4. Practice and you'll learn.
5. The tank dropped in the shaft.
6. It was cool along the beach.
7. At last, they won a match.
8. The skid mark crossed the line.
9. He rode out in the fog.
10. It could save a lot of time.

Level V 2400 BPS

1. Now set the pot to boil.
2. He got to be a pest.
3. You can say what you please.
4. It wasn't worth a lot.
5. The school will close at noon.
6. He was a crack shot.
7. We'll adjourn before eight.
8. Further out it's very deep.
9. His remarks are too dense.
10. Please return that bar stool.

Level VI 2300 BPS

1. Set Debbie's car for speed.
2. Palm trees grow very tall.
3. Prudent thieves look harmless.
4. His partner took a jar.
5. She gets to watch parties.
6. Chicken farms are for eggs.
7. The old red barn burned down.
8. Let's sit in the cool bar.
9. The man tore his dark suit.
10. Jack took part of the fish.

Level VII 2100 BPS

1. The cook has a red shawl.
2. The tall man took your seat.
3. The next show starts in March.
4. Hand me the blue teapot.
5. He foiled the plot this time.
6. The key is stuck in the lock.
7. If you try now, it should work.
8. That is certainly worth a thought.
9. Two will be about right.
10. It was a worthless sham.

Level VIII 1700 BPS

1. Take the net down and move it.
2. They fought the fire all night.
3. Show you left foot to the doc.
4. The kernel dropped to the ground.
5. It was damp and humid after dark.
6. He shot down the steep slope.
7. The pear tree was blighted and brown.
8. A huge hawk soared overhead.
9. You can judge that for yourself.
10. A few were made out of pine.

Level IX 1400 BPS

1. He bought their best skis.
2. Practice and you'll learn.
3. The tank dropped in the shaft.
4. It was cool along the beach.
5. At last, they won a match.
6. The skid mark crossed the line.
7. He rode out in the fog.
8. It could save a lot of time.
9. Now set the pot to boil.
10. He got to be a pest.

ANCHOR MATERIAL: CONNECTED TEXT

He had an immense capacity for work, and equal talent for obtaining the best from others, an almost impeccable judgment of men and a genius for making prompt firm decisions. Chester Nimitz was one of those rare men who grow as their responsibilities increase. Originally, Nimitz set his sights on West Point, but a Congressman told him there would be no appointments from his district for several years and suggested he try for the Naval Academy. Nimitz did and won the Annapolis appointment over thirteen other youths. In nineteen hundred and four, graduate of Annapolis, Nimitz saw battleship, cruiser, submarine and gunboat duty before being named Commander of the U.S. Pacific Fleet. Unlike many other World War commanders, Nimitz never wrote his memoirs. He said he felt such books often contained many critical remarks and self praise at the expense of others.

A program for
DADER: Data Analysis and Dump of Event Records¹

I. Program Input

- (1) Magnetic tape containing subject ID and event data.
- (2) Namelist SETUP
 - WDUMP = TRUE, print dump of subject data
 - = FALSE, skip dump printing
 - TPRINT = TRUE, print event data for each test
 - = FALSE, skip event data printing.
 - along with informational messages
- (3) Punched card data defining information pertinent to the statistical analysis.

	<u>Card Column</u>	<u>Description</u>
	1-2	Test number (01-90)*
	4	Ear designation (L=Left, R=Right)
Set 1	6	Row index of output arrays
	8	Column index of output arrays
	10	Test answer (not used)
	11-12	Test number
	14	Ear designation
Set 2	16	Row index of output arrays
	18	Column index of output arrays
	20	Test answer (not used)
	.	
	.	
	.	
	.	
	.	
	70	

*A value of 99 stored in the test number slot signals the end of data for the current subject ID.

Seven sets of data are placed on each card until all the defining information has been inputted for the current subject ID. Each time a new subject ID occurs a complete set of punched card data is read in.

¹This program was developed under Contract F19628-68-C-0155 as an integral portion of the reported effort. Certain details were designed to satisfy specific analytic requirements of the study.

II. Program Output

A. General Output

- (1) Dump printout of each record of subject data as read from the magnetic tape. (if "WDUMP = TRUE")
- (2) Printout of each set of punched input data cards.
- (3) Printout of each trial showing the individual event numbers together with their times. (labeled "Test No. ___"; if "TPRINT = TRUE")
- (4) Printout of the statistical summary calculated for each subject ID.
- (5) Printout of informational and diagnostic messages.

B. Diagnostics and Informational Messages

"CANNOT PROCESS BYTE"
An illegal byte has occurred.

"CAN NOT FIND EVENT 1."
Event 1 does not exist for this test. An event 1 is inserted .5 seconds before first event of the test.

"TEST DID NOT BEGIN WITH EVENT 1."
Check the event sequencing. Events out of order are discarded.

"DUPLICATE EVENT NUMBER ____."
When duplicate event numbers occur only the last occurrence is processed.

"EVENT SEQUENCE ERROR ON THE __TH EVENT."
Events out of sequence are discarded.

"THE LAST ____ TESTS OF HEADER ____ WERE NOT PROCESSED."
Not enough tests in set to perform a statistical analysis.

III. Major Subroutines

A machine language routine retrieves the bit information from tape. The principle Fortran routines which may be easily modified are modular as follows:

- APRINT - output and printing routine
- ALMAN - outputs as cards raw and corrected matrices
- MAIN - central logical and data handling routine. Most modification to suit individual needs will be done here.

III. Major Subroutines (Continued)

Certain routines incorporated in the program check particular logical conditions of this experiment, while others relate the computer system itself for checks such as synchronous read errors, etc.

BLANK PAGE

BIBLIOGRAPHY

1. Black, J.W. and Haagen, C.H. Multiple-choice Intelligibility Tests. JSHD, 28:77-87, 1963.
2. Bocca, E. "Clinical Aspects of Cortical Deafness," Laryng., 68:301-309, 1958.
3. Bocca, E. "Factors Influencing Binaural Integration of Periodically Switched Messages," Acta Otolaryng., 53:142, 1961.
4. Bocca, E. "Binaural Hearing: Another Approach," Laryngoscope, 65:1164-1171, 1955.
5. Bocca, E., Calearo, C. and Cassinari, V. "A New Method for Testing Hearing in Temporal Lobe Tumors," Acta Otolaryng., 44:219-221, 1954.
6. Bocca, E., Calearo, C., Cassinari, V. and Migliavacca, F. "Testing Cortical Hearing in Temporal Lobe Tumors," Acta Otolaryng., 45:289-304, 1955.
7. Bocca, E. and Calearo, C. "Central Hearing Processes," in Jerger, J. (ed.), Modern Developments in Audiology. New York: Academic Press, Inc., 1963.
8. Broadbent, D.E. "The Role of Auditory Localization in Attention and Memory Span," J. Experimental Psychology. 47:191-196, 1954.
9. Bryden, M.P. "Ear Preference in Auditory Perception," J. Experimental Psychology, 65:103-105, 1963.
10. Calearo, C. "Binaural Summation in Lesions of the Temporal Lobe," Acta Otolaryng., 47:392, 1957.
11. Calearo, C. and Antonelli, A. "Cortical Hearing Tests and Cerebral Dominance," Acta Otolaryng., 56:17, 1963.
12. Calearo, C. and Lazzaroni, A. "Speech Intelligibility in Relation to the Speed of the Message," Laryng., 67:410-419, 1957.
13. Carhart, R. "Clinical Determination of Abnormal Auditory Adaptation," Arch. Otolaryng., 65:32-39, 1957.

14. Clarke, F.R. "Confidence Ratings and Second Choice Responses," Signal Detection and Recognition by Human Observers, J.A. Swets, editor. New York: John Wiley and Sons, 1964.
15. Clarke, F.R. and Bilger, R.C. "The Theory of Signal Detectability and Measurement of Hearing," Modern Developments in Audiology, J.Jerger, editor. New York: Academic Press, Inc., 1963.
16. Denes, P. and Pinson, E. The Speech Chain. Bell Telephone Laboratories, Inc., 1963.
17. Dirks, D. "Perception of Dichotic and Monaural Verbal Material and Cerebral Dominance for Speech," Acta Otolaryng., 58:73, 1964.
18. Egan, J. and Clarke, F.R. "Source and Receiver Behavior in the Use of a Criterion," Signal Detection and Recognition by Human Observers, J.A. Swets, editor. New York: John Wiley and Sons, 1964.
19. Farrimond, T. "Factors Influencing Auditory Perception of Pure Tones and Speech," JSHR, 5:194, 1962.
20. Flanagan, J.L. Speech Analysis, Synthesis and Perception. New York: Academic Press, Inc., 1965.
21. Goldman, S.O. and Katz, J. "The SSW Test: Dichotic, Diotic, and Monaural" Paper read before the 42nd Annual Convention of the American Speech and Hearing Assn., Washington, 1966.
22. Goldstein, R. "Hearing and Speech Follow-up in Left Hemispherectomy," JSHD, 26:126, 1961.
23. Goldstein, R., Goodman, A. and King, R.B. "Hearing and Speech in Infantile Hemiplegia Before and After Left Hemispherectomy," Neurology, 6:869-876, 1956.
24. Hecker, M.H., Stevens, K.N., and Williams, C. E. "Measurement of Reaction Time in Intelligibility Tests," J. Acoust. Soc. Amer., 39:1188, 1966.
25. Heinz, J.M. and Stevens, K.N. "On the Properties of Voiceless Fricative Consonants," J. Acoust. Soc. Amer., 33:589-596, 1961.
26. Helm, Stanley. AFCRL Polymodal Vocoder Modification and Improvement Program. Final Report, AFCRL, Office of Aerospace Research, USAF, Bedford, Mass., 1966.

27. House, A., Williams, C., Hecker, M., and Kryter, K.D. "Articulation Testing Methods: Consonantal Differentiation with a Closed Response Set," J. Acoust. Soc. Amer., 37:158, 1965.
28. House, A., and Stevens, K.N. "Auditory Testing of a Simplified Description of Vowel Articulation," J. Acoust. Soc. Amer., 27:882-887, 1955.
29. House, A.S. and Stevens, K.N. "Analog Studies of the Nasalization of Vowels," JSHD, 21:218-232, 1956.
30. House, A.S., Stevens, K.N., Sandel, T.T., and Arnold, J.B. "On the Learning of Speech-like Vocabularies," J. Verb. Learn. and Verb. Behav., 1:133-143, 1962.
31. Inglis, J. "The Influence of Motivation, Perception and Attention on Age-Related Changes in Short-Term Memory," Nature, 204:103, 1964.
32. Jerger, J. "Observations on Auditory Behavior in Lesions of the Auditory Pathways," AMA Arch. Otolaryng., 71:797-806, 1960.
33. Jerger, J. "Audiological Manifestations of Lesions in the Auditory Nervous System," Laryng., 70:417-425, 1960.
34. Jerger, J. "Bekesy Audiometry in Analysis of Auditory Disorders," JSHD, 275-281, 1960.
35. Johnson, D. Psychology of Thought and Judgement. New York: Harper, 1955.
36. Kimura, D. "Cerebral Dominance and the Perception of Verbal Stimuli," Canad. J. Psychol., 15:166-171, 1961.
37. Kimura, D. "Cerebral Dominance in Temporal Lobe Damage," Canad. J. Psychol., 15:156-165, 1961.
38. Kryter, K.D., and Whitman, E.C. "Some Comparisons Between Rhyme and BB Intelligibility Tests," J. Acoust. Soc. Amer., 37:1146, 1965.
39. Lehiste, I. and Peterson, G. "Linguistic Considerations in the Study of Speech Intelligibility," J. Acoust. Soc. Amer., 31:280-286, 1959.
40. Lieberman, A.M., Cooper, F.S., Harris, K.S., and MacNeilage, P.F. "A Motor Theory of Speech Perception," Proc. Speech Comm. Sem., RIT, Stockholm, September, 1962.

41. Lieberman, A.M., Delattre, P.C., Cooper, F.S., and Gerstman, J.L. "The Role of Consonant-vowel Transitions in the Stop and Nasal Consonants," Psychol. Monographs, 68:379-381, 1954.
42. Lieberman, A.M., Cooper, F.S., Shankweiler, D., and Studdart-Kennedy, M. "Why are Speech Spectrograms Hard to Read?" Amer. Annals of the Deaf, 113:127-133, 1968.
43. Linden, A. "Distorted Speech Tests and Binaural Resynthesis Tests," Acta Otolaryng., 32:58-66, 1964.
44. Luterman, D., Welsh, O., and Melrose, J. "Response of Aged Males to Time-Altered Speech Stimuli," JSHR, 9:226, 1966.
45. Matzker, J. "Two New Methods for the Assessment of Central Auditory Functions in Brain Disease," Ann. Otol. Rhinol. Laryngol., 68:1185-1197, 1959.
46. Miller, G.A., Heise, G.A., and Lichten, W. "The Intelligibility of Speech as a Function of the Context of the Test Materials," J. Exp. Psychol., 41:329-335, 1951.
47. Mostofsky, D.I. and Green, E. "Two-Category Judgement Task with Available Anchors." Proceedings of the 74th Annual Convention of the American Psychological Assn., 1966.
48. Owens, E. "Intelligibility of Words Varying in Familiarity," JSHR, 4:113, 1961.
49. Palva, A. "Filtered Speech Audiometry," Acta Otolaryng. Supp. 210, 1966.
50. Pollack, I. "The Information of Elementary Auditory Displays," J. Acoust. Soc. Amer., 24:745-749, 1952.
51. Pollak, I. and Decker, L. "Confidence Ratings, Message Reception and the Receiver Operating Characteristic," Signal Detection and Recognition by Humans. New York: John Wiley and Sons, 1964.
52. Quiros, J.B.de "Accelerated Speech Audiometry, An Examination of Test Results," Translations of the Beltone Institute for Hearing Research. Chicago: 4201 W. Victoria St., 1964.
53. Rosenblith, W. and Stevens, K.N. "On the DL for Frequency," J. Acoust. Soc. Amer., 25:980-985, 1963.
54. Savin, H.B. "Word Frequency Effect and Errors in Perception," J. Acoust. Soc. Amer., 35: 200-206, 1963.

55. Shapiro, I. and Naunton, R. "Audiological Evaluation of Acoustic Neurinomas." Paper read before the 42nd Annual Convention of the American Speech and Hearing Association, Washington, D.C., 1966.
56. Smith, C.P. "Vocal Response Synthesizer," J. Acoust. Soc. Amer., 37:1, 170-171, 1965.
57. Speaks, C. and Jerger, J. "Method for Measurement of Speech Identification," JSHR, 8:185, 1965.
58. Tiffany, W.R. and Bennett, D.N. "Intelligibility of Slow-Played Speech," JSHR, 4:248, 1961.
59. Traul, G.N. and Black, J.W. "The Effect of Context on Aural Perception of Words," JSHR, 8:363, 1965.
60. Welsh, O. and Luterman, D. "The Effects of Aging on Responses to Filtered Speech." Paper read before the 42nd Annual Convention of the American Speech and Hearing Association, Washington, D.C., 1966.
61. Voiers, W. "Performance Evaluation of Speech Processing Devices III. Diagnostic Evaluation of Speech Intelligibility. Final Report AFCRL-67-0101, March, 1967.
62. Yaggi, Lawrence. Full-Duplex Digital Vocoder. Scientific Report #1, AFCRL, Office of Aerospace Research, USAF, Bedford, Mass., 1962.

BLANK PAGE

DIGITAL DATA ACQUISITION FACILITY

The Digital Data Acquisition System¹ can be designed to accept an unlimited number of independent on/off events from one or more experiments and record them on magnetic tape in a binary format with parity written as bit number 9. An event in a typical laboratory might represent a response from a subject, the correctness (or incorrectness) of the response, a stimulus, or reinforcement, etc. These might originate from one or more animals responding in one or more environments, or laboratories simultaneously. In general, an event is anything that may be represented by the opening or closing of a switch. When a rat presses a pedal this may be captured as event #1, and release of that bar as event #2. Alternatively a press in box #A may be represented as event #1 and the press in box #B as event #2. The user is free to assign his own event codes, subject to the limit available on the system. Time, in a 7-bit binary format, is recorded along with the data information.

Each event is presented to the system through a mercury wetted reed relay which acts as a buffer to isolate the data lines from the sensitive logic of the system. For this system the data must have the specifications which will operate the transistor driver of the reed relay and are as follows: a logical zero is equal to +0.7 volts and a logical one is equal to -15 to -23 volts and have a minimum duration of two milliseconds.

¹Design, assembly, and stock components used in the system and described in this report were products of Massey-Dickinson Company, Saxonville, Massachusetts. The incremental tape recorder was a 1400/360 9 track IBM Compatible unit manufactured by Kennedy Company, Altadena, California.

The output of the reed relay for each event operates or sets a memory (flip flop). This serves as a storage element for the data system since the piece of data will not be immediately accepted but must wait for a finite time to elapse for the magnetic tape deck to write the time or any previous data which might have occurred previous to that of the event of interest.

When any of the event memories are set, a "data alarm" circuit is enabled. Essentially this is a large Or gate. Any event(s) that has occurred will operate this data alarm gate, and initiate a "scan" activity.

The scan is made up of several counter stepper cards which can be compared to a stepping switch. The scan is in the rest or home position when no data is occurring. When any piece of data or time-overflow occurs (to be discussed later), a high speed digital multiplex clock is enabled. Each data point, time byte, overflow and rest has its own position in a scan. The clock pulse from this digital multiplex clock causes a scan to advance to the first active position and in this case is the time. The time byte has a memory associated with it which is set on the occurrence of data.

On the output of each memory is an And gate. One leg of the And gate is connected to the output of the memory. The second leg is connected to the appropriate stage or position of the scan. When the scan enters a position where the memory is set, a "data present" exists; that is, a signal that indicates there is data and that it is in its particular position.

The "data present" tells the write and step control of the tape deck that the data or time is in position to be punched (recorded). After the tape deck has accepted that particular data or time character, the scan is forced out of that position and proceeds at a high rate of speed (50 KC) until it reaches the next active or set memory of the appropriate event which has occurred. The scan clock is automatically frozen when a scan comes upon a stage where the data memory is set and is held there until the tape deck has completed its writing cycle. The data memory which was just entered is reset.

Each event or data point has a code associated with it. The inputs for the encoder (or code generator) are derived from the output of the And gate which tells the system that the scan has reached an active data point. This method guarantees that only the code associated with the active data point will be present during that writing cycle since the scan can only be in one data position at a time.

The system has a 7-bit generator (1, 2, 4, 8, 16, 32, 64). A clock with a variable rate, (i.e., resolution), generates a stream of pulses which are accumulated in the clock register. Since the clock register can accumulate up to a count of 127 ($1 + 2 + 4 + 8 + 16 + 32 + 64$), the 127th pulse "going away" (or 128th) pulse is called an overflow. A clock overflow is recorded as a unique character in the data system similar to that of an event. To measure the time between events it is only necessary to (1) note the time recorded with the first event of interest, (2) subtract it from the total number of overflows which have occurred between the two events and (3) add to the time which is recorded with the second

event of interest. It is possible, therefore, to measure response latencies or event-to-event durations over many hours by simply counting the number of overflows that have occurred together with the time associated with those events. This technique does not require a long clock register as in a system without an overflow, nor does it impose excessive use of tape or other recording media.

On the occurrence of any piece of data the time register is sampled or interrogated. Effectively, the system looks at the clock at the instant (or within $\frac{1}{2}$ clock cycle) of the occurrence of the event. This information is transferred to a storage register which holds the time information until the tape deck and its timing circuitry have a chance to record the time flag (i.e., a unique character signaling that the next entry is a time byte). The storage register is reset after that character has been written. This guarantees that the storage register will be available for the next interrogation of the clock register which occurs on the next data scan, and which when recorded, will be the clock value itself.

It is possible, and most probable, that more than one piece of data or event will be picked up or accepted on a given scan cycle. A scan cycle looks at all the scan positions when any one piece of data occurs. Since it takes a finite time for any piece of data to be written, whenever one or more events do occur in a given scan cycle, the same time characters will be associated with the one or more events that have occurred in that cycle. (See Table 1)

It must be realized, however, that the time recorded with an event(s) may be skewed by an interval whose length depends upon how many pieces of data have been written in a scan. If only one event occurs in a given scan cycle, then the time associated with that event is accurate to within $\frac{1}{2}$ a clock cycle. If two or more events occur in a given scan then there is a time skew associated with the events. Because each event has a position in the scan it is not possible to know which event occurred first, second or nth, since they are picked up on the serial order of position and do not reflect priority of occurrence.

In a 10-event system having a stepping rate of 500 per second, the time associated with the data will be very close to its true value if the selected clock resolution is greater than 10 milliseconds (assuming the probability of all 10 events occurring at once).

A bit check or parity network is included in the tape deck and is recorded as the 9th bit in each byte. The parity network looks at the other 8 bits and records a logical one if the number of logical ones is even and a logical zero if the number of bits is odd.

After every 4096 bytes have been written (plus a few), an inter-record (IR) gap will be automatically entered on the tape. If 4096 characters occur during a writing cycle the entry of the IR gap will be delayed until the scan has returned to the rest or home position. It is entered after a two millisecond delay to allow the tape deck to complete its timing cycle. The system is returned to normal when the "gap in process" signal goes away.

Other common hardware features

An echo check alarm circuit is incorporated in this system. If the head current does not agree with the written information than a parity alarm is generated by the tape deck and is counted by the data system. An audible alarm will sound upon the occurrence of four parity errors and will go off after 32 write cycles resulting in a beeping alarm.

A steady alarm indicates that an end-of-tape or broken tape condition has been detected. A photocell system is used to sense the presence of tape.

A TEST push button allows the user to test the entire system by setting all the event storage memories. It is advisable to test the system this way before use, to guarantee that it will, in fact, cycle through. It is also recommended that the system be tested in this way after loading a new tape and that a computer program be written to include a print of all the recorded data.

Manual entry of data

A manual identification panel is included in the system to allow the investigator to enter manual identification codes. Such codes might be used to identify the experiment, subject number, etc. The entry is accomplished by depressing the appropriate push button representing the corresponding bits of a byte. The required number of bits are entered on tape by depressing an ENTER (or operate) button. It is recommended that each character be entered at least twice to insure that the data has been recorded correctly as it is possible, but highly unlikely, that the manual data position has a tape drop-out, and since the entire program

which is to follow is dependent upon this manual identification code it should be guaranteed that the code is entered correctly.

It is necessary to guarantee that the manual identification codes are unique, that is, that their respective codes are different from any other event code. (See Table 2). Rather than consuming valuable channels, an alternative is also possible, viz., to enter a unique manual data identification code telling the computer that "the following block of information begins a manual identification code only, not data to be acted upon." This code would also be entered at the end of the manual identification block. This latter procedure allows the use of previously used codes, or their combination for manual codes, because they are block identified. (See Table 3)

When any bit button is depressed the entire system is frozen. The manual data entry buttons must be released before the data system will proceed to enter any electrical data. This is done either by depressing the master release push button, or by pulling out the individual bit buttons which have been pressed. On occasion, an investigator might intentionally wish to leave a manual entry button depressed in order to effect a system lock and prevent any intermediate tape write commands.

Computer programming requirements

Several routines should be available as a standard procedure for processing the recorded data. First, a DUMP program should be prepared. This program enables a binary (byte) printout of the tape to assure that the system is writing and that the event codes are functioning appropriately. Another dump, converting the binary information in

TABLE 2 SAMPLE EVENT CODE CONFIGURATION FOR 10 EVENTS

Binary Weight	1	2	4	8	16	32	64	128	p*
Channel No.	1	2	3	4	5	6	7	8	9
OVERFLOW**	0	0	0	0	0	0	0	1	0
TIME FLAG	0	0	0	0	0	0	1	1	1
(lower limit) TIME	0	0	0	0	0	0	0	0	0
(upper limit) TIME	1	1	1	1	1	1	1	0	0
EVENT #1	1	1	0	0	0	0	0	1	1
2	0	1	0	0	0	0	0	1	1
3	1	1	0	0	0	0	0	1	0
4	1	0	1	0	0	0	0	1	0
5	1	0	1	0	0	0	0	1	0
6	1	1	1	0	0	0	0	1	0
7	1	1	1	0	0	0	0	1	1
8	1	0	0	1	0	0	0	1	1
9	1	0	0	1	0	0	0	1	0
10	0	1	0	1	0	0	0	1	0

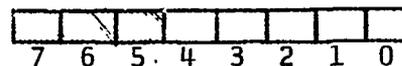
*Parity (Odd)

**128 x Clock Resolution

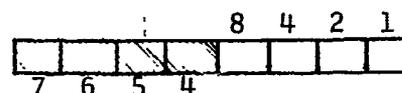
Table 3 SAMPLE GUIDE FOR ENTERING MANUAL DATA

(Consider that the computer requires a unique code to indicate the beginning of a block of ID data, the end of the ID data block, and the ID itself.)

- 1) Depress bit buttons 7, 6, 5 (begin ID code) and operate ENTER twice



- 2) Depress bit buttons 4, 5, 6, 7 (code accompanying ID)

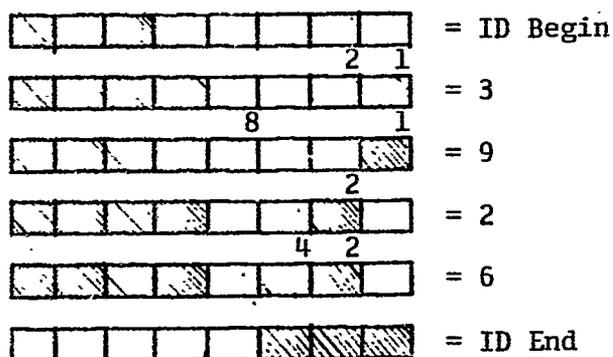


- 3) Enter first ID digit in BCD form according to binary weight using bits 0, 1, 2, 3 ONLY, then operate ENTER twice

(Repeat above as needed. Remember not to use RED release button for intermediate release of the single bits, but rather manually release the individual buttons.)

- 4) After last ID is entered, release all but bits 2, 1, 0 (end ID code) and operate ENTER twice
- 5) Press RED release button thereby releasing all bit buttons on panel. Events and overflows will now be entered on tape.

For example:
To insert ID No. 3926



readable digital form is also valuable for preliminary inspection. While subsequent programs can often be written in macro languages (PL-I, Fortran), the DUMP programs will have to be written in Machine Assembly or basic language. In many cases there will be little need to further inspect a record if the record of the dump can be judged to be satisfactory. Next, a program to merge the records of several tapes (i.e., keep building a master record) should be available. Quite often, only a fraction of the reel will be used and it will be inefficient to maintain a quantity of partially used tapes each to be analyzed in separate passes, and each containing lengths of unused tape. Finally, the necessary statistical or formatting programs (appropriate for the unique experimental situation) can be written. These routines can be prepared as separate phases of a single program, or may be left as independent programs.

In writing the programs it is important to think beyond the immediate study under analysis. Modifications or program rewriting at a later date is at best an expensive and time consuming operation. If options, various input formats, variable event sequences, etc. are even remotely anticipated, it is best to have the program capable of handling them at the outset. Similarly, possibilities of human experimenter error should be considered and appropriately anticipated in the program routines especially if more than a single person is likely to operate the data acquisition system.

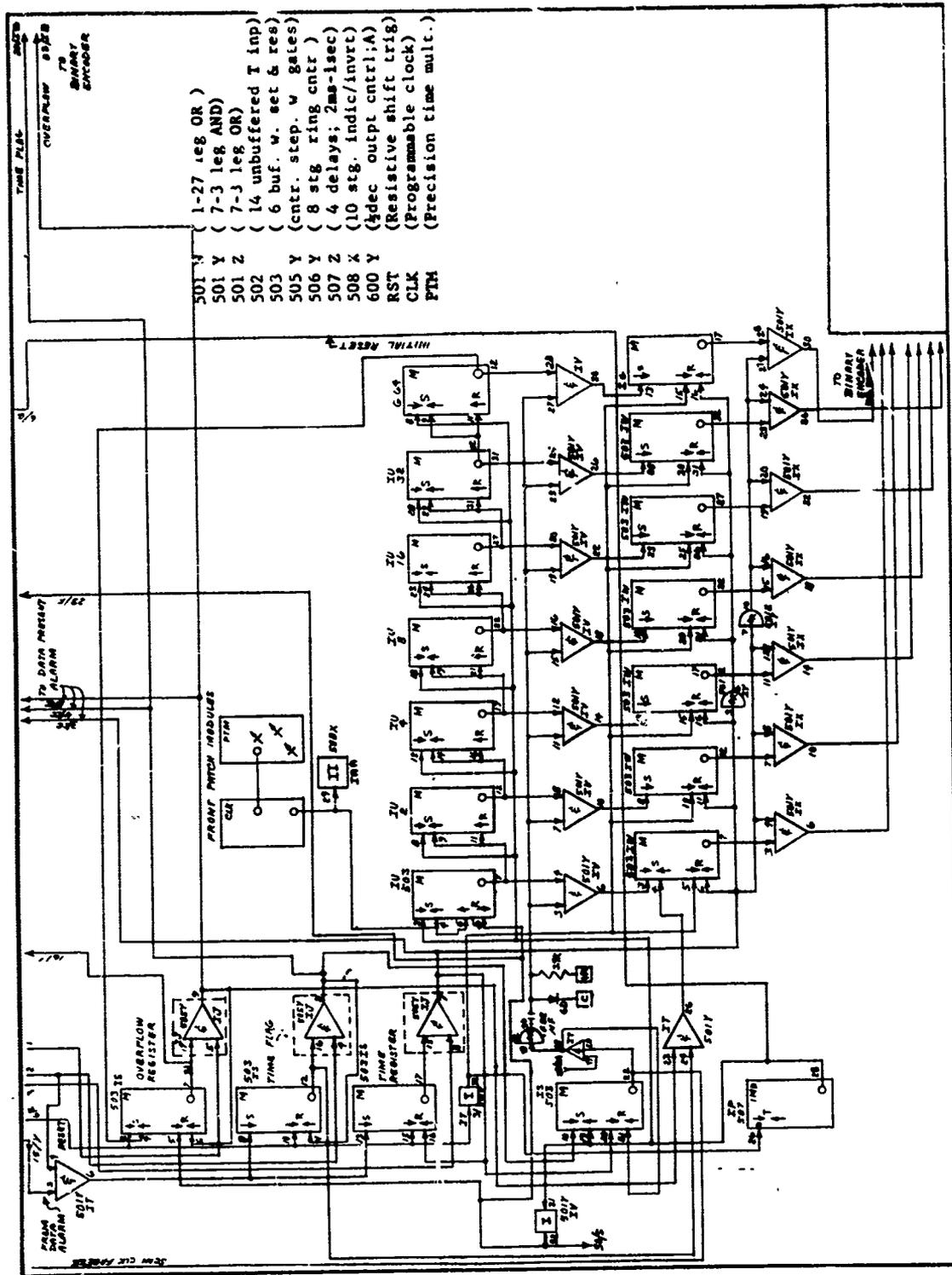


FIGURE 1
Data Acquisition System: Time Clock Section

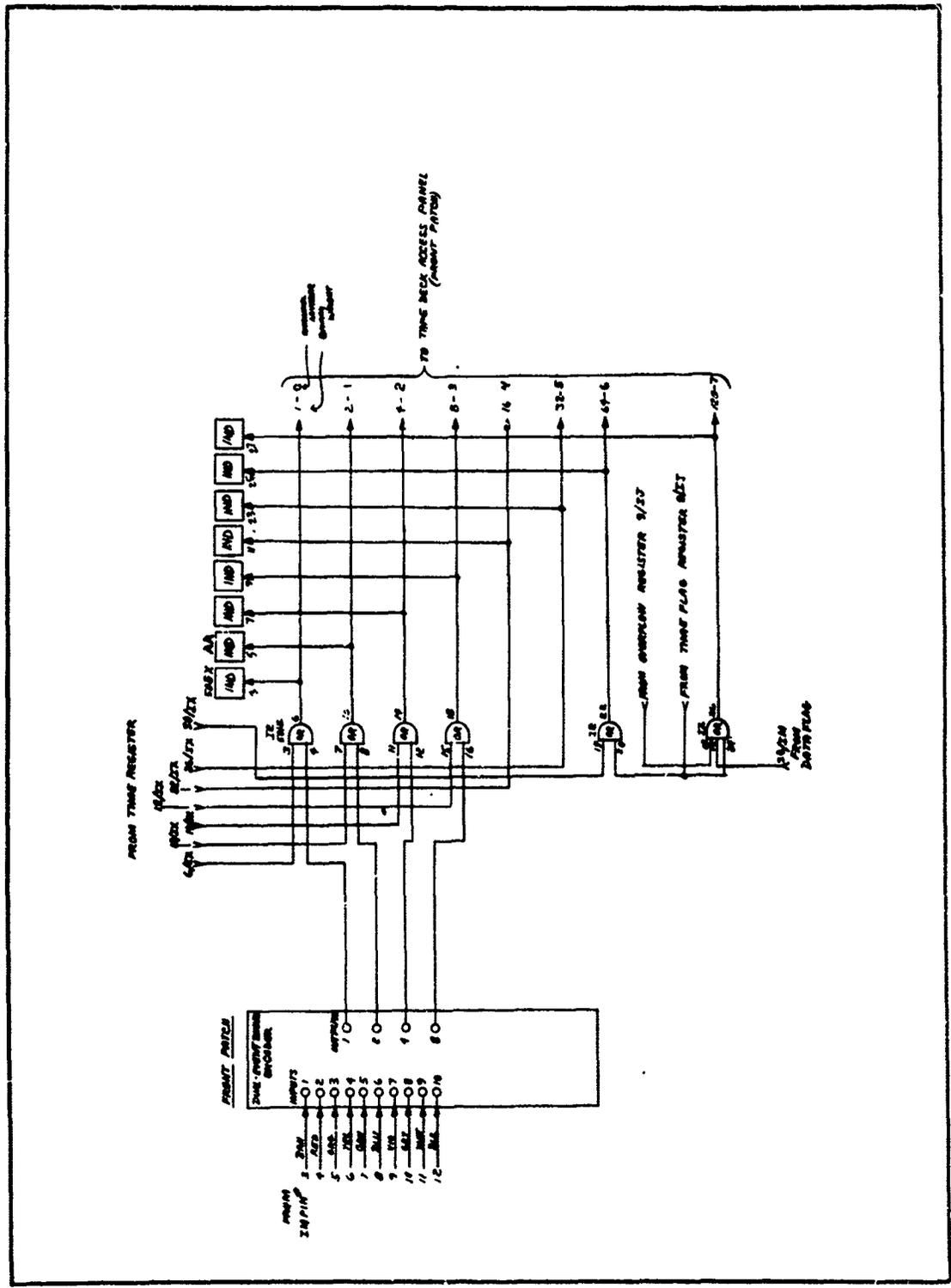


FIGURE 2
Data Acquisition System: Binary Encoder Section

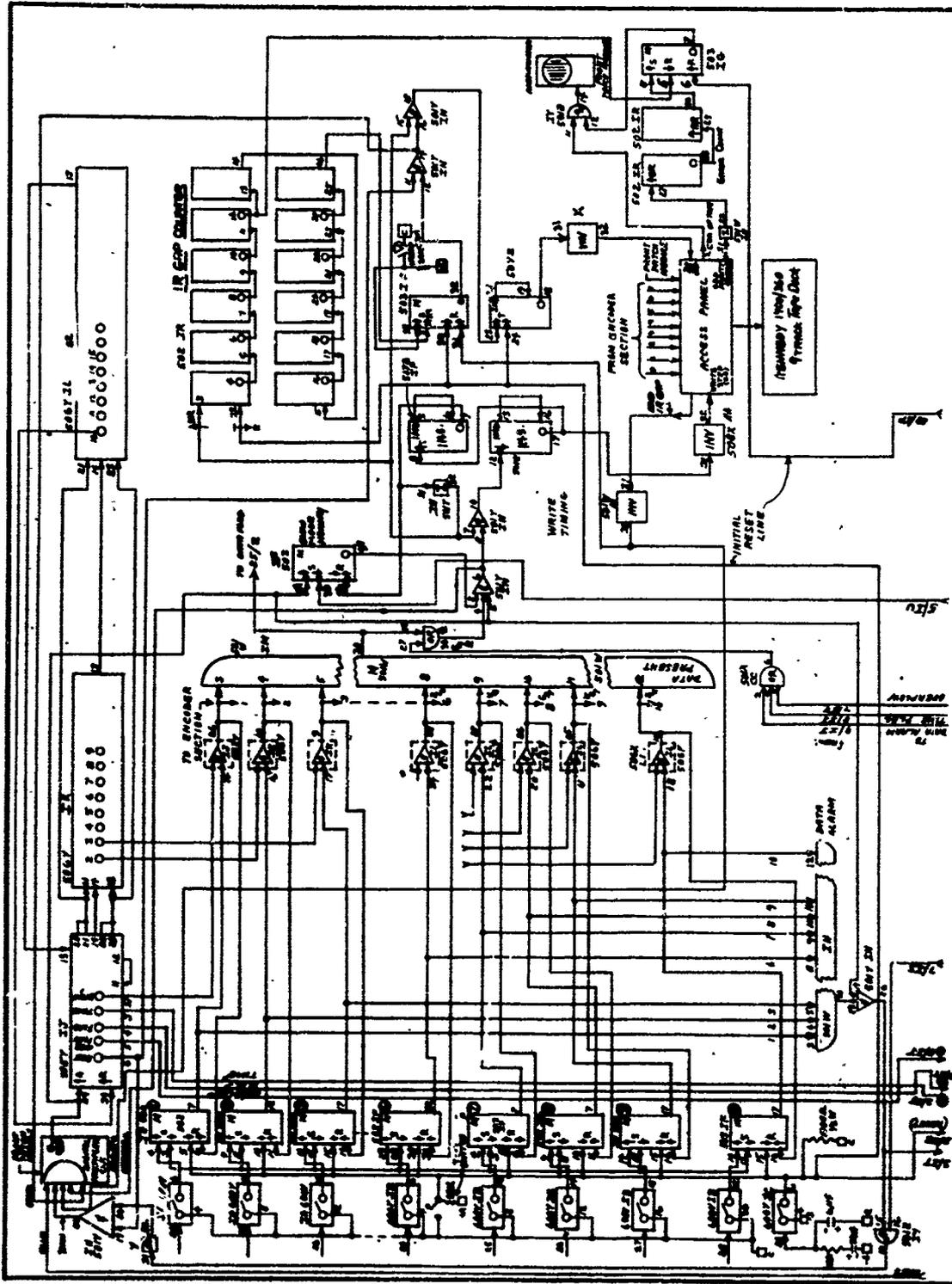


FIGURE 3
Data Acquisition System: Alarm and Write Section

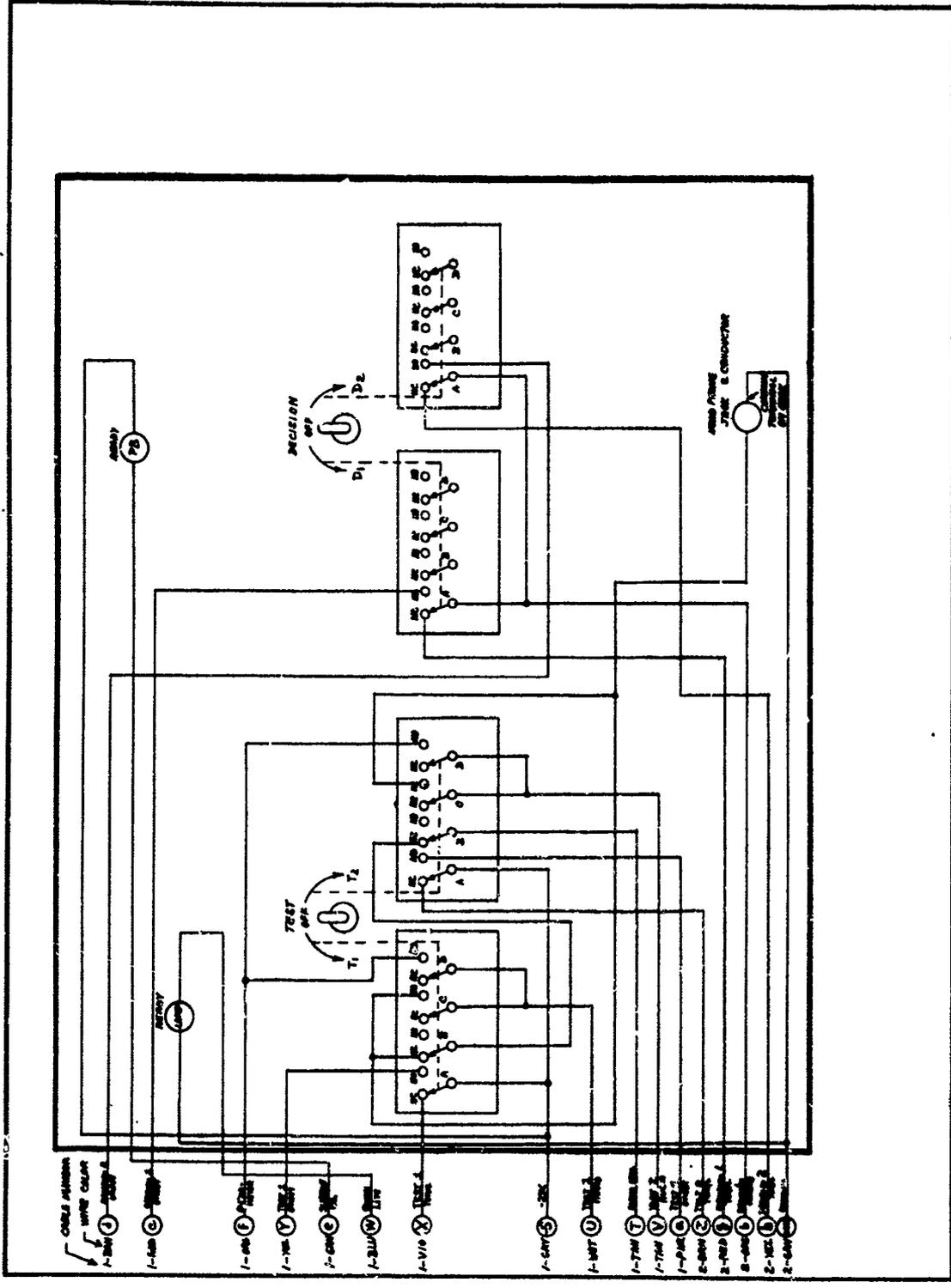


FIGURE 5
Details of Subject Response Station Circuitry

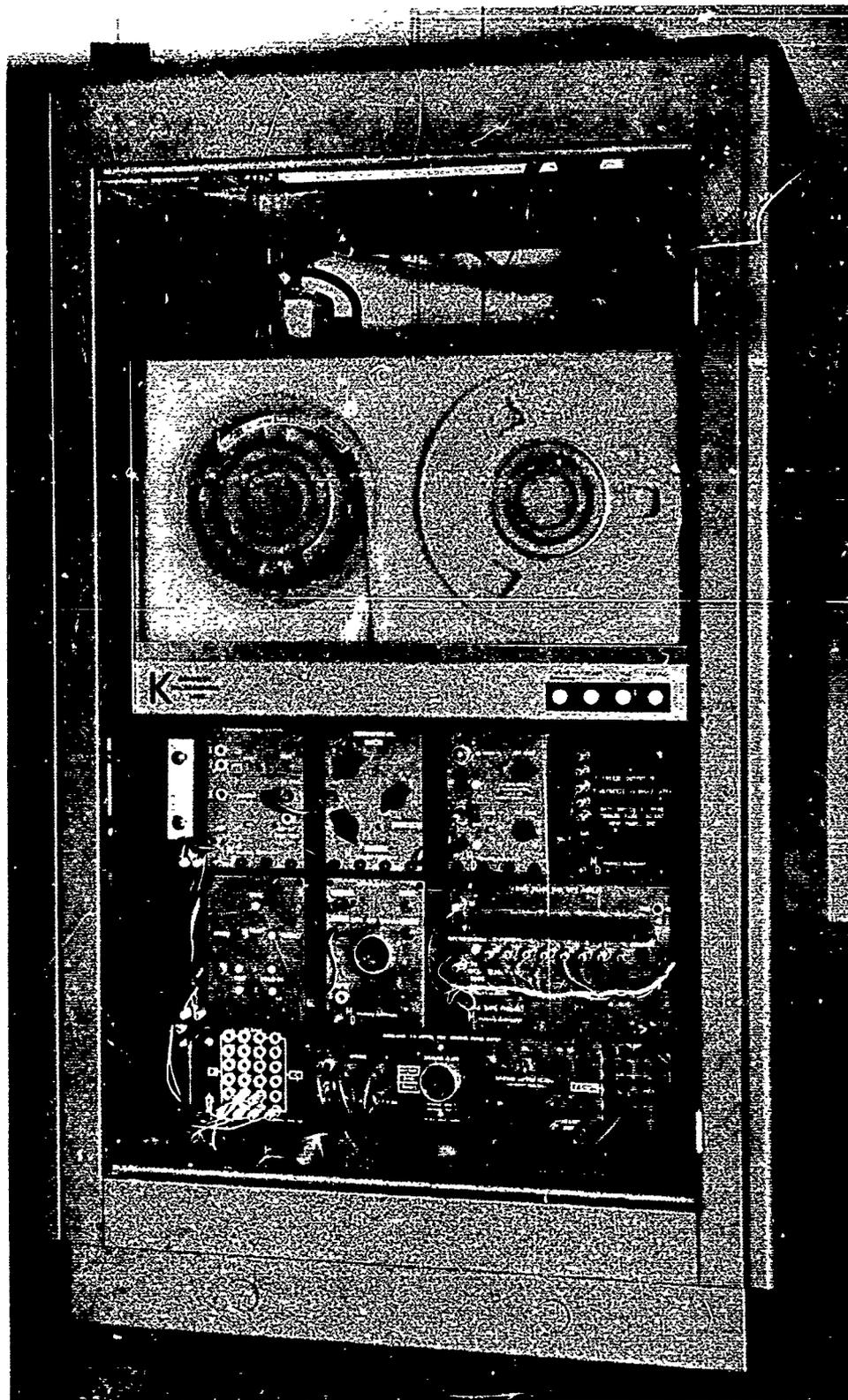


FIGURE 6
Photograph of Completed System. Console houses digital
tape deck together with logic modules and power supply.

BLANK PAGE

THE EFFECT OF UTILITY ON INTELLIGIBILITY TESTING

Numerous studies have demonstrated the modification of performance in psychophysical tasks as a function of response consequences. When appropriately varied, the administration of rewards and punishment can substantially alter a subject's perception of the experimental stimulus. This is true not only of stimuli which may be expected to be sensitive to environmental manipulation, but has been shown to affect even such "objective" situations as thresholds of visual intensity (Blackwell, 1953). It seemed therefore, totally reasonable to expect that any estimate of a subject's correct recognition of less-than-optimal speech--and thereby any estimate of the intelligibility of the speech producing system--will, in part, be a function of any accompanying rewards. Specifically, one might expect that where the utility of a correct response is increased, a reduction in error might be effected. Indeed, intuitively, one suspects that where payoff is high intelligibility increases. One need only refer to familiar anecdotal experiences of listening to an address imbedded in a distorting foreign accent. The attentive and careful listener soon improves his comprehension talents. Similarly, when the cost of an error is high, intelligibility may also be expected to increase. Again one need think only of a message transmitted to the pilot of a supersonic craft, who is less likely to confuse "mire" with "fire" in flight than he would in the comfort of an acoustic room of a laboratory.

The object of this experiment was to investigate the effect of announcing a monetary reward on the subsequent test performance of untrained Ss. Three quantized levels of vocoded speech were selected, viz., 3200, 2300, and 1400 bps.

A total of 20, 27, and 23 Ss were tested at these levels respectively. The materials were those used in Voiers' DRT tests; No. 101-list 3; 113-list 5; and 107-list 1, respectively. To these tests were added noise (via Grason Stadler Audiometer) resulting in a recorded test with S/N ratio of +6db. This value was selected in preference to others tested at db levels of 0, +2, +4, +8; the preference depended largely on providing a difficulty level closest to test results having a mean intelligibility score of 50.

Groups of subjects, with sample sizes of 5 to 20, were tested simultaneously at the study carrels housed in the Boston University Audio Visual Center. The age range of the participants was large, and no attempt was made to adjust, control, or otherwise compensate for the heterogeneity of potentially relevant variables. Each S was seated at a carrel, which contained a headset (either Telex, Educator, or Dynamic 270L) and a supply of DRT test booklets. The subjective loudness of the earphones was adjusted prior to the study and subsequent analyses did not suggest any contribution of control exerted by headset difference. When all headsets were in place the following prerecorded instructions were played:

"You are about to participate in a perception experiment. The results of this session will be used only for the internal purposes of the experiment and should in no way be interpreted as an intelligence or ability test. The experiment will consist of a series of pre-recorded words. These words will be presented to you through the earphones you are now wearing. The words will be spoken at a fairly rapid rate, and will not sound as clear to you as this message. These words also appear in the answer booklets which you have been given. You are asked to identify the word you hear by making a horizontal line through the corresponding word of the word-pair on your sheet. Please remember to draw the line horizontally through the word. Do not try to go back to a previously spoken word. Even if you are not completely sure, select a choice from each word pair as it is presented. If there are any questions, would you signal now by raising your hand. If not we shall begin."

Each subject participated in a single test session consisting of four presentations. The first series was administered primarily to familiarize him with the nature of the task. The second presentation was then administered and repeated for a third series. The warm-up effect and gain from the second series was estimated.¹ Finally, the subject was given these instructions:

"This is the final presentation. It is very important that you do well. It is to your benefit, as well, that you do as best as you can; because at the end of the experiment we will offer a reward up to \$1.00 which will be determined by your proficiency and by your improvement from performance on the previous trials. The exact amount will be individually determined later. But remember, the higher your score, the larger will be the amount of the reward."

The tests were scored using Voiers' Correction

$$\left[\frac{(R-W)}{T} \times 100 \right]$$

with omissions counted as errors.

The duration of a test session was approximately $\frac{1}{2}$ hour.

¹Continued testing for additional series under similar instructional conditions did not give evidence of comparable intelligibility improvement or residual warm-up. These determinations were derived from a series of pilot studies not included in this report.

Analysis of the data did not reveal difference attributable to the "high utility" instruction conditions. In some cases, the analysis did not give evidence of a pronounced warm-up effect. This is especially surprising considering the use of experimentally naive subjects. The explanation probably rests on two circumstances:

- (a) The extremely high subject variability, which as the error term for main effects, reduced the probability of detecting subtle changes, and,
- (b) The relatively asymptotically high performance level to begin with, which despite increased quantization does not result in corresponding intelligibility loss, provides but little room for demonstrable improvement.

Several subjects were then retested, and strongly encouraged to perform at their "very best," coupled with promise of doubled rewards. These subjects, originally stabilizing at scores below 50, then were retested at 1400 and 2300 bps, and in each instance improved to a point at least 1.5 sd above the respective group mean. This encouraging finding suggested a modified replication of the above experiment.

Under the same environmental conditions, 3 new independent groups of Ss ($n = 9, 13, 10$) were tested at 1200, 2300, and 2600 bps respectively. The age range of the subjects were comparable, as were other intellectual and socio-cultural factors. The initial procedures for test administration were the same as in the previous experiment. Under "high utility" conditions, the instructions were repeated and a sincere appeal for their best performance was added. They were also told that payment would be made immediately following the experiment.

The results of an Analysis of Variance with trend (Table 1) indicated that (a) group differences exist and (b) given the magnitude of change, no significant difference could be asserted even though improvement on trial #3 can be systematically observed. (Table 2). If statistically demonstrable, once again the relatively low ceiling available for improvement could not be meaningfully significant in any real sense. That the effect of the utility instructions is a real one, may be seen from a secondary analysis, in which each subject's performance on each test administration was transformed to ranks. The analysis of variance on ranks was then performed, with provision for trend (Page, 1963). The significance ($P < .05$) of the improved performance under the conditions of the final trial was demonstrated for the 2300 and 3600 bps subjects. The effect appears sufficiently real and replicable to warrant an intensive evaluation under varied conditions of stimulus presentation and reward payoffs. This demonstration resulting as it does from a relatively small sample size and under relatively adverse testing conditions is all the more convincing of the potential role of utility in a meaningful evaluation of intelligibility assessment.

TABLE 1
ANALYSIS OF VARIANCE FOR DRT AT DIFFERENT
QUANTIZING LEVELS

Source	df	Sum Sq	Mean Sq	F
Trials	2.	42.5625	21.2813	1.98
linear	1.	22.5625	22.5625	2.05
res	1.	20.0000	20.0000	1.92
Groups	2.	585.0000	292.5000	14.91*
Subjects	29.	568.8125	19.6142	1.83
Grp x Trial	4.	66.0625	16.5156	1.54
linear	2.	2.9212	1.4606	0.13
res	2.	63.1413	31.5706	3.03
Error	58.	622.0625	10.7252	
linear	29.	319.5161	11.0178	
res	29.	302.5464	10.4326	

* $P < .001$

TABLE 2
 MEANS (\bar{X}) AND STANDARD DEVIATIONS (σ) FOR DRT PERFORMANCE

BPS	Series 2	Series 3	High Utility
1200 \bar{X} σ	36.8 (3.0)	38.1 (3.6)	38.8 (3.2)
2300 \bar{X} σ	43.8 (4.1)	41.8 (3.0)	45.4 (2.9)
3600 \bar{X} σ	39.4 (3.3)	39.0 (3.0)	40.0 (4.9)

REFERENCES

Blackwell, H.R. "Psychophysical Thresholds," Engineering Research Bulletin, No. 36 (University of Michigan, Ann Arbor), 1953.

Page, E.B. "Ordered hypotheses for multiple treatments: A significance test for linear ranks," Journal of American Statistical Association, 58:216-230, 1963.

* * *

ACKNOWLEDGMENT

The assistance of Mr. Alan Noguee and Miss Helene Perler in the conduct of these studies is gratefully acknowledged.

SELECTED SCALING PROCEDURES

Introduction

Systematic differences in the level of a subject's performance are commonly described with reference to the statistical significance of the magnitude of such differences in response. Of additional interest is the perceived difference among the respective stimulus magnitudes. The assumption that the subjects' scaling of these stimulus values will coincide with the physical metric is only rarely tenable. The more likely situation suggests that subjects will determine inter-stimulus scale values which will markedly depart from specifiable or equal-interval dimensional units. While the underlying process used by subjects in arriving at such determinations are yet unknown, the re-scaling of stimulus values into subjectively defined distances is of great importance for any serious selection of stimuli subsets. Two strategies are advanced. The first approach offers an indirect estimate of scaled stimulus values.¹ The estimate may be extrapolated from conventional analysis of variance data matrices, and is stated in general form adaptable to a variety of focused interests. It relies upon maximizing the ratio of two quadratic forms. The scheme was not applied to the data described in Section I because of the noticeable zero slope which characterizes those data sets, and for which re-scaling would be but an academic exercise. Subject to the

¹ John E. Alman was primarily responsible for deriving the formal statement of this model. His contribution is gratefully acknowledged.

constraint that a linear relationship is assumed, the analysis is straightforward. The assumption does not violate any fundamental intuitive sense, and at the very least suggests an effective model which may be further developed for other assumed relationships.

The second approach utilizes a direct multidimensional scaling technique based on triadic judgments. Presentations of triads have been previously used with visually scalable materials. The apparent lack of such activity with speech stimuli may rest on the need for an "auditory memory"; since a successive rather than simultaneous presentation of stimuli is by necessity required. Two such studies were completed and analyzed. The results are highly encouraging from a methodological standpoint. The generalizability of the data must, however, be tempered with great caution considering the few subjects which were involved. The advantages which such "direct" scaling procedures offer include (a) efficient use of presentation trials by maximizing the amount of information obtainable in any single trial; (b) the possibility of using the technique with a variety of subjects; (c) the relative ease of data analysis (at least the mechanical processing) in view of the availability of numerous library routines; and (d) the ability to structure relevant perceptual dimensions of the speech stimuli, and the deficiencies of the subject judges. These advantages were clearly manifest even from the limited experimentation reported here.

Scale Values and Quadratic Forms

Consider an ordered set of k stimuli to which numerical values can be assigned. The numerical values may have physical meaning or may be codes to represent the order characteristic of the stimuli, i.e., 1, 2, ..., t . If the latter codes are used as a quantification of the stimuli they can be thought of as forming an equal-interval scale, although it is clear that the assumed linearity of the scale may merely reflect ignorance about the psychological values of the scale intervals. It is possible, of course, for the stimuli to have numerical values on some physical scale, but with the suspicion that the physical scale may be non-linear with respect to the psychological scale the stimuli present to the subjects of the experiment.

Regardless of which of the above may be presumed to hold, we can suppose that the stimuli can be ordered and that the numerical values to be associated with each can be represented by the vector, (v_1, v_2, \dots, v_t) . Let there be associated with each stimulus a response, X_j , and define

$$y_{ip} = v_1x_1 + v_2x_2 + \dots + v_tx_t \quad (1)$$

where $x_j = X_j - \bar{X}$. The index p represents the p th subject; the index i the i th replication or condition. $p = 1, 2, \dots, n$ and $i = 1, 2, \dots, r$. The y 's may be thought of as derived data and subjected to analysis of variance following the particular experimental design in use.

We suppose there is a particular component of the design that is of interest, say, C . To test for significance we would form the F -ratio for the component C . We now wish to find the values of the

v_j to maximize the F-ratio for C. But F is the ratio of the sum of squares for error, multiplied by the inverse ratio of the corresponding degrees of freedom. The latter is fixed for a particular design, hence, it is sufficient to maximize the ratio of the sums of squares. Taking the sum of squares of the y's generates the quadratic form $v' M v$ where M is the sum of products matrix, xx' .

Suppose C represents the sum of products matrix $\overline{xx'}$, where the mean values represent the component "of interest" in the analysis, and correspondingly E is the sum of products matrix for error. The ratio of the sums of squares can be written as

$$\theta = \frac{v' C v}{v' E v} \quad (2)$$

or,
$$\theta v' E v - v' C v = 0$$

Differentiating with respect to v and factoring yields

$$(\theta E - C) v = 0$$

or,
$$(\theta I - E^{-1} C) = 0 \quad (3)$$

A non-trivial solution exists for (3) only if

$$\left| \theta I - E^{-1} C \right| = 0 \quad (4)$$

Solution of the determinable equation above yields eigenvectors that provide the values of the v's.

The preceding outlines the general scheme of maximizing the ratio of two quadratic forms. Two well-known realizations of the scheme are principal component factor analysis and the discriminant function. If each of n subjects is exposed to each stimulus once, the data form an $n \times t$ matrix. Using analysis of variance terminology we choose to maximize the ratio of the sum of squares among row means to the sum of squares for the row \times column interaction. This is equivalent to maximizing individual differences, represented by the variation among row means, and is therefore equivalent to factoring the correlation matrix obtained by treating each column of the matrix as a separate variable.

A second realization of the general scheme is a design in which subjects are divided into groups, each group being exposed to the stimuli under different conditions. We choose to maximize the ratio of the sum of squares for conditions to the sum of squares for error. This is the classical discriminant problem, first stated by Fisher in 1936 (1936) for the two group case, later generalized to multiple groups by Rao (1965) and others.

The solution implied by (4) is quite general since C and E can represent a sum of squares for an effect and its corresponding error term in whatever analysis of variance model fits the occasion. Abelson (1960) discusses scales derived from three-way analysis of variance tables, showing that the solutions are formally identical to the discriminant function solution. In general we can say that, whatever the experimental design with which we work, we are interested in discriminating among some set of mean values; hence,

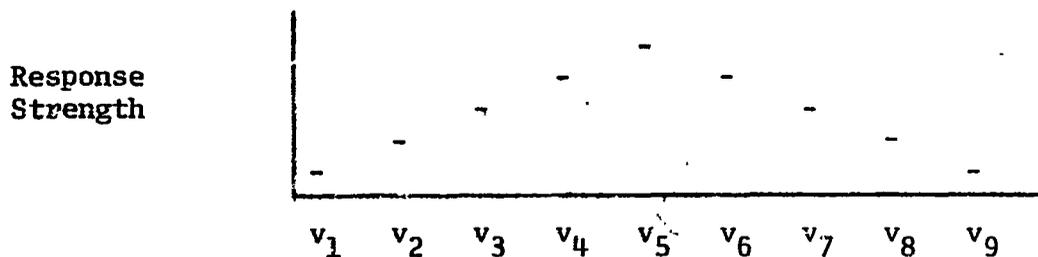
it is appropriate to use the term discriminant function in a quite general sense.

In both of the above realizations, it is not necessary to assume an order relationship among the stimuli; they can be ordered quite arbitrarily. However, the stimuli may be such that a monotonic relationship must exist; in this case we can establish the vector

$$0, v_1, v_2, \dots, 1$$

thus scaling the undetermined values of the v 's to be between zero and one. In this case, due to the value of unity to the right, the ratio of two quadratic forms generates a set of non-homogeneous equations the solution of which provides the required values of the v 's.

In the current study, the stimulus-response values can be considered as ordered but not monotonic. The two end points are considered as "anchors" and the stimulus values are presumed to be ordered as a function of the "distance" from the anchor points. Idealized, the stimuli can be arranged triangularly so that the scales are linear from each end point to the mid-point, thus:



If it is felt that the monotonicity shown schematically must exist and that the end points must be "zero points," we can place restrictions on the scaling such that the stimuli are labeled as below:

$$0 \quad v_1 \quad v_2 \quad v_3 \quad 1 \quad v_4 \quad v_5 \quad v_6 \quad 0$$

thus reducing the degrees of freedom for the scale to six. Furthermore, if logically the set of stimuli must be symmetrical about the mid-point, we can reduce the degrees of freedom to three by setting $v_1 = v_6$, $v_2 = v_5$, and $v_3 = v_4$. Less restrictive than either of the above, of course, is the scale obtained by setting the end points to zero and permit seven undetermined coefficients. However restrictive we choose to make the scale, an appropriate criterion would be to maximize the correlation between the scale value and the response variable. This is equivalent to maximizing the sum of squares due to regression, subject to a suitable restraint on the vector of v's. Graphically, the process may be conceived of as adjusting the values horizontally to coincide with the regression line, thereby altering the interstimulus distances from their original physically based spacing.

The sum of squares due to the regression of the response variable on the scale can be formally written as

$$\theta = \frac{(\text{sum of products of } x, v)^2}{\text{sum of squares of } v}$$

The sum of products term is

$$\sum_j v_j (\bar{x}_j - \bar{x}.) \quad j = 1, 2, \dots, k$$

where k is the number of undetermined scale values. The square of the sum of products becomes the quadratic form,

$$v' (\bar{x} \bar{x}') v$$

hence the solution, placing the restriction that $v'v = 1$, is formally equivalent to factoring the sum of products matrix $A = \bar{x} \bar{x}'$.

The above proposal represents the assumption of a linear relationship between the stimulus and the response. To be sure, a linear model will not always be appropriate over the entire range of the stimuli. The assumption does, however, seem reasonable if the range is taken from an anchor to the point midway between the anchors. (Note that the a priori definition of the midpoint is taken from the physical scale of the stimulus value, or post hoc from the results of the subjects' performance. The psychological midpoint may not agree with physical midpoint.) This is equivalent to two solutions each using five observational points. If the model is appropriate to the data we would expect a positive slope to the left and a negative one to the right. The question of symmetry is open. The lack of symmetry may be manifest either in a shift of the peak point from center or in different slopes around a central midpoint. It would appear that symmetry could be expected when the physical midpoint coincides with the psychological one. The two solution model, does not, however, make any such assumption. Indeed, it may be applied even when the physical midpoint is known to be other than the central value of the stimulus set. The model has particular utility insofar as it demands no additional testing time of subjects to derive scale values, nor does it substantially increase statistical analysis above the generally applied analysis of variance models.

Direct scaling by incomplete method of triads

In addition to establishing the feasibility of multidimensional techniques to the study of speech stimuli, the program of research hoped

to stimulate the use of this model for programmatic investigations of the dimensions of speech quality, and the perceptual handicaps of selected listeners to such materials. That the procedures, are in fact feasible, is evidenced by the completion of the studies without any adverse effects on the subjects or inconvenience to the experimenter. In light of the absence of any pronounced difficulties, it appears stranger still that this technique has not enjoyed a visible measure of popularity.

Among visually presented materials, chromatic stimuli, for example, have been satisfactorily described by the color cone, and the subsequent results from multidimensional scaling techniques have yielded descriptions of the perceived stimuli in accordance with these defined dimensions. For speech materials, however, no such analogous model exists, and the notion of a "typical" contour by normal or pathologically damaged listeners is not known. The eventual (though highly ambitious) goal is to provide for speech materials, a description of the auditory space as was done for visually perceived color by Helm and Tucker (1962).

Two independent studies were conducted, both utilizing the method of triadic presentations. The respective stimuli were variously processed tapes of the 'Nimitz passage' (see Appendix B of Section I). Each of six tapes was recorded on a standard quality cassette cartridge and played back on standard quality units. Precise acoustic control of the testing environment was not attempted, nor were precise audiological examinations of the subjects conducted, although no apparent hearing loss was evidenced.

Details of the technique and its development are reviewed elsewhere (Torgeson, 1958) though it may be noted here that the six stimulus tapes

were presented to the subject three at a time in all combinations of triads, for a total of $(n(n-1)(n-2)/6)$ presentations. At each presentation the subject is asked to indicate which two sound most similar and which two sound most dissimilar. With six stimuli, a total of 20 triad presentations and two judgments per presentation permit 60 interstimulus distances to be obtained for each subject.²

The resulting similarity judgments are then combined over subjects in a square matrix. The lower half of this matrix was then used as input to a multidimensional scaling program, FASCALE.³ Briefly, this program contains a number of subroutines with numerous options, among them: factor analysis, including principle axis, quartimax and varimax rotation; Kruskal's multidimensional scaling analysis with stress computation; Shepard diagrams for both the Euclidean and non-Euclidean solutions (in addition to the listing of the interpoint distances); Guttman-Lingoes smallest space analysis; and Young and Householder's T-scale analysis for departure from Euclidean metric.

The results of the study reported in Section I point to the contribution of both the nature of the material as well as the disability to the estimate or perceived quality. Accordingly, Experiment 1 used three groups of subjects: normals (n=5), aphasics (n=5), and right hemisphere brain-damaged non-aphasics (n=3). The basic analog quality

²The services of Ron VandenBossche in the conduct of these experiments are valued and gratefully acknowledged.

³A description of the program specifications and features has been published elsewhere (Guthrey et al, 1968). Although written in FORTRAN the program is quite machine dependent. These data were processed by the Michigan State University Computer Institute, East Lansing, Michigan, where it was originally written.

recording of the Nimitz passage as used in the earlier study was presented in six modified states: (1) clear (i.e., no alterations of the analog recording); (2) reversed; (3) +3db noise; (4) time-expanded to 50% with signal deterioration only and without pitch change; (5) time compressed to 40% (increasing the presentation rate of words to the limit of normal comprehension);⁴ (6) vocoded (1400bps).

The results, while not yielding to a simple or obvious interpretation, have certain pronounced implications. First, it was noted that despite the relatively few observations, discriminably different distances were derived for the different groups of subjects. It was noted that three factors adequately account for the variability, although the factors do not in all likelihood bear an isomorphic relationship to intelligibility, quality, and preference. In addition, clusters of interpoint distances were noted for the separate groups which had comparable agreement among the groups. In particular, small distances for clear-compressed-expanded and vocoded-reversed as contrasted with reversed-expanded-compressed were found. The findings do, however, contain a number of unexplained inversions and unexpected clusters which have not yet yielded to parsimonious interpretation.

The various subsidiary analyses did not substantially illuminate these complexities. It seems reasonable to assume either that the distances cannot truly be represented in Euclidean space, or that, though non-Euclidean may be imbedded in a Euclidean space of a greater number of dimensions, just as the two dimensional non-Euclidean surface of a

⁴Processing for expansion and compression was performed at the University of Louisville Center for Controlled Recording.

sphere can be isometrically imbedded in ordinary three-space. It does, however, appear possible to refine the analyses to a point where the auditory spaces for the separate groups can be constructed with some assurance of the stability of these contours. Without further study and an increased sample, these conclusions can only be regarded as tentative and suggestive.

To look closer at perceived distances for intra-vocoded materials a second study was conducted. Experiment 2 used the Nimitz passage with five normal college students. The speech samples were processed at the following levels, respectively: (a) analogue; (b) 3200 bps; (c) 3000 bps; (d) 2600 bps; (e) 1700 bps; (f) 1400 bps. The stimuli were found to distribute themselves comparably for the separate individual subjects. Furthermore, the interpoint distances gave evidence of a larger-than-usual gap in the mid-range of stimuli, where indeed a perceptible gap in vocoding levels appears. With the limited number of subjects tested, it can be stated with certainty that the stimuli do in fact scale themselves systematically in several dimensions and that to the extent to which auditory memory is requisite, the ability of untrained subjects is hardly taxed in the performance of these tests. It is, of course, difficult to assess the contribution of built-in anchors which may affect the scaling; anchors which are inherent in the subset of values which were used. Previous studies in other sensory modalities suggest, however, that while specific statistical indices may be subject to change, the general profile is not as labile.

Taken together, both these experiments indicate the promise which the triadic method holds for describing the multidimensional space in which speech materials are best defined. Of particular interest are the considerations of evaluating psychologically relevant variables such as quality, and perceptually relevant factors such as organismic pathologies. It is perhaps difficult at this time to speculate on the eventual reconstruction of the human performance in the judgment of such materials, either degraded or unmodified, but the approach does seem to offer yet another methodological and analytic alternative to investigatory paradigms in speech system evaluation programs.

REFERENCES

- Abelson, R.P. "Scales derived by consideration of variance components in multi-way tables" in H. Gulliksen and S. Messick (eds.) Psychological Scaling, New York: John Wiley & Sons, 1960.
- Fisher, R.A. "The use of multiple measurements in taxonomic problems," Ann. Eugenics, pp. 179-188, 1936.
- Guthrey, S.B., Spaeth, H.J., and Thomas, S.. "FASCALE, A FORTRAN IV Multidimensional Scaling and factor analysis program," Behavioral Science, 13 No. 5:426, 1968.
- Helm, C.E. and Tucker, L.R. "Individual differences in the structure of color perception," American Journal of Psychology, 75 No.3: 437-444, 1962.
- Rao, C.R. Linear Statistical Inference and its Applications. New York: John Wiley & Sons, 1965.
- Torgeson, W.S. Theory and Methods of Scaling. New York: John Wiley & Sons, 1958.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Educational Research Corporation 10 Craigie Street Cambridge, Massachusetts 02138		2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	
		2b. GROUP	
3. REPORT TITLE ALTERNATIVE STRATEGIES IN THE EVALUATION OF SPEECH SYSTEMS			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Scientific Final. 15 February 1968 - 14 July 1969. Approved 27 August 1969			
5. AUTHOR(S) (First name, middle initial, last name) David I. Mostofsky			
6. REPORT DATE August 1969		7a. TOTAL NO. OF PAGES 125 plus iv	7b. NO. OF REFS 70
8a. CONTRACT OR GRANT NO. F19628-68-C-0155		9a. ORIGINATOR'S REPORT NUMBER(S) ERC-P57-6901	
b. PROJECT NO, Task, Work Unit Nos. 4610-02-01		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AFCRL-69-0357	
c. DoD Element: 6240545F			
d. DoD Subelement: 674610			
10. DISTRIBUTION STATEMENT 1. Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.			
11. SUPPLEMENTARY NOTES TECH, OTHER		12. SPONSORING MILITARY ACTIVITY Air Force Cambridge Research Laboratories (CRB) L.G. Hanscom Field Bedford, Massachusetts 01730	
13. ABSTRACT <p>This report summarizes the results of a program of research concerned with the perception of degraded speech in normal and pathological listeners. The comparisons were derived from performance in a two-category judgment task where the anchor values remain accessible to the subject and under the control of the experimenter. Response decisions and frequency of anchor testings are recorded together with decision times in computer compatible tape format. In addition, studies were designed to examine selected features of utility effects and applications of unidimensional and multi-dimensional scaling procedures. The respective sections of this report describe (1) the development and applications of the modified, non-verbal psychophysical technique; (2) the design of a data collection system; (3) variations in response utility; and (4) selected scaling procedures. These investigations were shown to be particularly appropriate for assessments of communication with vocoded materials in addition to suggesting alternative experimental paradigms for the study of audiological and perceptual problems.</p> <p style="text-align: center;">END</p>			

DD FORM 1473
1 NOV 65

UNCLASSIFIED

Security Classification