METHODS OF PREDICTING USER ACCEPTANCE OF VOICE COMMUNICATION SYSTEMS

by

William D. Voiers, Principal Investigator

with the staff of

Dynastat, Inc., 2704 Rio Grande, Austin, Texas 78705

Marion F. Cohen, Project Scientist
Ira L. Panzer, Project Scientist
Alan D. Sharpley, Project Scientist
John N. Eddins, Jr., Chief Laboratory Technician
Martha S. Bettis, Publication Coordinator

FINAL REPORT

Contract No. DCA100-74-C-0056

10 June 1974 - 30 June 1976

15 July 1976

Contract Monitor: Major Orley L. Lake, USAF
Defense Communications Engineering Center

[redacted]

Prepared for

DEFENSE COMMUNICATIONS AGENCY
DEFENSE COMMUNICATIONS ENGINEERING CENTER
1860 WIEHLE AVENUE
RESTON, VIRGINIA 22090

SUMMARY

During the past five years a number of important developments in the field of narrowband digital voice communications have been achieved through the sponsorship of various government and Department of Defense agencies. To implement the coordination and evaluation of these efforts, a consortium of representatives of the Army, Navy, Air Force, Defense Communications Agency, National Security Agency, and Advanced Research Projects Agency was established by the Assistant Secretary of Defense (Telecommunications). The need for valid and reliable methods of predicting user acceptance of the various narrow band systems was recognized at the outset by the Consortium. It was acknowledged that a high degree of intelligibility, though necessary, is not a sufficient condition of user acceptance. Other more subjective factors also contribute heavily to the user's acceptance of a communication system. Although the technology of intelligibility measurement was already highly developed, no comparable technology existed for evaluating the subjective aspects of the user's reaction to system processed speech. The present project was undertaken to meet the need for such a technology. It resulted in the development and standardization of two valid, reliable and cost effective methods of evaluating the "quality" or overall acceptability of voice communication systems.

The Paired Acceptability Rating Method (PARM) was developed to serve both as a research tool and as an interim method to meet the immediate evaluation needs of the Consortium. The results of research with PARM yielded valuable information concerning the major sources of error in acceptability test results and indicated the means to their control. In particular these results showed that stable listener differences in subjective

origin constitute the major source of extraneous variance in acceptability ratings and that control of this source can be achieved through the use of appropriately selected "probe conditions." They showed further that listener differences can be most effectively evaluated by means of standard probe conditions located in the midrange of the acceptability continuum.

Various results of research with PARM contributed to the development of the Quality Acceptance Rating Test (QUART). QUART permits evaluation of the overall acceptability of a communication system and also yields information regarding the perceptual qualities which determine the degree of acceptance accorded the system.

Research conducted with QUART has provided important, if still tentative, insights concerning the nature and number of elementary perceptual qualities that determine the user's acceptance of a communication system. Subject to the results of additional research, QUART can yield predictions of acceptability based not only on the listeners direct evaluation of acceptability, but also on his evaluation of the degree to which a system is characterized by various perceptual qualities. Such predictions will be minimally affected by the personal "taste" or value systems of individual listeners or samples of listeners. QUART rating of systems with respect to various elementary perceptual qualities can be expected to have substantial diagnostic value.

Cross validation of PARM and QUART was accomplished by correlating acceptability ratings of representative systems by a sample of communication-involved military personnel with PARM and QUART ratings of the same systems by a large sample of professional listeners.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES

1.0 HISTORY OF THE PROBLEM

A number of significant advances have taken place in the methodology of speech intelligibility evaluation during the past 20 years. These are represented in particular by the Fairbanks Rhyme Test (Fairbanks, 1958), the Modified Rhyme Test (House, et al, 1965), and the Diagnostic Rhyme Test (Voiers, 1971). Such tests, to the extent that they evaluate the useful information content of a transmitted speech signal, yield results which have important implications for the overall acceptability of the signal.

Although intelligibility is unquestionably an important factor in the overall acceptability of voice communication systems, highly intelligible speech may not be acceptable in some circumstances of human communication. For example, whispered speech (synthetic or natural) can be highly intelligible, but is essentially devoid of the properties normally connoted by the term "quality." While possibly acceptable in special circumstances, whispered speech is obviously maladapted to many others.

A need clearly exists for practical, scientifically valid methods of evaluating communications equipment and devices in terms of factors other than speech intelligibility. The term "quality" is commonly used in reference to such factors, variously including and excluding intelligibility and speaker recognizability. However, quality has yet to be defined in a scientifically rigorous manner, which possibly accounts for the fact that generally acceptable methods of evaluating speech "quality" in an engineering context have also yet to be developed.

It will simplify matters to define the issue as one of overall system acceptability, and to address the issue from this point of view. Once the means of evaluating overall acceptability have been developed, it then becomes appropriate to attempt to identify the perceptual and physical acoustic correlates of acceptability. Before a valid and reliable measure of acceptability can be developed, however, several issues must be dealt with. Among the most important of these is the issue of how the errors inherent in all psychophysical procedures are to be controlled. It is appropriate, therefore, that the various types of error and the means of controlling them be reviewed at the outset.

1.1 The Control of Measurement Error

1.1.1 Random Sampling Errors - A diversity of random effects are potentially operative in the acceptability evaluation situation. However, four major sources of random variation most generally account for the bulk of the practically significant random variation in test results. Grossly, they can be identified as interindividual listener differences, intraindividual listener differences, interindividual speaker differences and intraindividual speaker differences. Of these, intraindividual speaker differences are of least immediate concern, since the use of recorded speech materials, combined with systematic selection of these materials provides rigorous control of this factor. The others, however, merit more extensive consideration.

1.1.1.1 Sampling Errors Attributable to Interlistener Variation - Listener factors, both systematic and random, are potential sources of error in any psychoacoustical experiment or test. Their impact upon test results is likely to be especially significant where a listener's rating or judgment of a stimulus property is in some degree a matter of personal taste or preference. Other things equal, methods of acceptability evaluation which solicit a direct expression of the listener's acceptance or preference will tend to be particularly susceptible to random sampling error associated with listeners. The most direct means of reducing this component of evaluation error is by increasing the size of the listener sample, but there are other means of reducing listener sampling error. Individual differences in response tendency may be independently evaluated to provide a statistical basis for the adjustment of data yielded by "deviant" subjects. For example, a listener's ratings of a standard set of reference conditions can be used to determine the extent of his tendency to rate more leniently or stringently than the typical or normative subject. His responses to experimental conditions may then be adjusted accordingly.

1.1.1.2 Sampling Error Attributable to Intralistener Variation - Errors of significant magnitude may arise from random variation in the response characteristics of a given listener. This type of variation can be reduced by replication in accordance with well-defined statistical principles. As in the case of inter-listener differences, however, seemingly random errors may have systematic origins. Depending upon the nature of the listener's task, factors such as fatigue, habituation, and learning, may contribute to intralistener variation in an acceptability rating situation. Generally, however, such effects are amenable to experimental control through careful experimental design.

1.1.1.3 Sampling Error Attributable to Interspeaker Variation - Speaker differences, particularly as they interact with system characteristics, are also potential sources of error in the prediction of system acceptability. Unfortunately, the literature dealing with this problem is quite limited. Yet to be specified are the speaker characteristics of greatest relevance to acceptability testing. Modern digital speech processing systems, vocoders in particular, are quite sensitive to speaker differences in pitch (Voiers and Smith, 1972), and to other yet-to-be-identified speaker characteristics (Voiers, et al, 1973) insofar as speech intelligibility is concerned. But it remains to be determined that the individual speech characteristics on which other aspects of acceptability depend are subject to the interaction of speaker and system characteristics.

1.1.1.4 Sampling Error Attributable to Intraspeaker Variation - It has been observed by many investigators that the intelligibility of an individual's speech varies with a number of factors, for example with level of vocal effort (Williams, et al, 1966). Inasmuch as intelligibility is an important condition of overall acceptability, it is to be expected that system acceptability measurements will be subject to some degree of variation with intraindividual speech variation. Ultimately some consideration should be given to this issue in determining the suitability of a system in the operational situation, though resolution of this issue is beyond the scope of the present project. While the effects of intraindividual speech variation are not systematically investigated, here, they are rigorously controlled by the choice of speech materials used, by instructions to the speakers, and, more generally, by the circumstances of the recording situation.

1.1.2 Adaptation Level Variation and Systematic Error - Helson (1959) has shown that much of the extraneous variation observed in the results of psychophysical experiments is ultimately attributable to variation in the individual's adaptation level (AL) for simple or complex stimulus properties.[1] His

[1] "Adaptation Level" is used in a relatively loose sense throughout this report. Certain systematic shifts can occur in the range of a listener's responses as a result of factors other than true adaptation level changes. In the case of ratings of system acceptability, such differences may result from different conceptions of the communication situation, which factor may account for observed systematic differences between ratings by professional listeners and by system users who are more familiar with the circumstances under which a system under evaluation might be actually used.

judgment of the brightness of a light, the heaviness of a lifted weight or the loudness of a sound is directly dependent on his adaptation level or subjective origin for each of the stimulus properties involved. Thus, individual differences in the response to a given stimulus event can in many cases be explained on the basis of individual differences in adaptation level for the relevant stimulus property or properties.

In summary, adaptation level phenomena have important implications for the precision of methods for evaluating speech acceptability, particularly where absolute, as well as relative, measurements of acceptability are involved. On one hand, residual AL shifts may contribute to interlistener variation. On the other hand, transient or intra-experimental shifts may increase intralistener response variation.

1.2 State of the Art in Acceptability Evaluation

Other investigators who have dealt with the problem of speech acceptability or "quality" evaluation have been sensitive to the error phenomena discussed in the previous section, and the solutions they have offered generally reflect special concern with one or several of these types of error.

The isopreference method of Munson and Karlin (1962) represents a major contribution to the study of acceptability evaluation. In this method, both a variable test parameter (loudness) and a variable reference signal (high fidelity speech and additive random noise) are used in a forced pair comparison task. The method yields a set of isopreference contours enclosing an area which represents the optimum setting of the test system with respect to loudness and noise level. From the set of isopreference contours, a "transmission preference level" is determined for the test signal, that level being simply the signal-to-noise ratio (S/N) of the reference signal that is isopreferent to the test signal.

Among the desirable features of the isopreference method are high reliability, unidimensionality of results, and the use of a physical reference scale. The method provides extremely rigorous control of adaptation level. It is, however, somewhat maladapted for use in circumastances which involve other than the simplest types of signal degradation. The use of additive random noise as the method of signal degradation may serve among other things to invite judgments of S/N ratio rather than of overall acceptability.

Rothauser, et al, (1967) developed a modification of the isopreference method in which only the reference signal is varied. This modification is substantially simpler to implement than the original method. It involves a preliminary test to determine both the optimum loudness for test signal presentation and the range of S/N ratios for the reference signals and uses the S/N ratio at the point of isopreference as its indicant of speech acceptability. An assumption underlying the Rothauser modification is that speech "preferability" varies as a monotonic function of S/N. The use of a simple reference for preferability measurements, i.e., noise-degraded speech, is desirable in that the standard can be easily described and reproduced by other laboratories. But, as in the Munson-Karlin method, the danger exists that subjects will tend to assume that their judgments are to be based primarily on the noisiness of the system under test rather than on the totality of its subjectively relevant characteristics. Individual differences in listener preference characteristics remain a major obstacle to the generalization of results, as the developers of this method acknowledge.

The relative preference method (Hecker and Williams, 1966) uses several fundamentally different types of distorted speech as references, specifically: peak clipped and bandpassed speech with reverberant echo, lowpassed speech combined with lowpassed white noise, bandpassed speech, and high fidelity speech. In a typical test run, the test system is compared with each reference condition, and the reference conditions are compared with each other. From the comparisons among reference conditions, a ten-point preferability scale is constructed. Then, from the comparisons involving the test system and each of the reference conditions, a preferability rating (1 to 10) is determined for the test system. It should be noted, however, that the coarseness with which the reference systems are scaled may be detrimental to the efficiency and precision of the method. The evaluation of any one system becomes effectively a function of degree to which the test system is preferred to a single reference condition. For example, a fairly high quality system will quite possibly be preferred to the lowest three reference conditions in all comparisons involving them. Likewise, it will always be judged less preferable than the highest reference condition (high fidelity speech). In this circumstance, the preference value assigned the system under evaluation may depend primarily on the frequency with which it is judged to be preferable to the fourth reference condition alone, which condition involves not only a particular degree but a particular type of degradation. Moreover, the confounding of degree and type of degradation in the reference signals invites a diversity of artifacts, the full implications of which have yet to be evaluated. The relative preference method would in any case appear to make extremely inefficient use of the listener's time and of the data he yields.

The unit variance method of Voiers, et al (1965) incorporates a number of novel theoretical and practical features, but was designed primarily to cope with a limited class of systems (vocoders) and could not, without some modification, be used with other types of systems. It is, in any case, extremely cumbersome to prepare, administer, and score. Moreover, it shares with other "isometric" methods a susceptibility to sampling error associated with listeners.

A simplified pair comparison method described by Coulter (1974) appears to provide relatively reliable rankings of systems. Like other pair comparison methods, however, it is maladapted to situations involving conditions of widely disparate acceptability. Like the unit variance method, it involves an extremely tedious process for the preparation of test materials.

Distinct from the relative or preference methods are the absolute methods, several of which (Richards and Swaffield, 1959; Rothauser, et al, 1971; Grether and Stroh, 1972) may be discussed as a group, since they share a number of crucial features. In all of the variations of this method the subject is directed to describe his impressions of the acceptability of the speech test signal in terms of a set of ordered categories. Typical category labels are "Unsatisfactory," "Poor," "Fair," "Good," and "Excellent." Some variations of the basic method involve a continuous scale on which selected points are labeled; others provide the subject with examples of the extreme categories in order to "anchor" his subjective scale; still others present the subject with either all, or a representative sample, of the test signals in order to orient him to the relevant range of qualities.

The absolute preference methods are often characterized by low reliability, presumably due to interindividual differences in preferred characteristics, subjective scaling factors, and adaptation level or subjective origin. Given adequate control of these variables, however, the absolute methods have a number of theoretical advantages in addition to the practical advantages of simplicity and economy. In particular, they yield "absolute" rather than relative measures of acceptability.

An investigation by McDermott (1969) contributed significantly to the methodology of speech acceptability evaluation. In this investigation, preference data and similarity judgments were obtained from relatively large samples of listeners for a set of 21 speech transmission conditions. The results demonstrated the feasibility of predicting preferability or acceptability from judgments made with respect to other subjective dimensions, a number of which were involved in judgments of similarity. An especially significant aspect of this demonstration was the finding that similarity data, unadjusted for listener idiosyncrasies, could be used to predict the results of preference judgments which were statistically adjusted for listener idiosyncrasies. This finding suggests the means of circumventing what is perhaps the most formidable obstacle to the development of valid, practical methods of acceptability evaluation: the elementary fact that listeners tend far more to agree on what they hear than on how well they like what they hear. More importantly, McDermott's results raise the possibility that measurements of what individuals perceive to be the distinguishing features of processed or transmitted speech can serve as valid bases for the prediction of acceptability by listeners, independently of the values placed on these features by the individual listener.

2.0 BASIC APPROACHES TO THE PROBLEM--PROPOSED SOLUTIONS

2.1 Basic Approaches

In light of McDermott's results, it appears that the problem of predicting system acceptability can be solved in more than one way. Two basic approaches can be distinguished.

2.1.1 Isometric Approach to Acceptability Evaluation - One approach to acceptability evaluation is the "isometric" approach, in which an evaluative or affective reaction is directly solicited from the listener. The validity of this approach rests heavily on the assumption of representative sampling--the assumption that the listener sample is representative, both qualitatively and quantitatively of the population of interest from the standpoint of personal preferences or tastes. To the extent that a listener sample values the same perceived system qualities, and to the same degree, as the typical member of the population of interest, accurate prediction of the acceptance reactions of that population can be achieved with the isometric approach. To the extent that the value systems of the two groups differ, predictions based on isometric data will necessarily be less accurate.

2.1.2 Parametric Approach to Acceptability Evaluation - A second approach is the "parametric" approach in which the experimental listener's perception, rather than his evaluation of a system or condition is used as a basis for predicting the acceptance reactions of the population of interest. The validity of the parametric approach rests on two assumptions:

1. That whatever their various preferences with respect to the perceptual qualities of transmitted speech, the experimental listener sample

and the population of interest have in common the capacity for discriminating these qualities.

2. That correlation exists--at the normative, if not the individual, level--between the perceived characteristics of transmitted speech and degree of acceptance by the population of interest.

It follows from these assumptions that even the listener who does not value (or negatively values) the perceptual qualities most valued by the population of interest can provide information concerning the degree to which an experimental speech signal is characterized by those qualities. Such information can, in turn, be used to predict the acceptance reactions of the population of interest.

Prerequisites of the development of a parametric method of acceptability prediction are (1) the development of means of measuring the relevant perceptual qualities and (2) the determination of relations between these qualities and the evaluative or affective reactions of the user population.

2.2 Proposed Solutions

To meet both the near-term and longer-term needs of DCA Narrowband Voice Consortium, both the above approaches were experimentally investigated. The end products of these investigations were the Paired Acceptability Rating Method (PARM) and the Quality Acceptance Rating Test (QUART).

2.2.1 Paired Acceptability Rating Method (PARM) - PARM is a state-of-the-art method which utilizes the isometric approach. It was initially conceived to serve as an interim method in order to meet an immediate practical need. As such, it presents a number of the problems typical of isometric evaluation methods, but it is designed to permit rigorous control and the evaluation of the major types of error commonly encountered in psychophysical experiments. The information it has yielded regarding the relative magnitudes of the various types of systematic and random error has resolved a number of issues regarding the optimal design of acceptability tests from the standpoints of scientific validity and cost effectiveness. The availability of such information greatly facilitated the refinement of PARM and the development of the Quality Acceptance Rating Test. PARM will undoubtedly contribute to further refinements in the technology of acceptability evaluation.

2.2.2 Quality Acceptance Rating Test (QUART) - QUART utilizes a combination of the isometric and parametric approaches, but was designed, subject to the results of further research and development, to function entirely as a parametric method of predicting user acceptance. It solicits an evaluative response from the listener, but also requires him to characterize a system-condition in terms of various perceptual qualities.

Both methods have been validated against a set of criterion data yielded by a large sample of operational communications personnel drawn from the Air Force, Navy, and Army. Details of these validation studies are described in subsequent chapters, following a description of the criterion data and the method of its collection.

3.0 VALIDATION OF ACCEPTABILITY EVALUATION METHODS

It is commonly observed that the acceptability of processed speech depends upon the experience, orientation and needs of the listener. Thus the reactions of the communications engineer who is heavily involved in the development of a speech processing or transmission technique are often found to be quite different from those of the casual listener or the potential system user. It is extremely important to insure that the results yielded by any acceptability evaluation method permit valid predictions of the reactions of the population of individuals who will use a system or device in the operational situation. It is essential, therefore, that the correlation between the reactions of laboratory listeners and potential system users be known. To permit the determination of this correlation, a survey was undertaken in which a large sample of potential system users was presented speech materials as processed by various state-of-the-art narrowband and broadband voice communication systems. Both the affective and perceptual reactions of the "target sample" to these systems were solicited, using, among other things, the QUART Rating Form described in Chapter 5.

3.1 Collection of Validation Data

3.1.1 The Targe Sample - A total of approximately 130 military and civil service personnel, all of whom were potential users of military communications equipment and systems, participated in the survey. From the total somewhat heterogeneous sample of available respondents, a relatively homogeneous subsample of 90 respondents was segregated for purposes of validating PARM and QUART. Only male military personnel, both officers and enlisted men, were included in the final sample. All had survived various informal checks for understanding of the task and for self consistency in performing the task.

3.1.2 Data Collection from the Target Sample - Following a brief explanation of the purposes of the survey, and of the nature of this task, Target Sample respondents were presented the following materials to which they responded as indicated.

Speech Materials	Response
One-sentence sample of each of 26 laboratory and system conditions as spoken by each of three male speakers.	Yes or no response to the question: "Would transmission of this quality be generally acceptable for purposes of routine communications in the job you presently perform?"
Twelve-sentence sample of each of 26 laboratory and system conditions as spoken by one male speaker (CH or LL).	Rating of each system-condition on 12 perceptual qualities plus rating of acceptability on a 100 point scale.
One-sentence sample of each laboratory and system-condition, as above.	Yes or no response to the question: "Would transmission of this quality be at least minimally tolerable for purposes of routine communications in the job you presently perform?"
Twelve-sentence sample of each laboratory and system-condition as above, but spoken by alternate male speaker (CH or LL).	Rating of each system on perceptual qualities and acceptability as above.
One-sentence sample of each laboratory and system-condition, as above.	Yes or no response to the question: "Would transmission of this quality suffice at least for purposes of emergency communications in the job you presently perform?"

Data obtained by the foregoing procedures are ultimately of interest from several points of view and are discussed more fully, elsewhere. Most immediately, however, they are of interest for purposes of validating PARM and QUART as used with "professional" listeners. In this connection two classes of results are of greatest relevance. These are, first, the results based on the respondents' binary judgments of system acceptability and, secondly, the results obtained from the respondents' ratings of the various laboratory and system-conditions. The development of appropriate criterion measures from these results is the primary issue to which this section is addressed.

3.2 Selection of an Acceptability Criterion Measure

The ultimate concern of a using agency is to determine the proportion of the user population for which a system equals or exceeds some level of acceptability. On the face of it, therefore, one potential criterion of system acceptability is provided by F(A), the estimated proportion of the user population for which a given communication system or condition is considered generally acceptable for purposes of routine communication. However, F(A) has several shortcomings which limit its usefulness and validity in this application. Most obvious is that F(A) provides no discrimination of relative acceptability for systems which are found acceptable or unacceptable by the entire sample of listeners or respondents involved in a given evaluation. It permits no distinction between two or more systems of sufficient but differing degree of acceptability. More generally, F(A) permits precise evaluation of relative acceptability only over a relatively narrow range of the acceptability continuum and fails to provide adequate discrimination at one or both extremes of the continuum.

The major underlying reason for F(A)'s limitations as an acceptability criterion is familiar to statisticians in the behavioral and biological sciences, and becomes evident when one examines the relevant statistical principle. Given the assumption that individual acceptance thresholds with respect to one or more underlying perceptual continua tend to be normally distributed, F(A) then represents an estimate of:

$$P(A) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}} dx$$

where P(A) is the proportion of the user population for which the system-condition is acceptable and x is the position of a system-condition on an underlying psychological continuum.

It is to be expected that x can be closely approximated by the average (or a linear transformation thereof) of a sample of listener acceptability ratings R(A). Figure 3.1 confirms this expectation, where F(A) is seen to have the expected sigmoidal relation to R(A), average acceptability rating. Specifically, F(A) is the median (for three male speakers) percentage of Target Sample members who indicated general acceptance of a system for routine voice communications and R(A) is the average acceptability rating (on a scale of 0-100) assigned the system by the same sample of respondents. (Since most of the system-conditions were found minimally acceptable for emergency use, data with respect to these criteria are of limited value in the present application. No further use was made of them for purposes of this investigation.) The curve shown in Figure 3.2 was obtained from the regression of T(A) on R(A), T(A) being the corresponding normal deviate (with arbitrary mean of 50 and standard deviation of 21.48) for each of the obtained values of F(A).

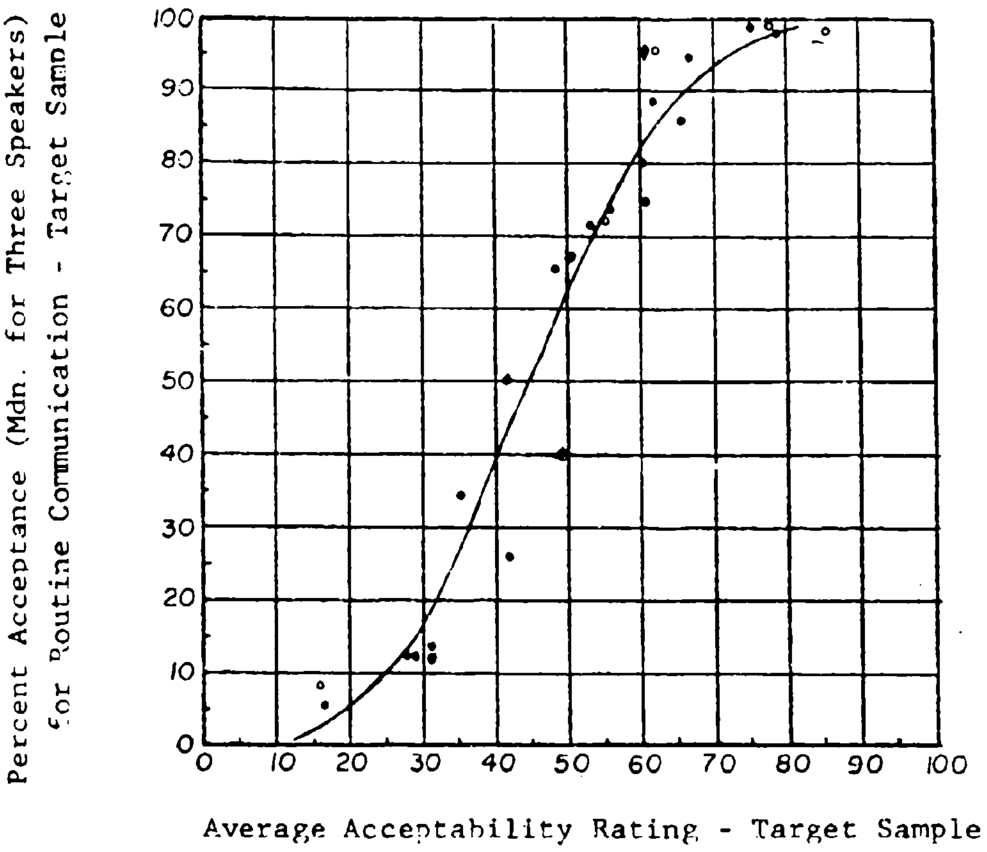

Fig. 3.1 Percent Acceptance as a Function of Average Acceptability Rating

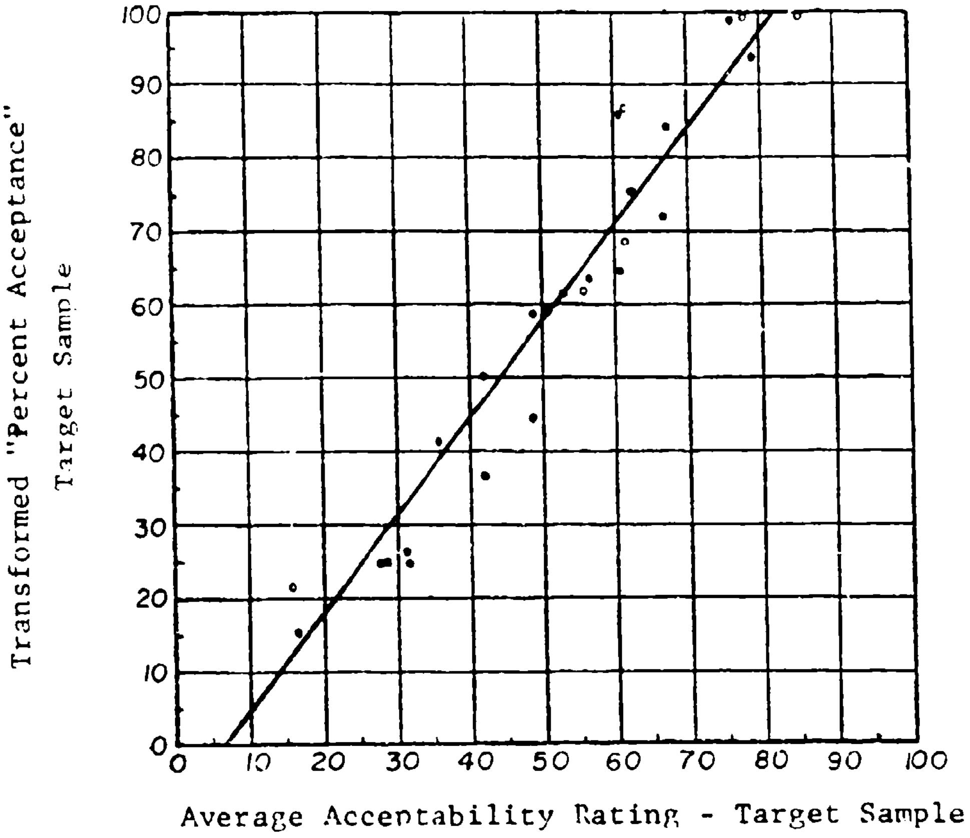

Fig. 3.2 Transformed "Percent Acceptance" as a Function of Acceptability Rating

In view of the high correlation which R(A) exhibits with T(A), and of its other desirable properties--high reliability, sensitivity of system differences over the full range of the acceptability continuum, adaptability to use with small samples, and Gaussian distribution--R(A) is clearly the best choice as a criterion of system acceptability to the target sample. Accordingly it is used as the primary basis for the cross-validation of PARM and QUART.

4.0 INVESTIGATION OF THE ABSOLUTE RATING APPROACH TO ACCEPTABILITY EVALUATION

Most methods of comparing voice communications systems from the standpoint of speech quality or acceptability have been derived in one way or another from the classical "Method of Pair Comparisons" (Guilford, 1954). However, practical considerations of time and economy have usually precluded the use of procedures which take full advantage of the potential power and sensitivity of this method. The classical method requires a single judge or subject to make many comparisons (i.e., 100 or more) of each member of all possible pairs of stimuli or conditions under evaluation. Alternatively, the method can be adapted for use with a great many subjects (i.e., 100 or so), each of whom judges each pair of conditions only once.

Although variations of the method have been developed to cope with the case of multiple judgments by multiple judges, these variations are somewhat cumbersome to use and yield results that cannot easily be generalized to the population of interest. In particular, these methods are poorly suited for use in circumstances involving small crews of judges or subjects and small numbers of judgments by each subject. No matter how precisely the reactions of a small panel of judges are evaluated, the size of the panel remains the major determinant of the generality of the results.

In the major variants of the classical method, the judge's task is simply to order the members of each pair of conditions with respect to some physical or psychological continuum such as frequency, loudness, brightness, or aesthetic acceptability. The binary data generated by this procedure are normally subjected to a transformation (e.g., "phi-gamma" or arc sin) designed to place all of the systems under consideration on an equal interval scale, the unit of which is based on intra- or inter-subject "discriminal dispersion," or other unit of psychological distance. Such transformations are feasible, however, only when relatively large numbers of judgments (say, greater than 100) are made by each judge for each pair of conditions. Normally, such scales have arbitrary origins and are thus not ratio-preserving.

Some simplication of the pair comparison method can be achieved by the sacrifice of the equal interval property, as, for example, where the figure of relative merit is simply the percent of time that each system or condition is preferred. With such figures of merit, only the ordinal properties of the acceptability scale are preserved (i.e., scale values are not linearly related to the underlying scale of acceptability). In any case, the pair comparison method in all variations is optimally suited for comparative evaluation of relatively similar conditions. Somewhat arbitrary procedures must be resorted to in scaling widely disparate conditions, particularly where one condition is universally favored or rejected. The classical method and its major variants are, as such, not optimally adapted for the evaluation of systems or conditions from an absolute standpoint.

Outside information is normally necessary to transform relative values obtained from pair comparison data to values on an absolute scale which has a psychologically meaningful zero point. One means of effecting this transformation is to employ some of the absolute rating procedures in which each condition of interest is judged in isolation using two or more ordered categories, e.g., like-dislike. Since data yielded by absolute judgments or ratings can themselves be used to scale stimuli, use of the pair comparison method for purposes of routine evaluation of system acceptability would seem, at best, to provide an uneconomical solution.

The absolute rating approach has several features to recommend it for present purposes. Although often regarded as intrinsically less reliable than various comparative methods, the absolute methods can greatly simplify the scaling problem. There is, moreover, the possibility that the seemingly poor reliability of absolute ratings derives from potentially controllable factors, in particular, interindividual differences and intraindividual shifts in adaptation level. This was a major consideration in the design and development of PARM.

There is little question that AL phenomena are operative in any speech rating situation and may give rise to significant variation in listener performance. What remained to be determined in the present case, were the practical implications of the various components of AL. A major part of the research described in the following sections was addressed directly or indirectly to this issue.

4.1 Development of the Paired Acceptability Rating Method (PARM)

PARM was designed to provide a practical, reliable, and valid method for relative and absolute evaluation of the acceptability of voice communications systems. It is an absolute rating method, but it utilizes a format that permits comparative evaluation of experimental systems or conditions. Each system-condition to be evaluated is presented under circumstances in which the listener has the opportunity, if so directed, to compare it (in two temporal orderings) with every other experimental condition involved, and with one or several "anchors" or reference conditions. For the purposes of PARM, however, listeners were not asked to make comparative ratings. The temporal ordering of conditions was designed to provide uniformity of context, as represented in particular, by the immediately preceding condition.

4.2 Experimental Evaluation of PARM

4.2.1 Materials, Method and Procedures - The test materials comprising PARM consist of a master corpus of six-syllable, phonemically controlled sentences (see Appendix A) from which a sample, or subset, is drawn for purposes of a given test administration. Although the number of experimental conditions and the number of speakers may be varied at the experimenter's discretion, a three-speaker module presented via each of four experimental transmission conditions and two reference conditions, or anchors, was employed for purposes of the present series of investigations.

From the listener's standpoint, PARM involves two successive utterances of each of 30 sentences by each speaker. The listener's task is simply to rate each utterance from the standpoint of transmission quality or acceptability, using a scale from 0 to 100. A rating of 100 indicates perfectly acceptable transmission quality, a rating of 0, totally unacceptable quality, a rating of 50, "half good enough," and so on.

The manner in which the test speech materials are presented to the listener is schematized below:

First Utterance	Second Utterance
1H	1L
2B	2A
3D	3C
4B	4H
.	.
.	.
.	.
.	.
27H	27B
28C	28D
29A	29B
30L	30H

where the numbers from 1 to 30 identify the sentence uttered and the letters identify the anchors and individual system-conditions being evaluated. Specifically, the letter H identifies the high anchor, L, the low anchor. The letters A-D identify the systems or conditions being evaluated. Where more than one speaker is used, the test speech materials for each speaker are divided into two halves and presented in a counter-balanced fashion i.e.,

$$S_{a1}$$

$$S_{b1}$$

$$S_{c1}$$

$$S_{c2}$$

$$S_{b2}$$

$$S_{a2}$$

where the letter subscript identifies the speaker and numerical subscript identifies the subset of test sentences spoken by that speaker.

4.2.2. Test Design and the Control of Adaptation Level - From the above discussion of adaptation level theory, it should be evident that the reliability of absolute ratings depends heavily on the effectiveness with which adaptation levels of individual listeners are controlled over the course of a single test as well as from one test to the next. It is clearly desirable that individual differences in residual AL be effectively minimized, whether by experimental or statistical means. Two aspects of the design of PARM are directly addressed to this problem. First is the manner in which speech samples for the various system-conditions under test are temporally ordered. Each system-condition is presented in the context of (i.e., following) every other system-condition under test. Context is thus very nearly uniform across the system-conditions being evaluated in a given PARM.

An additional contextual feature of the original version of PARM is provided by two "anchors," a high anchor and a low anchor, each of which is heard preceding (and following) each system under evaluation on the same number of occasions. The selection of anchors, particularly the low anchor, was a matter of special concern. It was considered important, first, that the anchors represent more extreme levels of acceptability than those likely to be encountered in any system-condition subjected to evaluation, and secondly, that neither anchor be uniquely distinguished by one or more perceptual qualities characteristic of a particular type of system-condition or form of speech degradation While the case of the high anchor presented no particular problem in this connection, the case of the low anchor was more complicated. Following semantic differential investigations (see Section 5 for description of the semantic differential method) involving several candidates, a low anchor was obtained by tandemming the following system-conditions:

Linear predictive coder (LPC), Longbrake, at 2.4 kbps with 1% BER; HY-2 channel vocoder at 2.4 kbps and CVSD at 9.6 kbps with 5% BER. Gaussian noise was added to give a processed speech/noise ratio of 26-28 dB lowpassed at 4 kHz. This anchor was characterized by an average acceptability rating of approximately 20 (100 point scale) and, as nearly as possible, a "perceptually neutral" status.

4.2.3 Scoring PARM Data

4.2.3.1 Standard Procedure - In principle, the scoring of PARM data is a relatively straightforward procedure. The indicated figure of merit for each condition is simply the average of the ratings accorded the condition by the listening crew. Where more than one speaker is involved, additional scores consisting of the averages associated with each speaker may also be obtained. Tests of the significance of intercondition difference may be accomplished by means of some form of analysis of variance in the case of appropriately designed experiments. Alternately, differences among haphazardly selected conditions may be tested by means of the Newman-Keuls test or a related type of test. A specimen presentation of PARM results is provided in Figure 4.1. Shown in the figure are the average ratings of system-conditions and anchors for individual listeners and for the crew. Shown in the lower part of the figure is the difference matrix used in evaluating the significance of differences with the Newman-Keuls test (see Winer, 1972).

4.2.3.2 Special Problems - Ideally, the contribution of individual differences in subjective origin and scale to the variance of rating results are small by comparison with the contributions of systematic factors. With relatively large listening crews (30 or so listeners), this situation may prevail. However, the

RATINGS FOR SYSTEMS BY LISTENERS (AVERAGED ACROSS SPEAKERS)

LISTENER #	HI ANC	SYS C	SYS D	SYS B	SYS A	LO ANC	MEAN	MEAN-A	ANCHOR
43	81.9	45.7	47.8	44.2	42.6	17.7	46.6	45.1	49.8
56	76.8	57.0	54.4	43.1	52.4	24.6	52.2	53.0	50.7
63	78.3	53.3	51.4	49.7	44.3	22.0	49.6	49.7	50.2
02	75.6	49.0	47.8	48.0	48.6	19.9	48.1	46.4	47.7
09	76.7	45.3	40.8	40.1	48.9	20.2	45.3	43.8	46.4
31	82.6	54.2	43.6	43.8	42.4	23.4	48.0	45.5	52.9
52	71.4	42.1	44.4	43.0	37.4	20.8	43.2	41.[illegible]	46.1
68	79.4	50.3	42.5	45.5	41.0	23.7	47.1	44.8	51.8
75	78.2	57.6	54.7	53.0	48.9	23.6	52.7	53.5	51.0
96	73.4	41.5	40.6	39.4	43.1	24.9	43.8	41.2	49.2
MEAN	77.4	49.4	46.6	45.5	45.0	22.1	47.7	46.7	49.6
S.E.	1.09	1.80	1.67	1.35	1.45	.75	1.01	1.37	.63

NEWMAN-KEULS TEST FOR DIFFERENCE BETWEEN PAIRS OF SYSTEM MEANS.

SIGNIFICANCE \ DIFFERENCE BETWEEN
OF DIFFERENCE \ PAIRS OF MEANS

	HI ANC	SYS C	SYS D	SYS B	SYS A	LO ANC
HI ANC	\	28.0	30.6	31.9	32.5	55.3
SYS C		\	2.6	3.9	4.4	27.3
SYS D			\	1.3	1.9	24.7
SYS B		•		\	.5	23.4
SYS A		•			\	22.9
LO ANC						\

S.E.(AVG. DIFF.) = .91

• P<.05
•• P<.01

Fig. 4.1 Specimen Set of PARM Results

economics of routine system evaluation makes it desirable to minimize the crew size requirement. Experimental evaluation of listener differences in adaptation level with commensurate adjustment of individual listener data for differences in subjective origin, offered one means to this end.

An individual's rating of the high and low anchors, common to all PARM sets, provided the basis for evaluating AL differences. To the extent that a listener is atypically lenient in his ratings of both anchors, it is a reasonable hypothesis that he is likewise atypically lenient in his ratings of the experimental systems or conditions being evaluated--that his subjective origin, or AL, is atypically low. To the extent that his ratings of the high and low anchors deviate in opposite directions from the respective normative values for the two anchors, it is appropriate to hypothesize that his subjective scale is atypically expanded or constricted depending on the manner of deviation. His ratings of the anchors can thus provide a basis for "correcting" his responses to the systems under evaluation.

It is convenient in the above connection to represent the response of the typical or ideal listener to system-condition, i, in terms of an equation of the form:

$$\bar{R}_i = \bar{A} + \bar{B}X = \bar{A} + B\ (\bar{R}_i - \bar{A}),$$

where R_i is the average or ideal rating of system-condition, i, $\bar{A}$ is the ideal listener's subjective origin; B is a slope or scale factor, (which is "1" by definition the case of the ideal listener) and X is the perceived difference between the system-condition in involved and the ideal subjective origin. To the extent that the response of a given listener, R_i differs from that of the ideal listener, R_i, such differences may be attributed to individual variation with respect to subjective origin, A, and slope or scale factor.

Given that perfectly reliable means were available for determining individual subjective origins and slope factors, the response of an individual listener, R_{ij} can be transformed to its ideal equivalent by appropriate scale and origin adjustments, i.e.,

$$\bar{R}_{ij} = A_j - (A_j - \bar{A}) + \frac{\bar{B}}{B}(R_{ij} - A_j),$$

what remains to be determined is a means of estimating A_j and B_j it was hypothesized that the individual's subjective origin deviates from the norm if the average of the ratings he assigns to the two anchors deviates from the ideal of 50. It was hypothesized that his subjective scale deviates from the norm if the difference between his average ratings of the high and low anchors deviates from 58, a historical average for Dynastat crews.

The first of these hypotheses was tested by examining the correlation between $\bar{A}_o$ and $\bar{A}_s$. Here, $\bar{A}$ is the average of many ratings made by an individual listener. $\bar{A}_o$ is the average of the ratings given by a listener to the two anchors (historically, 50) and $\bar{A}_s$ is the average of the ratings given by the same listener to the four system-conditions represented in a particular PARM. Over the course of a succession of such tests, the median coefficient of correlaiton (in this instance, also the regression coefficient) was .70.[1] The implication of this

---

[1]This assumes equal variances for average system rating and average anchor rating, which condition prevailed during the major part of this investigation. During the later stages of the investigation, the variance of anchor ratings decreased somewhat due to ill conceived instructions given the listeners concerning "typical ratings" for the two anchors.

finding is that individual differences in $\bar{A}_o$ do reflect individual differences in adaptation level, but provide less than perfectly reliable indications of such differences. Thus the most appropriate correction for individual differences in subjective origin is something less than the difference between an obtained individual value of $\bar{A}_o$ and the ideal or normative value of 50. Specifically, the indicated correction of an individual's rating of system-conditions is, on this basis, $.70\ (\bar{A}_o - 50)$. Given for example, $A_o = 60$, the best estimate of the individual's "true" subjective origin is 57, [i.e., $.70(60-50)+50$]; the indicated adjustment of his ratings of individual system-conditions is a uniform reduction of 7 points.

To test the hypothesis that variations in subjective scale contribute significantly to the variance of PARM ratings, the differences between each individual's ratings of the high and low anchors were correlated with the standard deviation of his ratings of the four system conditions involved in each PARM (The greater a listener's standard deviation, the finer his subjective scale and the greater his slope relative to the typical or normative listener). Computed on large samples (16-20) of listeners on a number of PARMs, the median coefficient of correlation was found to be .30. From these results it was concluded that interanchor rating differences reflect individual differences in subjective scale and can thus be used as a basis for a scale factor correction.

Given the normative interanchor rating difference is 58, a listener who has an interanchor difference of 68 has a finer subjective scale (steeper slope) than the average. If interanchor rating difference were a perfectly reliable indicant of an individual's subjective scale, transformation of scale would be accomplished simply by

$$\frac{58}{AD_o}\left(R_o - \bar{A}_t\right) ,$$

where AD_o is the observed anchor difference for a single individual, R_o is his response to a given condition and A_t is his true subjective origin. In fact, an observed deviant AD warrants an estimate that the individual's subjective scale is increased by $.30 \left| AD_o - 58 \right|$; that his "true" interanchor difference (AD_t) is $58 + .30 \left| AD_o - 58 \right|$. The appropriate scale adjustment factor thus becomes

$$\frac{58}{AD_t} = \frac{58}{58 + .30\ (AD_o - 58)}$$

On the basis of these findings the following equation was developed as an interim means of correcting rating data for individual differences in subjective origin and scale

$$R_i' = \bar{A}_o - .70(\bar{A}_o - 50) + \frac{58}{58 + .30(AD_o - 58)} \left[R_o - 50 + .70(\bar{A}_o - 50) \right]$$

where R_i' is the estimated rating of an ideal listener, $\bar{A}_o$ is the observed average rating of the two anchors by a given listener, AD_o is the observed difference in ratings of the two anchors, and R_o is the observed or actual rating of a condition by a given listener.

If, for example, an individual listener rates the high anchor 89, the low anchor 41, and a given system-condition 63, his adjusted rating of the system-condition, R_i', is calculated as:

$$65 - .70(65-50) + \frac{58}{58 + .30\ (48-58)} \left[63 - \left(50 + .70(65-50) \right) \right]$$

$$= 65 - 10.5 + \frac{58}{55}\ (63-60.5)$$

$$= 54.5 + 2.5 = 57.0$$

Application of the above equation serves two distinct but related functions. On one hand, it serves to reduce the effects of sampling errors which may express themselves as crew differences, particularly in cases involving small listening crews. On the other hand, it reduces the listener component of variance within crews. This effectively increases the sensitivity or power of tests for significance of differences between systems rated in separate PARMs, given the assumption of independent listener samples. Although scale adjustments may operate to increase the sensitivity of significance tests conducted on systems evaluated in the same PARM, origin adjustments will have no effects on the sensitivity of such tests.

Further research on the issue of individual differences in subjective origin and scale is clearly called for. The above adjustments served effectively, however, for the immediate purposes of the Narrow Band Consortium.

The efficacy of adjustments for subjective scale and origin differences was evident on many occasions over an extended period, in particular as such adjustments substantially increased the replicability of test results, both within and across crews. However, after six months or so, during which the listening crews had intensive exposure to PARM on a regular basis, various discrepancies in PARM results began to emerge. In particular, individual system-conditions which were subjected to repeated evaluation in varying context occasionally received inconsistent acceptability ratings. The possibility that such inconsistencies arose from contextual differences was explored but rejected. No malfunction of the playback equipment could be detected.

Although it might have been expected that the above adjustments for origin and scale shifts would offset the effects of long term adaptation level drifts, a complicating factor emerged: many subjects evidently learned to identify the anchors and to rate them in an extremely consistent manner. This tendency was undoubtedly enhanced by the fact that early in the project the subjects were apprised of the "typical ratings" for the two anchors. This attempt to "homogenize" the listening crews proved to be ill advised. The tendency of a number of listeners to assign ratings of 80 and 20 to the high and low anchors, respectively, regardless of their actual subjective scales and origins significantly reduced the sensitivity of anchor rating to individual differences in subjective origin and scale. Adjustments based on ratings of the anchors appeared to become less and less efficacious with the passing of time.

In a further attempt to find the reasons for the observed discrepancies in PARM results, a number of PARM sets evaluated over the course of the preceding six months, were reevaluated one or more times. With rare exceptions, acceptability ratings of individual systems were lower on reevaluation than on initial evaluation. Moreover, the size of the drop appeared to be related to the dates on which the evaluations took place. From these and other data it was possible to define a trend which indicated, for example, that a system-condition evaluated in late September would receive an average acceptability rating nearly nine points lower than when previously tested in June.

To verify the above trend, the multiple correlation between PARM rating and Diagnostic Rhyme Test diagnostic scores was computed for various classes of system-conditions. Multiple correlations ranging from .60 to .70 were obtained, depending upon the class of system-conditions involved. Examination of the differences between actual PARM ratings and predicted ratings revealed a pronounced trend as a function of the date of the PARM evaluation. Actual PARM ratings generally exceeded predicted ratings for system-conditions evaluated early in the six month period, but consistently fell short of predictions during the later stages of the period. The trend of these deviations as a function of PARM test date was quite consistent with the trend derived from PARM test-retest comparisons. Further confirmation of the trend was provided by test-retest results involving single system-conditions in different contexts.

Figure 4.2 represents a somewhat arbitrary combination of these various estimates of the trend, greatest weight being given to test-retest for complete PARM sets. Whatever its validity, the cause of the trend is yet to be determined. Its value for purposes of future PARM evaluations is open to question. In any case, one lesson learned from this experience is that periodic checks for longterm "adaptation level drift" should become a standard aspect of PARM procedures. As will be shown elsewhere, listener differences in subjective origin and scale tend to be extremely stable over the course of a single PARM, over a daily

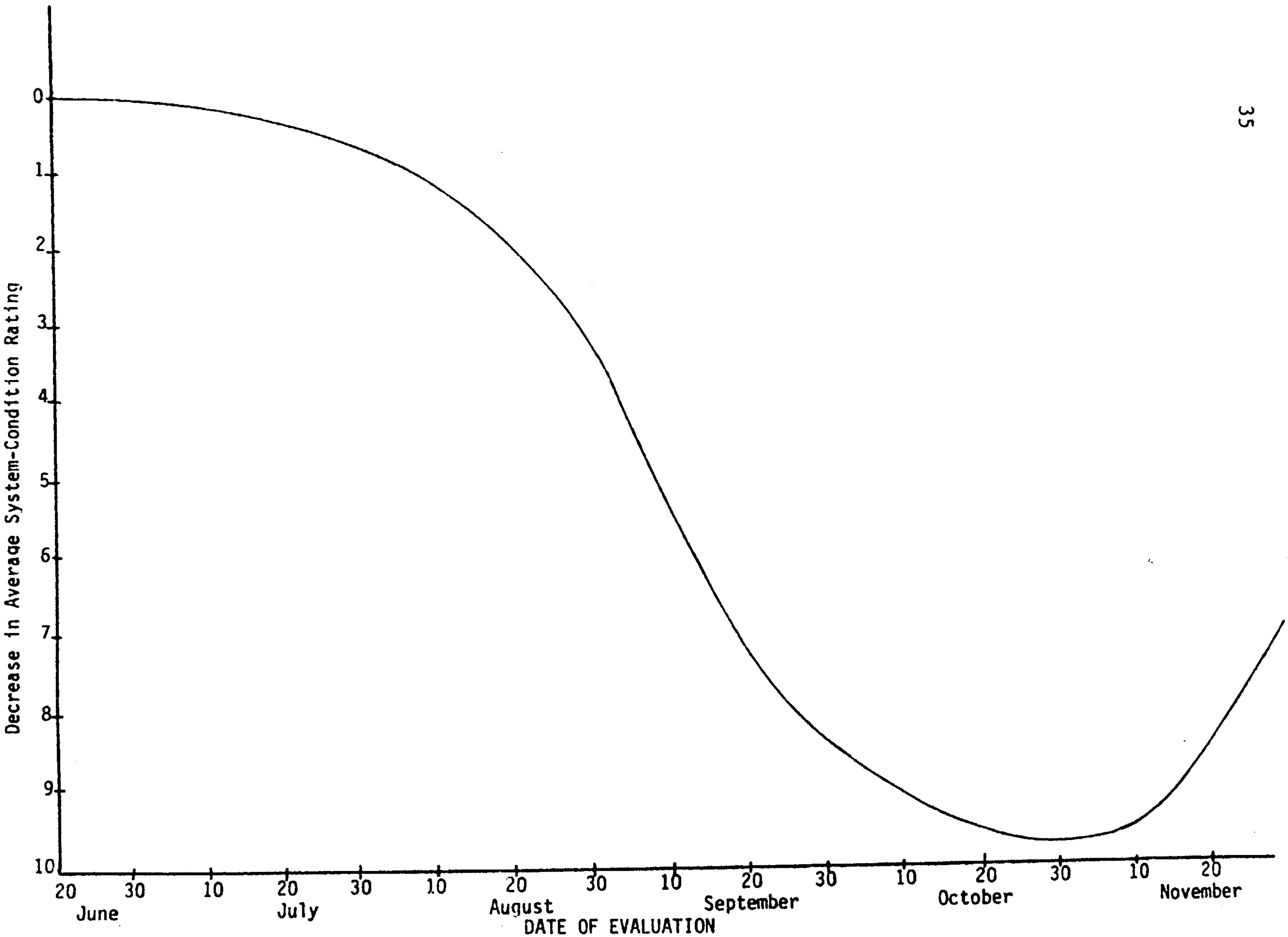

Fig. 4.2 Estimated Change in PARM Ratings as a Function of Date of Evaluation

rating session, and over somewhat more extended intervals of time. However, the possibility of longer term trends must be recognized and provided for in future PARM projects.

It should perhaps be remarked that longterm AL drift became evident only after the crews involved had been exposed to PARM for several months, during which period they were subjected to an extremely heavy PARM schedule. It is possible, that longterm AL drift will prove to be less of a complicating factor with less arduous testing regimens, but resolution of this issue must await the results of further research.

4.2.4 Reliability of PARM - A test is said to be reliable to the extent that it yields replicable or self-consistent results. The reliability of a test is a measure of freedom from error and, ultimately of resolving or discriminating power. Reliability varies in a predictable manner with test length in particular, and with redundancy in general. Since test length is a matter of some economic consequence, detailed examination of the reliability of PARM is appropriately a matter of major concern.

Efficiency in the use of testing time and resources depends heavily on the manner in which redundancy is utilized in a test. Ideally, it is allocated among the various test parameters in such a way as to equalize the sampling errors associated with these parameters. If, for example, the sampling error associated with speakers were found to be extremely pronounced in a test of system performance, the most direct remedy would be an increase in the sample of speakers and (assumming constraints on the total amount of data collected per speaker) a decrease in some other dimension of redundancy.

More comprehensive treatment of the relevant principles of experimental design is not feasible here, but the general principle is that redundancy be allocated in proportion to the intrinsic variability (variance) associated with a test parameter.

PARM is potentially susceptible to a diversity of extraneous effects, both systematic and random. Recognition of this fact is implicit in various symmetries that characterize the design of PARM. The issue to be resolved at this point, however, is whether PARM, as initially designed, makes optimal use of its redundancy. Described below is a series of investigations which bear on this issue and, more generally, on the reliability of PARM results. Because PARM test materials are impractical to assemble without the special facilities available at DCEC, it was necessary to draw the data for these studies primarily from operational system evaluations performed under the terms of Contract No. DCA100-75-C-0034. Inevitably, this served to impose various constraints on the design of validation experiments, but did permit reasonably rigorous treatment of the major issues. Except where noted otherwise, data used for these investigations were yielded by operational tests, identified as 2M, 7M, 8M, and 32M. Among them they provided a fairly representative sample of state-of-the-art digital voice systems. All were 6-speaker (male) tests, each involving four system-conditions and two anchors.

4.2.4.1 Components of PARM Variance - The design of PARM is such that PARM results are amenable to analysis of variance in which the testable effects are (among others) listeners, speakers, trials, and system-conditions. It is thus possible, to estimate the contributions of all of these effects to the variance of

PARM results. The principle employed in deriving such estimates is embodied in the relation:

$$MS_E = t\sigma_E^2 + \sigma_e^2$$

where MS_E is the mean square for an effect or treatment, (e.g., listeners) σ_E^2 is an unbiased estimate of the true variance associated with the effect, σ_e^2 is the random component and t is the number of occasions, e.g., number of ratings made by a listener, on which each state of E is represented (not to be confused with the degrees of freedom associated with the effect). Thus,

$$\sigma_E^2 = \frac{MS_E - \sigma_e^2}{t}$$

is the estimated contribution of E to the variance of a single observation. In turn, the estimated variance of an average of t observations is given by $t\sigma_E^2$. Where E is an undesirable or extraneous component, it is clearly desirable to minimize t. If, for example, σ_E^2 were the component of variance attributable to speakers in an acceptability rating experiment, increasing t would serve to increase the contribution of speaker sampling error to the test results. A reduction of t, with a commensurate increase in the number of speakers would serve to decrease the speaker effect and, generally, to increase the reliability of the test without increasing its length.

Examination of data from four representative PARM sets yielded the results presented in Table 4.1. Shown for each PARM set are estimates of the contributions of the indicated effects to the total variance of listener ratings of four system-conditions. Specifically, $t\sigma_E^2$ is an estimate of the variance

TABLE 4.1 COMPONENTS OF VARIANCE IN THE RESULTS OF FOUR REPRESENTATIVE PARM SETS

	EFFECT	DF	ERROR POOL	t	$t\sigma_\epsilon^2$					σ_ϵ^2				
					PARM 2	PARM 7	PARM 8	PARM 32	AVG.	PARM 2	PARM 7	PARM 8	PARM 32	AVG.
1	LSTNR	19	11 12 14 15	120	1968.1	6221.2	3132.7	4020.3	3835.6	16.4	51.8	26.1	33.5	32.0
2	SPKR	5	11 12 13 15	400	422.0	.0	.0	353.6	193.9	1.1	.0	.0	.9	0.5
3	CONTEXT	4	11 13 14 15	480	.0	.0	42.5	17.4	15.0	.0	.0	.1	.0	0.0
4	SYSTM	3	12 13 14 15	600	37292.4	8337.4	6648.1	3770.6	14012.1	62.2	13.9	11.1	6.3	23.4
5	LSTNR x SPKR	95	11 12 15 0	20	.0	.0	.0	.0	0.0	.0	.0	.0	.0	0.0
6	LSTNR x TRIAL	76	11 14 15 0	24	.0	.0	.0	.0	0.0	.0	.0	.0	.0	0.0
7	LSTNR x SYSTM	57	12 14 15 0	30	368.0	300.7	290.0	164.3	280.8	12.3	10.0	9.7	5.5	9.4
8	SPKR x CNTXT	20	11 13 15 0	80	.0	.0	.0	.0	0.0	.0	.0	.0	.0	0.0
9	SPKR x SYSTM	15	12 13 15 0	100	33.8	27.3	.0	30.8	23.0	.3	.3	.0	.3	0.2
10	CNTXT x SYSTM	12	13 14 15 0	120	.0	.0	.0	.0	0.0	.0	.0	.0	.0	0.0
11	LSTNR x SPKR x CNTXT	380	15 0 0 0	4	6.6	14.1	11.6	11.0	10.8	1.6	3.5	2.9	2.7	2.7
12	LSTNR x SPKR x SYSTM	285	15 0 0 0	5	6.8	4.0	8.5	.8	5.0	1.4	.8	1.7	.2	1.0
13	SPKR x CNTXT x SYSTM	60	15 0 0 0	20	15.3	71.2	43.7	73.8	51.0	.8	3.6	2.2	3.7	2.6
14	LSTNR x CNTXT x SYSTM	228	15 0 0 0	6	3.6	.0	.0	.0	0.9	.6	.0	.0	.0	0.2
15	LSTNR x SPKR x CNTXT x SYS	1140	0 0 0 0	1	28.8	31.4	28.1	33.2	30.4	28.8	31.4	28.1	33.2	30.4
16	TOTAL	2399	0 0 0 0	1	109.7	108.7	78.9	84.9	95.6	109.7	108.7	78.9	84.9	95.6

contributed by an indicated effect to an average PARM rating for the case of PARM as presently constituted. Estimates of σ_E^2, the contribution of each effect to the variance of a single unit of observation, are also shown to indicate the intrinsic variability associated with each effect. Column t shows the number of unit observations, or "trials" involving each level or case of the effect (e.g., each listener) involved. "Error pool" identifies the effects for which sums of squares were pooled to obtain an estimate of the error variance in each instance. For purposes of this analysis, it is assumed that all second and higher order interactions are insignificant--a rather strong but necessary assumption, considering that all the involved effects are fixed rather than random effects.

Although the results vary somewhat from PARM set to PARM set, some important consistencies are evident. Compared with listeners and listener x systems, all of the other extraneous effects are of negligible consequence. Much of the inherent redundancy of PARM thus appears not to be used to best advantage.

In particular, the results bearing on the importance of context are consistent with earlier findings (Voiers, 1974) that the immediately prior condition has little effect on the PARM rating of a given condition. The effect of speakers appears to be negligible, suggesting that listeners are not generally biased in their ratings by the quality of the speaker's voice. There is some indication of interaction between speakers and systems, suggesting that the various systems are not equally receptive to all voices. However, the magnitude of this interaction is not substantially greater than the random effect, as estimated by the interaction, listeners x speakers x context x systems.

Taken together, these results suggest that the reliability of PARM could be substantially increased, at no cost in total amount of data collected, by increasing the number of listeners and proportionally decreasing the amount of data collected from each listener, e.g., by dispensing with the requirement of "all possible pairings of systems-conditions." (Alternatively, the length and cost of PARM could be reduced at no cost in reliability.) However, further research on this issue is in order before instituting extensive changes in the design of PARM.

4.2.4.2 Split-half Reliability of PARM - Assuming that short-term contextual factors have virtually no impact on PARM ratings, as is indicated in Table 4.1, the second half of a PARM effectively replicates the first. The question then becomes one of whether such replication is in fact necessary. To the extent that the two halves yield equivalent results, a negative answer to this question is warranted. Two aspects of first-half - second-half equivalence are of interest. It is of interest to know, first, whether crew average ratings undergo systematic changes from the first-half to the second-half and second whether individual listeners maintain their relative positions in terms of the ratings they accord the system-conditions.

PARM sets, 2M, 7M, 8M, and 32M were used to resolve the above issues. Results of the analyses conducted for this purpose are presented in Table 4.2. Shown in the table are the average ratings given to four system-conditions by a crew of 20 listeners during the first half of each PARM and during the second half. From these results it appears that little or no rating drift occurs over the course of a PARM. In three of the four cases first-half - second-half differences were virtually non-existent. In the fourth case a larger, but statistically insignificant, difference was obtained. Further tests involving additional PARM sets failed to provide any more evidence of rating drift from first to second half.

TABLE 4.2 Split-half Reliability of Listener Ratings

(N = 20)

PARM	Mean System Rating: First Half	Mean System Rating: Second Half	Diff	"t"*	r_{ii}	$r_{\bar{I}\bar{I}}(8)$
2M	56.1	56.0	0.1	0.0	.79	.97
7M	49.0	49.0	-0.8	1.0	.89	.98
8M	51.6	51.6	0.0	0.0	.89	.98
32M	51.4	53.4	-2.0	2.17	.82	.97
Mean	52.0	52.5	.5			

*For 19 df, $P < .05$ for "t" < 2.09

Also shown in Table 4.2 are split-half coefficients of reliability for the four cases. Specifically, these are coefficients of correlation between the individual's average rating of the four systems for the first and second halves of each test. Though far from perfect, these correlations indicate a generally high degree of individual consistency from one half of a PARM test to the next. These results also bear on the problem of crew stability from one half to the next. Application of the Spearman-Brown Prophecy Formula (see Guilford, 1954, pp. 353-354) to these results provides the basis for estimating the correlation that would prevail between crew average ratings for the first and second halves of a PARM. The final column in Table 4.2 shows that for a crew of eight listeners, virtually perfect predictions of average (four) system ratings from one half of a PARM to the other could be achieved.

The most important conclusion to be drawn from these results is simply that AL's for listeners and, in turn, crews remain exceptionally stable over the course of a PARM. Data obtained from the second half of a PARM provide little additional information.

4.2.4.3 Effects of Utterance Position - Another redundant aspect of PARM stems from the fact that each system-condition is evaluated equally in the "first utterance" position and in the "second utterance" position. A comparison of the results obtained under these two conditions is thus of interest. This comparison is provided in Table 4.3. A significant systematic difference between first utterance and second utterance ratings is evident in three out of four cases. Other things equal, listeners evidently tend to rate systems more favorably when they are presented via the second utterance than presented via the first. The reasons for this difference are not clear, but

TABLE 4.3 Interutterance Differences and Correlation

(N = 20)

PARM	Mean System Ratings: First Utterance	Mean System Ratings: Second Utterance	Diff	"t"*	r_{ii}	$r_{\bar{i}\bar{i}(8)}$
2M	55.9	56.3	- .4	1.18	.94	1.00
7M	48.6	50.1	-1.5	4.10	.97	1.00
8M	51.2	52.0	- .8	2.96	.97	1.00
32M	51.9	52.9	-1.0	3.50	.98	1.00
Mean	51.9	52.8	- .9			

*For 19 df, $P < .01$ for "t" > 2.86

a reasonable hypothesis is that the greater familiarity of a sentence on second utterance enhances its intelligibility and in turn, its overall acceptability. (There are subsequent indications that inter-utterance rating differences decrease as listener gains greater familiarity with the corpus of test sentences.) But while listeners tended, systematically, to rate systems more favorably in the second utterance position than in the first, there is high correlation between listener, ratings in the two positions. At the listener level and the crew level, second utterance ratings are highly predictable from first utterance ratings. Thus, little additional information is provided by the second utterance data.

4.2.4.4 Intercondition Effects - In sections 4.2.4.3 it was shown that listener differences in first utterance ratings were highly correlated with listener differences in second utterance ratings. There is, however, an additional issue relating to interutterance dependencies which merits examination. This is the issue of the general effects of one stimulus condition on the rating of the immediately following condition. Adaptation level theory would lead to the prediction, other things equal, of a negative correlation between successive ratings by an individual listener. A highly rated initial condition should tend to depress the rating given the succeeding condition. A low quality initial condition should tend to enhance the perceived quality of the condition which follows it. Earlier research on this general issue has led to the conclusion that such effects are of generally negligible magnitude. However, a further investigation of the issue seemed warranted, and was accordingly undertaken. Data from four PARM sets (2M, 7M, 8M, 32M) were used for this purpose. These data consisted of second utterance ratings for which the preceding conditions were one or the other of the two anchors, effectively providing a "worst case" test of adaptation level stability. The test involved an analysis of variance with factorial design in which the main effects were

<u>system-condition</u>, <u>preceding anchor</u>, and <u>listener</u>. A separate analysis was performed for each of the four PARM sets (each set involved different system-conditions). In all cases, average ratings were higher when the preceding condition was the low anchor than when it was the high anchor. However, the magnitude of this effect and of the interaction of systems and context, though statistically significant (Table 4.4) in three instances, was generally quite small. Moreover, even smaller effects are to be expected when less extreme preceding conditions are involved. An example (PARM set # 7) is provided in Fig. 4.3 where the independent variable is the average first utterance rating of a preceding condition (system or anchor), the dependent variable is the average rating of the following condition, and the parameter is the identity of following condition. In no case does the average rating of the following condition vary substantially as a function of the average rating of the preceding condition, although the effect is statistically significant under extreme circumstances. These results are consistent with those of Parducci (1964) and Voiers (1974), to the effect that the extreme stimulus conditions experienced in an experimental situation do exert a pronounced effect on the subject's response to other stimuli, and that this effect tends to remain fairly constant throughout the course of a laboratory session. Subsequent exposures to extreme stimuli are not accompanied by substantial adaptation level changes. As Parducci (1964) has observed:

> "The relative permanence of this end-anchoring in simple laboratory situations may tend to obscure trial-to-trial changes in AL. It is as though the two extreme stimuli were constantly present as standards against which each of the successive stimuli are compared."

TABLE 4.4 Effects of Immediate Context (preceding condition) on PARM Ratings

	Source	Degree of Freedom	Error	F-Ratios for PARM Sets* 2M	7M	8M	32M
1.	SYSTEM	3	(5.)	68.3	18.2	14.5	5.9
2.	CONTEXT (preceding anchor)	1	(6.)	8.2	4.1	6.4	12.4
3.	LISTENERS	19	--				
4.	SYSTEM x CONTEXT	3	(7.)	1.4	.7	5.7	5.9
5.	SYSTEM x LISTENERS	57	--				
6.	CONTEXT x LISTENERS	19	--				
7.	SYSTEM x CONTEXT x LISTENERS	57	--				
	TOTAL	159					
	Mean rating difference ("low anchor preceding" minus "high anchor preceding")			1.8	.7	1.3	2.5

*For 3 and 57 degrees of freedom, $P < .05$ for $F \geq 2.76$ and $P < .01$ for $F \geq 4.13$; for 1 and 19 degrees of freedom, $P < .05$ for $F \geq 4.38$ and $P < .01$ for $F \geq 8.18$.

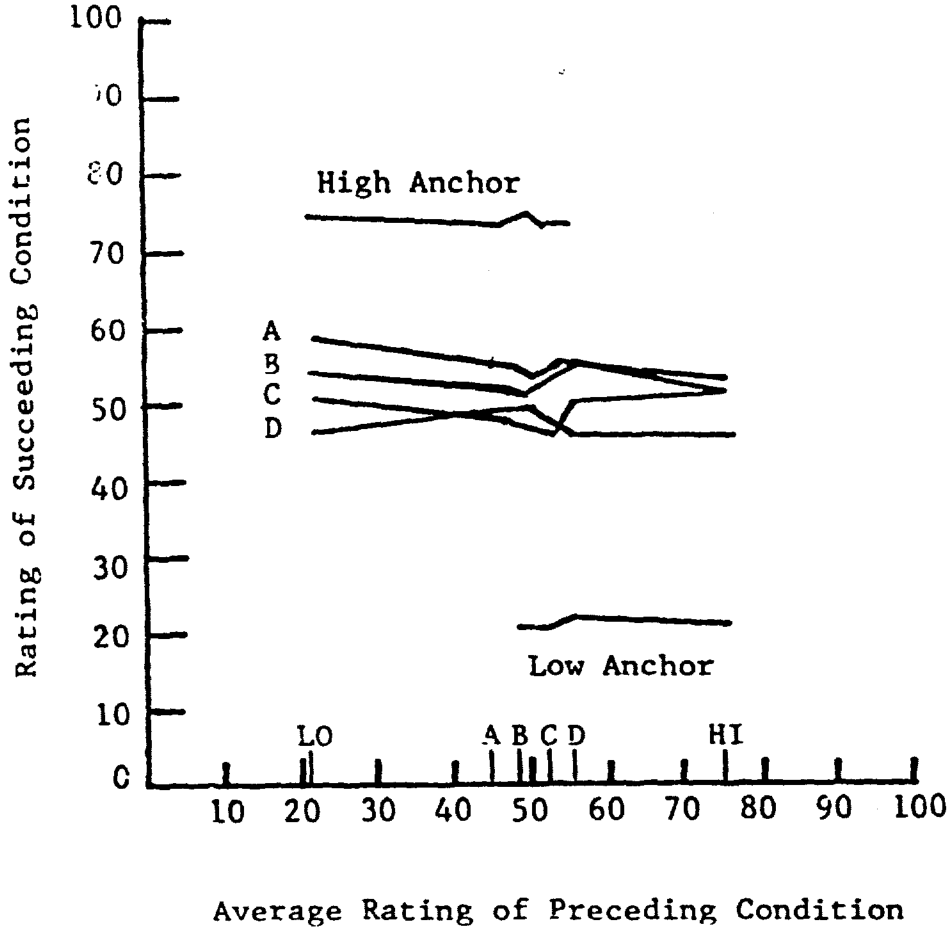

Fig. 4.3 Second Utterance PARM Ratings for Six Systems as a Function of the Preceeding System Rating (PARM 8M)

Seen in the above light, the practice of pairing all systems would appear to constitute a fairly inefficient use of resources. It would seem necessary, at most, to insure that all systems under evaluation were preceded on an equal number of occasions by each of the two anchors.

4.2.4.5 Inter PARM Reliability - From the results of the foregoing analyses it can be concluded that individual and crew adaptation levels, as measured by average system ratings, remain quite stable over the course of a PARM testing session. Intraindividual variation in PARM ratings is either negligible or adequately controlled by the design of PARM. Remaining to be answered are questions concerning listener and crew stability over longer periods of time. To resolve this issue, a crew of 20 listeners was subjected to two administrations of a representative PARM set (335A, 3 male speakers) during the same testing session. The first of these administrations was at the beginning of a routine 4½-hour testing session; the second, near the end. The crew participated in various other routine tests during the intervening period. Table 4.5 shows the average rating received by the four system conditions and two anchors under each administration.

Because of the possibility that ratings of the two anchors were subject to the extraneous influences discussed earlier, the two administrations were compared using data for the system-conditions only. A test for the significance of mean differences yielded a "t" of 0.95 which does not approach statistical significance. The coefficient of correlation between individual listener's mean system-condition ratings on the two administrations was .90. When the Spearman-Brown formula is applied to estimate the correlation to be expected between crew means on repeated administration, this coefficient becomes .99 for the case of an 8 member crew. The stability of PARM results over the course of a testing session appears, therefore, to be extremely high.

TABLE 4.5 Intrasession Stability of PARM Results

(N=20)

Condition	First Administration	Second Administration	Diff	"t"	r_{ii}	$r_{ii(8)}$
High Anchor	80.8	80.8	0.0			
D	54.3	55.0	-0.7			
A	42.0	41.8	-0.2			
C	41.7	43.4	1.7			
B	39.7	39.6	-0.1			
Low Anchor	20.9	19.9	-1.0			
MEAN (All conditions)	46.6	46.7	-0.1			
MEAN (Systems only)	44.4	44.9	-0.5	.95	.90	.99

For 19 df, $P < .01$ for "t" ≥ 2.09

4.2.4.6 Effects of Instruction - In view of the dramatic long-term changes in listener performance that occurred over the course of this project, it was of some interest to know the effects of instructions upon listener behavior in the PARM situation, particularly as the instructions received by individual listeners (and/or their comprehension of these instructions) varied somewhat over the period of time involved. Accordingly, an investigation was undertaken in which an attempt was made to evaluate the extremes to which listener performance might reasonably be affected by instructions. The speech materials used for this investigation were provided by PARM sets 180 and 181, both of which were subjected to a fixed amount of intermodulation distortion before presentation to the listeners. (This last feature is not relevant in the present context, having been introduced for purposes of another experiment.)

Two crews were employed. One crew was administered PARM set 180 on two occasions, being instructed on the first occasion to "rate as leniently as you conceivably ever have during the course of your experience with PARM." Following a 30-minute break, this crew was again administered PARM set 180, being instructed on this occasion to "rate as stringently as you ever conceivably have during your experience with PARM."

The second crew was administered PARM set 181 in a similar fashion, except that the time order of the two instructional conditions was reversed from that of the previous case. The results of this experiment are summarized in Table 4.6. From the table it appears that the instruction given the subject can, in the extreme, increase or decrease his effective adaptation level on the order of six rating points. Although the obtained correlation between averages for individual raters under the two

TABLE 4.6 Effect of Instructions on PARM Ratings

	Mean System Rating				
	"Stringent" Condition	"Lenient" Condition	Diff.	"t"	r
PARM 180(N=9)	30.4	42.6	12.2	4.54	.06
PARM 181(N=7)	29.2	41.1	11.9	8.47	.92

For 8 df, $P < .01$ for $t \geq 3.36$; with 6 df, $P < .01$ for $t \geq 3.71$

conditions was drastically reduced by a single deviant listener in the case of PARM 180, the truc correlation appears to be quite high: individuals and crews respond in a relatively uniform manner to instructions regarding the rating "set" they should adopt.

In view of the fact that differences in the instructions given subjects at different times in the course of this project never approached the extremes represented here, it seems highly unlikely that changes in listeners' conceptions of their task could have accounted to a significant extent for the long term adaptation level drift (implied by a 10-point drop in average ratings) described in Section 4.2.3.2.

4.2.4.7 Evaluation and Control of Listener Differences - From the various results described in the foregoing sections it is evident, on one hand, that individual differences in adaptation level represent the major source of sampling error in PARM ratings. On the other hand, there is substantial evidence concerning the stability of individual adaptation level, both over time and over a diversity of experimental conditions. Taken together, these results attest further to the feasibility of "calibrating" listeners and, in turn, of adjusting rating data to compensate for such differences. The use of high and low anchor ratings for such purposes was in fact instituted as part of the standard PARM scoring procedure quite early in the program. However, the question of whether ratings of anchors provide the optimal bases for evaluating the prevailing adaptation levels of individual listeners remained to be determined. Accordingly, further research on the issue was undertaken using data from PARM sets 2M, 7M, 8M, and 32M yielded by a crew of 20 listeners.

On the hypothesis that individual adaptation levels remain stable during the course of a single PARM, individual differences in ratings of the anchors and system-conditions should be correlated to some degree. The question then arises as how best to detect individual differences in adaptation level. Factor analysis provides a means of resolving this issue.

For each of the PARM sets, the correlations among individual listener's ratings of the two anchors and four experimental system-conditions were determined. The obtained correlation matrices were then subjected to a principle axis factor analysis. The results of these analyses are summarized in Table 4.7.

Uniformly high positive loadings of anchors and system-conditions on Factor I serve to identify this factor as <u>adaptation level</u> or subjective origin. The implication of this configuration of loadings is that listener differences in ratings of all conditions are subject to a common influence: knowledge of an individual's deviance in rating any one condition thus has value for predicting his deviance in rating any other condition. These results are consistent with earlier findings regarding the correlation between average anchor ratings and average system ratings, but they yield several important additional insights.

One inference to be drawn from the results in Table 4.7 is that the high and low anchors do <u>not</u> provide the best possible means of evaluating individual adaptation levels. The basis of this inference is to be found in the relatively low Factor I loadings of the anchors in all four cases. The high loadings of the system-conditions which fall near the midrange

TABLE 4.7 Factor Structure of PARM Ratings

(N=20)

	FACTOR LOADINGS									
	Factor I					Factor II				
PARM Set	2M	7M*	8M	32M	Mean	2M	7M*	8M	32M	Mean
Condition										
High Anchor	.36	.40	.67	.57	.50	.84	.88	.60	.77	.77
System A	.63	.91	.88	.87	.82	.41	.11	.14	.24	.17
System B	.86	.87	.89	.95	.89	-.02	-.16	.12	-.06	.03
System C	.88	.82	.94	.90	.89	.08	-.50	-.09	-.24	-.18
System D	.84	.81	.86	.88	.85	-.28	-.49	-.26	-.26	-.29
Low Anchor	.53	.54	.39	.62	.52	-.70	-.70	-.84	-.36	-.64
Percent Trace	.50	.56	.63	.66	.59	.24	.29	.19	.14	.21

* Original factor axes arbitrarily rotated.

of the acceptability continuum indicate that midrange conditions are better adapted for purposes of sensing individual differences in adaptation level. Factor I loadings in the .85 - .95 range serve, in fact, to suggest that a single "midrange anchor" could serve quite effectively for purposes of calibrating individual listeners. Knowing individual ratings of such an anchor would permit the investigator to account for (and adjust for) something on the order of 81% ($.90^2$) of the sampling error associated with individual differences in adaptation level. By contrast, the optimal combination of high and low anchors would, at best, suffice to account for approximately 52% ($.50^2 + .52^2$) of this component of variance.

An examination of the pattern of loadings on Factor II reveals this factor to be a subjective scale factor. Specifically, high loadings (though of opposite sign) uniformly exhibited by the high and low anchors indicates that listeners differ in terms of the subjective scales to which they reference their ratings. Other things equal, the listener who tends to be more extreme in rating at one end of the scale also tends to be more extreme in rating at the other end. Given no listener differences in adaptation level, one would thus expect to find a negative correlation (or factor loadings of opposite sign) between ratings of the high and low anchors. The pattern of Factor II loadings thus indicates that a substantial amount of the listener component of variance in PARM ratings can be attributed to individual differences in subjective scale and that the interanchor range for individual listeners can provide a means of controlling this subcomponent of variance. It should be noted, however, that the practical benefits of such controls will tend to be rather limited, except in circumstances involving system-conditions falling near the extremes of the acceptability continuum.

4.2.4.8 Evaluation and Control of Speaker Factors - As noted in Section 4.2.4.1, the magnitude of the speaker's contribution to PARM variance, is small, compared to the contribution of the listener. However, its statistical significance was an unresolved issue. Further analysis of the data from PARM sets 2M, 7M, 8M, and 32M yielded results which bear on this issue. They are presented in Table 4.8. In two of the four cases the main effect for speakers is significant at the .01 level. In all four cases the interaction of speakers and systems are significant. Evidently systems vary in their receptivity to individual voices. It should be noted, however, that the sample of speakers involved here was in no sense a random sample. Rather, it was deliberately selected to provide representation of extremes with respect to fundamental frequency. The practical significance of these results is, therefore, still open to some question. A less rigorous examination of data from a large number of PARM sets revealed that speaker variation, either within or between sexes, is rarely of magnitude comparable to that associated with listeners or system-conditions. However, further research on this issue is clearly in order.

4.3 Interim Conclusions and Recommendations for the Use of PARM

From the diversity of experimental results described in the preceding sections two major principles can be clearly discerned.

1. Listener differences account for the major component of the extraneous variance of PARM results. By comparison the contributions of other systematic factors is negligible.

2. The listener component of variance in PARM test results has its origin primarily in stable listener differences in subjective origin or adaptation level, which differences are eminently subject to statistical evaluation and control.

TABLE 4.8 Evaluation of Speaker Contribution to PARM Variance

	Effect	df	Error M.S.	F-ratios* PARM 2M	PARM 7M	PARM 8M	PARM 32M
1.	SYSTEMS	3	(5.)	80.55	21.85	17.81	15.45
2.	SPEAKERS	5	(6.)	7.76*	1.48	1.60	6.14*
3.	LISTENERS	19	--	-----	-----	-----	-----
4.	SYSTEMS x SPEAKERS	15	(7.)	4.01	5.56	3.29	6.04
5.	SYSTEMS x LISTENERS	57	--	-----	-----	-----	-----
6.	SPEAKERS x LISTENERS	95	--	-----	-----	-----	-----
7.	SYSTEMS x SPEAKERS x LISTENERS	285	--	-----	-----	-----	-----

* For 3 and 57 df, $P < .01$ for $F \geq 4.14$

For 5 and 95 df, $P < .01$ for $F \geq 3.24$

For 15 and 285 df, $P < .01$ for $F \geq 2.10$

Given that the means of controlling the listener factor can be found, PARM can be expected to provide extremely reliable estimates of system-acceptability for the population represented by the experimental listener sample. Realization of this expectation can be facilitated if cognizance is taken of a number of secondary or corollary principles that have also emerged from the results of research conducted during the period of this contract. The more important of these are discussed below.

4.3.1 Use of anchors, probes and reference standards - It is evident from an accumulation of results that the function of anchors and the function of reference standards in rating situations are quite different. Reference standards are properly used to achieve experimental control of extraneous variance in psychophysical experiments. To this end, the identity and function of reference standards are normally made explicit to the experimental subjects, who may or may not be required to evaluate the standards themselves.

By its mere presence an anchor exerts some degree of experimental control of adaptation level. Anchors can also be used to achieve some degree of statistical control of extraneous variance, in that the subject's response to an anchor may permit statistically evaluation of, and correction for, intra and inter-listener variation in AL. For such controls to be most effective, however, the listener must be unconstrained in his response to an anchor, as experience in the present project has confirmed. In the present case an attempt was made to experimentally reduce individual differences in subjective origin by apprising listeners of the historical ranges of the ratings given the high and low anchors. While this procedure was undoubtedly efficacious in some respect, subsequent results clearly indicate, that it substantially reduced the value of anchor rating for purposes of sensing residual individual differences in subjective origin.

Following the receipt of information concerning the historical ratings of the two anchors, some listeners effectively changed their subjective origins and response scales when responding to the anchors, but were unable to maintain the same frame of reference when rating the system-conditions involved. These findings attest to the validity of the adaptation level concept, for the listeners evidently continued to rate system-conditions in relation to stable adaptation levels, even while artificially changing their modes of response to the anchors. The value of anchor ratings for detecting AL differences was, however, greatly reduced under such circumstances.

It is possible that some benefit is to be realized by identifying the extreme anchors for experimental listeners without indicating "appropriate" ratings of these anchors. A wealth of evidence indicates that such procedures will effectively stabilize the rating behavior of the individual listener. There remains, however, the problem of stable listener differences in adaptation level, which differences make acceptability ratings highly susceptible to listener sampling error.

It will simplify matters, somewhat, if a terminological refinement is introduced at this point. Specifically, it is suggested that "anchor" be reserved for extreme conditions whose primary function is to exprimentally reduce intraindividual variation in adaptation level. The term, probe, will be reserved for conditions used primarily to sense interindividual differences in adaptation level, to the end of permitting retrospective statistical adjustments for such differences.

Conditions designed primarily to serve the anchoring function may, in fact, have some value as probes if no constraints are placed on the listener's responses to these conditions. However, the various results described above attest to the superiority of midrange conditions as probes. Whatever use is made of the extreme anchoring conditions, the inclusion of one or more midrange probes would thus seem to be highly desirable in the case of PARM or similar methods of acceptability evaluation.

In summary, the results available to date indicate that the reliability of PARM can be significantly enhanced by the use of two extreme anchors and one or more midrange probes.

4.3.2 Feasibility of Listener Selection as a Means of Enhancing the Reliability of PARM Results The contribution of listener factors to the variance of PARM results has been dealt with extensively in the preceding sections. The evidence, both implicit and explicit, leaves little doubt that control of this factor can significantly enhance the reliability of PARM. Anchors and probe conditions offer one means of achieving at partial control of this factor, but additional means are available. One is through the astute selection of listeners, the feasibility of which is attested to by a remarkable degree of stability over both the short and long term that characterizes the performance of the typical listener.

A series of studies has shown that the residual, or steady state, adaptation level of relatively unselected listeners can vary over a range of 20 points on the acceptability continuum. (The most tolerant listener among Dynastat's crew of 40 listeners consistently rates systems 20 point higher than the least tolerant listener on the crew). Because of the self consistency of the typical listener, however, it is possible to select a subsample of listeners for which individual AL's (as reflected in their ratings of a "probe PARM-set") have a relatively restricted range.

The desirability of a standard procedure for preselection of PARM listeners seems beyond question at this point. The possibility remains, however, that further refinement of PARM can be achieved by post-experimental selection, i.e., by means of procedures for determining that individual participants in a test have performed in a consistent fashion, and that their data have been accurately evaluated. One such procedure that has been employed with some success involves comparing the individual listeners actual rating of a system condition with an expected value derived as follows:

$$E_{ij} = \bar{A}_j + \bar{A}_i - \bar{A}_{ij}$$

where $\bar{A}_j$ is the average of all listeners ratings of the jth condition, $\bar{A}_i$ is the average of the ith listener's ratings of all conditions, and $\bar{A}_{ij}$ is the average of all listeners' ratings of all conditions.

$$S.D._i = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (A_{ij} - E_{ij})^2}$$

thus becomes a measure of the extent of the ith listener's variability with respect to himself and to the crew as a whole. It can serve effectively as a criterion for detecting listeners who have lost their places during the test, whose data have not been accurately transcribed, or who simply performed in a generally erratic manner during the test. However, it should be noted that $S.D._i$ is sensitive to true interactions of systems and listeners. It is also sensitive to individual differences in subjective scale and must, therefore, be used with some discretion when applied to data which have not been adjusted for such differences.[2] Somewhat arbitrarily an $S.D._i$ of greater than 7 has been employed with some effectiveness as a basis for post-experimental rejection of listeners in the present project.

In summary: The reliability of PARM results can be significantly enhanced by careful selection and calibration of listening crew members and by the astute use of systematic procedures for post-experimental rejection of inconsistently performing listeners.

4.3.3 Role of the Speaker - The relevant data available during the course of this project do not permit unequivocal conclusions concerning the importance of the speaker as a factor in PARM results. It can be said, at least, that speaker factors are of substantially less consequence in the acceptability rating situation than in the intelligibility testing situation. Inasmuch as intelligibility, is a correlate of acceptability, it is possible that speakers affect acceptability measurements primarily through their effects on intelligibility. Further research will be needed to resolve this issue. For the present, the use of multiple speakers is recommended.

4.3.4 Miscellaneous Experimental Considerations - Although it was reported in Section 4.2.4.5 that listener performance did not deteriorate or otherwise change to a significant degree over the course of a 4½-hour listening session, it should be noted that these results were obtained under more or less ideal conditions. Listeners participated in total of only four three-speaker PARMs during the course of this session. These PARMs were interleaved with several DRTs which resulted in "duty cycle" of approximately 40%.

---

[2] The introduction of this checking procedure antedated investigations of individual differences in subjective scale. Subject to the results of further research on such differences, the checking procedure can be easily modified to remove the effects of systematic scale differences.

Experience has shown that subject morale and performance deteriorate significantly if the PARM test load substantially exceeds the equivalent of five three-speaker PARMs during a normal 4½-hour session. On one occasion early in the course of this project a specially selected crew was administered a total of eight three-speaker PARMs during the course of a 4½-hour session. The reactions of the listeners to this procedure took the form of one resignation, one refusal to participate beyond the fifth or sixth PARM, and vociferous complaints from the remaining crew members. Inspection of the data revealed excessive "lost places" and general deterioration of performance beginning with the sixth or so PARM. Clearly, PARM makes extremely rigorous intellectual and attentional demands on the listener, and his capacity to maintain a stable level of discriminative performance is definitely limited. In view of this consideration the extraneous redundancy of PARM becomes an even more crucial issue.

In summary, modifications which lessen PARM's demands on the listener's attentive capacities are clearly desirable. In the meantime, listener exposure to the original version of PARM should be limited to the equivalent of five three-speaker PARMs per 4½-hour session, with or without interleaving of other tests such as the Diagnostic Rhyme Test. (By contrast with the 25-35% duty cycle that listeners can tolerate with PARM, a 50-60% duty cycle is comfortably tolerated in the case of the Diagnostic Rhyme Test.)

4.4 Predictive Validity of PARM

On the hypothesis that both PARM and QUART provide valid indications of system acceptability a high degree of correlation between the two measures is to be expected. In this connection it was noted first that the original professional listener sample used with QUART was, but for a difference in adaptation level, highly correlated with the Target Sample in its perception and evaluation of the sample of laboratory and system conditions employed. It was noted, further, that a number of factors undoubtedly operated to reduce the reliability and validity of the QUART data obtained from the target sample. Accordingly it was decided that a combination of data from the two samples would provide a more valid estimate of the "true" acceptability levels of the sample of conditions involved. From such a combination a superior criterion is provided for purposes of validating PARM.

Specifically, acceptability ratings of the system conditions by the original professional listener sample were transformed to yield a new variable with the same mean and standard deviation as the distribution of acceptability ratings by the target sample. The transformed value for each system condition was then averaged with the average acceptability rating accorded it by the target sample, and these averages used as criteria for testing the predictive validity of PARM.

During the term of this project, composite criterion data and PARM data were available for a sample of only 20 system-conditions. However, the results presented in Figure 4.4 leave little doubt as to the fundamental validity of PARM. An extremely high correlation would have been obtained but for the two deviant cases (CONUS Median Voice Grade and APC with 5% BER). In view of the time elapsed between the processing of the PARM speech test materials and the QUART speech test materials, it is a tenable hypothesis the systems involved were not functioning in the same fashion on both of the occasions in which they were involved.

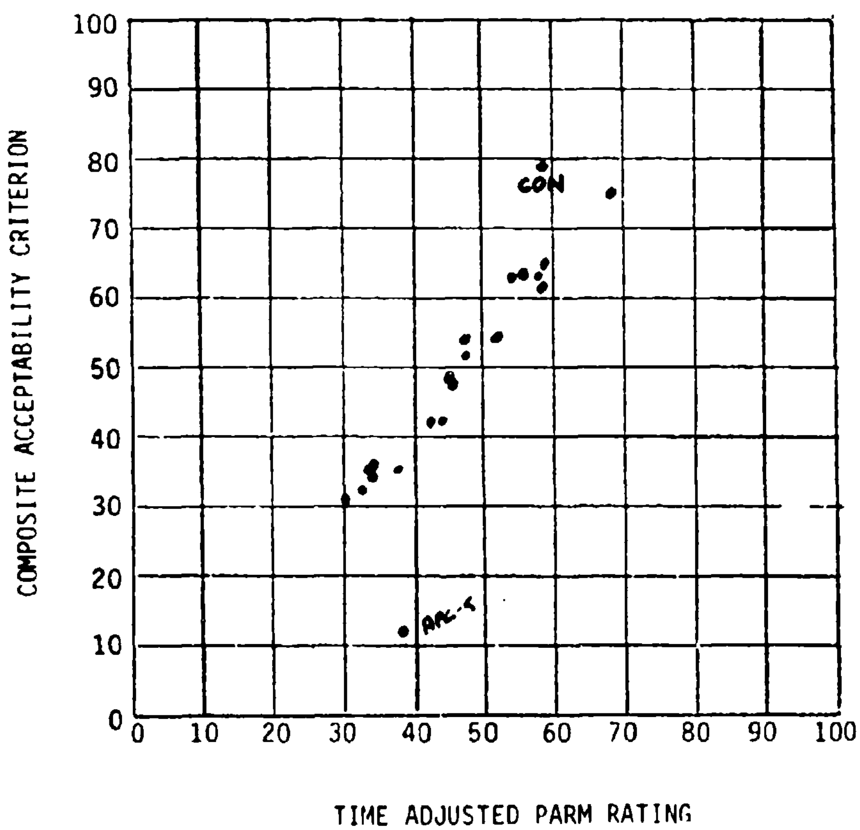

Fig. 4.4 Prediction of Composite Criterion Values
Time Adjusted PARM Ratings

4.5 Recommendations for Future Use of PARM

It is undoubtedly evident from the foregoing discussion that PARM, as originally conceived, is in need of some refinement before it can rival such speech evaluation isntruments as the Diagnostic Rhyme Test from the standpoints of robustness, reliability, and validity. Highly reliable results can be obtained from the DRT with minimum regard for the selection and management of the listening crew, but this is not yet the case with PARM. However, the means of achieving such refinement are rather clearly indicated by the results of research thus far performed, and a number of fairly specific recommendations can be made at this point.

4.5.1 Selection of Listening Crews - For all but the most preliminary evaluations, a listener crew of 10 or more carefully selected listeners is recommended. It is recommended that listeners be selected on the basis of performance on a probe PARM set, where the criteria for selection are self consistency and conformity with previously established norms for selected system-condition.

4.5.2 Selection of Speakers - It is recommended that a minimum of three male speakers, selected by means of a semantic differential voice rating form, (e.g., as used by Voiers, 1964) be used for routine system evaluation. Alternatively speaker selection may be based on data yielded by PARM, for a representative sample of system-conditions.

4.5.3 PARM Format - It is recommended that the inherent redundancy of PARM be substantially reduced and that other steps be taken to control intra-PARM listener variation. Specific steps to these ends should include:

1. Abandonment of the paired utterance feature.

2. Reduction in the number of presentations of all conditions.

3. Inclusion of one or more midrange probe conditions in all PARMs with post-experimental adjustment of each listener's data on the basis of his ratings of the probe conditions.

4. Increase in the number of system conditions included in each PARM set from four to six.

4.5.4 Statistical Control of Long-term Adaptation Level Drift - It is recommended that a standard PARM-set be periodically administered to PARM crews and that crew deviations from the normative response to the standard set be used as a basis for adjusting the data obtained from the crews during the particular epoch involved.

4.6 Overview

In the foregoing sections evidence with regard to the intrinsic validity and reliability of PARM has been presented. It is concluded that PARM can provide a highly reliable and valid measure of system acceptability to the population represented by the listening crew. Various recommendations have been made to increase its reliability, validity, and cost effectiveness. However, the effect of these recommendations is to dispense with a number of the features that distinguish PARM as an evaluation method.

Far from least among PARM's contributions to the technology of acceptability evaluation has been that of providing the means of determining which control features are important and which are trivial. Only through the use of such an instrument as PARM could one make this determination and confidently dispense with various of the controls which it originally incorporated. PARM as initially conceived, has thus served both as a valuable research tool and as an interim instrument for practical acceptability evaluation. Now perhaps, it should be abandoned in favor of modifications or new methods which take better advantage of the principles which it has served to elucidate.

The Quality Acceptance Rating Method (QUART), described in the next chapter represents one new method which was developed and refined largely on the basis of insights gained through experiments with PARM.

5.0 INVESTIGATION OF THE SEMANTIC DIFFERENTIAL APPROACH TO ACCEPTABILITY EVALUATION: DEVELOPMENT OF THE QUALITY ACCEPTABILITY RATING TEST.

5.1 The Semantic Differential Approach

The semantic differential approach was originally developed by Osgood (1952) to provide a comprehensive method of quantifying meaning. It has subsequently found application to a diversity of problems, the solutions to which require parsimonious, quantitative characterizations of complex cognitive processes. Most relevant in the present context is the usefulness of the method for characterizing the perceptual correlates of complex physical stimuli, for example, the perceptually distinctive characteristics of individual voices (Voiers, 1964) of passive sonar sounds (Solomon, 1958, 1959a, 1959b), and of complex visual forms (Elliott and Tannenbaum, 1963).

The classic semantic differential method involves a set of rating scales, each of which is defined by an antonymous pair of adjectives, for example, good:bad, black:white, and heavy:light. The respondent's task is to assign each concept, object or stimulus being investigated a value on each scale. Depending upon the problem being addressed, the basic procedure has been modified in various respects. For example, Voiers, (1965) has used pairs of word clusters rather than single-word pairs to define semantic continua, the choice of words comprising each cluster being based on results of preliminary investigations which, themselves, employed the semantic differential approach. Such clusters were designed to reduce the subject's uncertainty as to the nature of each perceptual continuum involved or as to the meanings of individual terms.

Although it is theoretically possible to determine the "semantic coordinates" of virtually any object or concept by using scales defined with such general terms as "good-bad," "large-small," "beautiful-ugly," and so on, the use of terms having more immediate relevance in a particular context (e.g., loud-soft, high-low, in the case of acoustical stimuli) can be expected to increase the precision and economy of the method. It is important, however, that technical jargon be avoided, except where it can be assumed that the subjects involved are fully acquainted with the meanings of the jargon expressions or terms. A major purpose of the semantic differential approach in a psychophysical context is, in fact, to develop a common language by means of which individuals can communicate their sensory-perceptual and effective experiences.

Regardless of the number of scales employed, subjects in semantic differential experiments most often respond in ways which indicate that a very limited number of orthogonal parameters (typically three) can account for the systematic component of their responses on the various scales. However, the use of a greater number of scales is desirable to insure a comprehensive inventory of the subject's perceptual reactions to the stimuli or cencepts involved. Normally, then, the semantic differential provides highly redundant characterizations of the subject's response. Factor analysis or a related technique is then employed to determine the number and nature of the underlying or implicit parameters of the subject's response to the stimuli or concepts involved.

A particularly useful property of the semantic differential approach is that it permits the simultaneous assesment of the affective or evaluative and the perceptual or non-evaluative aspects of a subject's response to the stimulus conditions involved. Thus, it can be used not only to identify the perceptual correlates of various types and degrees of speech signal degradation, but also to determine their interrelations with each other and with the evaluative aspect of the subject's response. It can be used, for example, not only to gauge the acceptability of processed speech but also to provide insights concerning the perceived characteristics which govern the listener's evaluative reaction to such speech.

5.2 Development of the Quality Acceptability Rating Test (QUART).

For the development and validation of QUART it was necessary, first, to obtain speech samples representing the diverse forms of speech processing and degradation likely to be encountered in communication situations of the present and foreseeable future. Speech materials representing various simple forms of degradation, plus materials that had been processed by various digital voice communication systems, were available for these purposes. These materials consisted of ninety six-syllable, phonemically-controlled sentences. Thirty of these were spoken by each of three male speakers. They were presented at an approximate rate of one sentence every four seconds.

In the first of a succession of pilot studies a semantic differential rating form involving 24 scales (see Figure 5.1) was used by several samples of listeners to describe their perceptions of the various types of speech processing and degradation and to indicate the degree of acceptability they would accord each type. Factor analysis of the results indicated the existence of four orthogonal parameters of the typical listener's response. It also provided some useful insights concerning the interrelations among various perceived system characteristics and system acceptability. Additionally, it revealed:

1. Several "silent" scales (i.e., scales for which listeners responses provided little or no basis for discrimination among the system-conditions involved.)

2. Several highly redundant scales.

3. Insufficient discrimination among some system conditions.

On the basis of these findings, a number of items were deleted or modified, and new items introduced.

Over the course of five additional pilot studies, the number of semantic rating scales was reduced to twelve, plus a 100-point acceptability rating scale. A rating form based on these scales is shown in Figure 5.2.

5.3 Experimental Validation of QUART

5.3.1 Materials, Method and Procedure

5.3.1.1 Experimental Conditions - To validate the QUART concept, generally, and System Rating Form III, in particular, speech samples representing 20 system-conditions and 6 forms of laboratory degradation were presented to 35 listeners, who used a version of System Rating Form III to indicate their perceptions and evaluations of these conditions. The conditions (and the abbreviations used in subsequent discussions) were as follows:

Laboratory Conditions

1. (H) Undegraded speech, lowpass filtered at 4 kHz.
2. (L) Speech processed sequentially by:
   a. A 2.4 kbps linear predictor with 1% bit error rate.

System ________
Rater ________
Date ________

SYSTEM RATING FORM 1A

LOUD INTENSE	( )	( )	( )	( )	( )	( )	( )	SOFT MILD
BABBLING GURGLING	( )	( )	( )	( )	( )	( )	( )	BUZZING DRONING
ANIMATED DYNAMIC	( )	( )	( )	( )	( )	( )	( )	LETHARGIC INERT
CONTINUOUS SUSTAINED	( )	( )	( )	( )	( )	( )	( )	INTERMITTENT INTERRUPTED
GROANING CREAKING	( )	( )	( )	( )	( )	( )	( )	SNAPPING CRACKLING
PASSIVE RESTING	( )	( )	( )	( )	( )	( )	( )	ACTIVE BUSY
CHIRPING TINKLING	( )	( )	( )	( )	( )	( )	( )	HOOTING BLEATING
NATURAL FAMILIAR	( )	( )	( )	( )	( )	( )	( )	UNNATURAL FOREIGN
LOW RUMBLING	( )	( )	( )	( )	( )	( )	( )	HIGH WHINING
ROUGH COARSE	( )	( )	( )	( )	( )	( )	( )	SMOOTH FINE
WARM COLORFUL	( )	( )	( )	( )	( )	( )	( )	COLD COLORLESS
JAGGED ABRUPT	( )	( )	( )	( )	( )	( )	( )	ROUNDED GRADUAL
PLEASANT PLEASING	( )	( )	( )	( )	( )	( )	( )	ANNOYING IRRITATING
ROARING THUNDERING	( )	( )	( )	( )	( )	( )	( )	HISSING RUSHING
THIN TWANGING	( )	( )	( )	( )	( )	( )	( )	THICK THUDDING
INTELLIGIBLE CLEAR	( )	( )	( )	( )	( )	( )	( )	UNINTELLIGIBLE HAZY
BOOMING THUMPING	( )	( )	( )	( )	( )	( )	( )	SCRAPING SCRATCHING
SHRILL PIERCING	( )	( )	( )	( )	( )	( )	( )	MELLOW MUFFLED
LARGE HEAVY	( )	( )	( )	( )	( )	( )	( )	SMALL LIGHT
HUMAN ALIVE	( )	( )	( )	( )	( )	( )	( )	MECHANICAL DEAD
SOLID SUBSTANTIAL	( )	( )	( )	( )	( )	( )	( )	HOLLOW FLIMSY
DANGEROUS THREATENING	( )	( )	( )	( )	( )	( )	( )	FRIENDLY REASSURING
BEAUTIFUL CLEAN	( )	( )	( )	( )	( )	( )	( )	UGLY DIRTY
STEADY STABLE	( )	( )	( )	( )	( )	( )	( )	FLUTTERING UNSTABLE

How would you rate this system on a 100 point scale of overall acceptability .... [ ]

Fig. 5.1 Preliminary QUART Rating Form

Dynastat

System ______________

Rater ______________

Date ______________

SYSTEM RATING FORM III-A

CONTINUOUS SUSTAINED	( )	( )	( )	( )	( )	( )	( )	INTERRUPTED INTERMITTENT
THUMPING THUDDING	( )	( )	( )	( )	( )	( )	( )	CLICKING TICKING
RATTLING PATTERING	( )	( )	( )	( )	( )	( )	( )	BUZZING DRONING
CRACKLING CLATTERING	( )	( )	( )	( )	( )	( )	( )	SQUISHING PLOPPING
NATURAL HUMAN	( )	( )	( )	( )	( )	( )	( )	UNNATURAL MECHANICAL
SIMMERING SEETHING	( )	( )	( )	( )	( )	( )	( )	CHIRPING CHEEPING
DIRTY CLUTTERED	( )	( )	( )	( )	( )	( )	( )	CLEAN UNCLUTTERED
SHARP PIERCING	( )	( )	( )	( )	( )	( )	( )	DULL MUFFLED
RUSHING GUSHING	( )	( )	( )	( )	( )	( )	( )	BABBLING GURGLING
GUTTURAL THICK	( )	( )	( )	( )	( )	( )	( )	NASAL THIN
UNINTELLIGIBLE GARBLED	( )	( )	( )	( )	( )	( )	( )	INTELLIGIBLE DISTINCT
FLUTTERING TWITTERING	( )	( )	( )	( )	( )	( )	( )	SCRATCHING SCRAPING

How would you rate this sytem on a 100 point scale of overall acceptability? .... (   )

(Assume that a typical telephone would receive a rating of 90)

Fig. 5.2 System Rating Form

b. An HY-2 channel vocoder.

c. A 9.6 kbps CVSD with 5% bit error rate.

d. A 4 kHz noisy channel which provided a processed speech/noise ratio of 22 dB in the passband.

3. (9 dB) Unprocessed speech with additive filtered white noise, providing a speech/noise ratio of 9 dB, measured in a 4 kHz passband.

4. (CLP) Peak clipped speech.

5. (Int.) Interrupted speech with an interruption rate to 150 ips and 50% duty cycle.

6. (2 kHz) Unprocessed speech lowpass filtered at 2 kHz.

System-Conditions

1. (4.8L-0) Linear predictor system at a 4.8 (2.7 kbps speech data) kbps transmission rate and 0% bit error rate (2.1 kbps used for error protection).

2. (3.6L-0) Linear predictor system at a 3.6 (2.7 kbps speech data) kbps transmission rate and 0% bit error rate (0.9 kbps used for error protection).

3. (2.4L-0) Linear predictor operating at 2.4 kbps.

4. (A-[illegible] A adaptive predictive coder operating at 8.0 kbps (four coefficients plus quantized error signal and pitch period indication).

5. (H-5) HY-2 channel vocoder (2.4 kbps).

6. (32C-0) Continuously variable slope delta modulation system (CVSD) operating at 32 kbps.

7. (16C-0) CVSD operating at 16 kbps.

8. (9.6C-0) CVSD operating at 9.6 kbps.

9. (P) Parkhill (20 dB S/N).

10. (A-C) Army vocoder in tandem with 16 kbps CVSD.

11. (C-A) CVSD in tandem with Army vocoder.

12.	(4.8L-5)	Linear predictor at 4.8 kbps (2.7 kbps) with 5% bit error rate (ber).
13.	(3.6L-5)	Linear predictor at 3.6 kbps (2.7 kbps) with 5% ber.
14.	(2.4L-5)	Linear predictor at 2.4 kbps with 5% ber.
15.	(A-5)	An APC with 5% ber.
16.	(H-5)	HY-2 vocoder with 5% ber.
17.	(32C-5)	CVSD at 32 kbps with 5% ber.
18.	(16C-5)	CVSD at 16 kpbs with 5% ber.
19.	(9.6C-5)	CVSD at 9.6 kbps with 5% ber.
20.	(CMV)	CONUS Median Voice grade link.

5.3.1.2 Listeners - The listening crew was composed of males and females between the ages of 18 and 29. All had survived a screening and training regimen which involved pure tone audiometry, the Diagnostic Rhyme Test, the Paired Acceptability Rating Method, and QUART, itself.

5.3.1.3 Speakers - Recordings by two male speakers, CH and LL, provided the speech materials for this investigation. CH is a relatively low-pitched speaker, LL a relatively high-pitched speaker.

5.3.2 Experimental Design and Procedure - Test materials spoken by the speakers were counterbalanced across listening crews. Approximately half the listeners heard the materials spoken by CH. Following a short break, they then heard the materials spoken by speaker LL. This order was reversed for the remaining listeners. In both cases, the laboratory processed speech materials were presented first and in the same order. Following the laboratory conditions samples representing the various system-conditions were presented in a randomly determined order in the case of one speaker and in the reverse order in the case of the other speaker.

A standard and an alternate version of the rating form was used. With both versions the subject's final task was to rate the system-condition involved on a 100-point scale of acceptability. The versions differed only in that the order and polarities of the rating scales were reversed in the case of the alternate form.

5.3.2.1 Instructions to subjects - A standard set of instructions (Appendix A) was read to each crew. Crew members were then encouraged to ask questions as needed to clarify their understanding of the task.

5.3.2.2 Familiarization with test materials - Prior to the rating session proper, the subjects were allowed to hear a sample sentence representing each of the 26 laboratory-and system-conditions. They were instructed not to rate these samples but simply to attend to them as a means of experiencing the range and diversity of speech qualities involved, and of establishing a reference frame in terms of which to make their ratings.

5.3.3 Analysis of Results - Since the interaction of speakers and systems was negligible, data for the two speakers were combined for purposes of the following analyses. No further analysis of data for individual speakers was undertaken for purposes of this investigation.

Each of the 12 semantic scales was assigned an arbitrary polarity. Numbers from "one" to "seven" were then assigned to the seven scale categories. Insofar as possible on an a priori basis, polarities were determined such that higher scale values were associated with favorable connotations, lower scale values with unfavorable connotations. An example is, "intelligible-distinct" which clearly has a more favorable connotation than "unintelligible-garbled." In some instances where both characteristics have unfavorable connotations (for example "chirping-cheeping" versus "simmering-seething") a neutral rating of "four" is the most favorable rating. To make fullest use of such bipolar scales, additional scoring procedures were introduced. Specifically, data for Scales 3, 4, 6, 9, and 12 were evaluated first in a normal manner and were then transformed to yield a second variable in each case. This second or derived variable was based on absolute deviations from the neutral rating of "four." Thus a total of 18 variables (including the acceptability rating) became available for purposes of characterizing listeners' reactions to the various laboratory and system conditions.

5.3.4 Results and Discussion - Table 5.1 presents the average rating received by each of the 26 conditions on each of the 13 primary variables and the 5 derived variables. Word pairs at the top and bottom of each column identify the upper and lower extremes of each continuum. System differences with respect to both primary and derived variables are evident, and various trends can be detected on close scrutiny.

Means, standard deviations, and F-ratios for conditions are presented for each variable in Table 5.2. Differences among the variables in terms of discriminating power are evident. Generally, those variables which involved evaluative reactions discriminate most effectively among the 26 conditions. However, all of the variables, both primary and derived, possess a high degree of discriminating power, as attested to by F-ratios which were significant at well beyond the .01 level in all instances.

TABLE 5.1 QUART Ratings of 26 System Conditions - Professional Listener Sample (N = 35)

		CONTN SUSTN	CLICK TICK	CLATR PATTR	CRAKL CLATR	NATPL HUMAN	CHIRP CHEEP	CLEAR UNCLU	SHARP PIERC	BABBL GURGL	NASAL THIN	INTLG DSTNC	FLUTR TWITR	BUZZ CLATR	SOISH CRAKL	SIMMR CHIRP	RUSH BABBL	SCRAT FLUTR	ACCPT
1	CVSD-32KB-0%	5.1	4.2	3.8	4.4	4.8	3.7	4.4	3.9	3.7	4.0	5.2	3.4	.2	.4	.3	.3	.6	71.1
2	CVSD-16KB-0%	4.4	4.2	3.5	4.3	3.6	3.6	3.4	3.9	3.6	4.4	4.2	2.9	.5	.3	.4	.4	1.1	60.7
3	CVSD-9.6KN-0%	3.7	4.4	3.3	4.6	3.2	3.5	2.8	3.7	3.6	4.4	3.5	2.6	.7	.6	.4	.4	1.4	50.6
4	CVSD-32KB-5%	4.4	4.6	3.2	5.3	3.2	2.9	2.3	3.6	2.8	4.2	3.8	2.0	.8	1.3	1.1	1.2	2.0	50.8
5	CVSD-16KB-5%	4.1	4.5	3.3	5.3	3.1	2.9	2.0	3.5	2.6	4.1	3.3	1.9	.7	1.3	1.1	1.4	2.1	46.1
6	CVSD-9.6KB-5%	3.9	4.6	3.4	5.5	2.6	2.6	1.7	3.3	2.8	4.0	2.7	1.7	.6	1.5	1.4	1.2	2.3	38.3
7	LPC-4.8KB-0%	4.1	3.8	4.0	4.1	4.2	4.1	3.8	3.4	4.3	3.6	4.5	4.0	.0	.1	.1	.3	.0	64.5
8	LPC-3.6KB-0%	4.2	4.0	3.8	4.1	3.7	4.2	3.5	3.2	4.4	3.6	4.3	4.0	.2	.1	.2	.4	.0	59.6
9	LPC-2.4KB-0%	4.1	3.7	3.9	3.9	3.6	3.9	3.5	2.9	4.2	3.5	4.0	3.9	.1	.1	.1	.2	.1	58.8
10	LPC-4.8KB-5%	2.2	4.5	5.2	3.9	1.9	6.0	1.6	3.5	6.4	3.6	2.2	6.3	1.2	.1	2.0	2.4	2.3	33.4
11	LPC-3.6KB-5%	2.3	4.3	4.9	3.9	2.1	5.5	1.9	3.4	6.2	3.5	2.4	5.5	.9	.1	1.5	2.2	1.5	33.9
12	LPC-2.4KB-5%	2.3	4.3	5.0	4.2	2.2	5.4	1.9	3.3	6.1	3.5	2.5	5.8	1.0	.2	1.4	2.1	1.8	34.3
13	HY2-2.4KB-0%	4.0	4.0	3.8	3.4	2.6	4.5	3.3	3.5	5.0	4.0	3.8	4.5	.2	.6	.5	1.0	.5	52.5
14	HY2-2.4KB-5%	3.1	4.2	4.9	3.8	1.9	5.2	1.9	3.3	6.2	3.4	2.6	5.7	.9	.2	1.2	2.2	1.7	35.1
15	APC--0%	4.4	3.7	3.9	4.1	3.7	3.8	3.7	3.4	3.8	3.8	4.7	3.8	.1	.1	.2	.2	.2	62.1
16	APC--5%	1.7	3.4	4.6	3.6	1.3	5.8	1.4	2.9	6.0	3.8	1.5	6.0	.6	.4	1.3	2.0	2.0	20.0
17	PARKHILL-20Db S/N	3.3	4.8	4.9	4.5	2.7	4.1	2.4	3.5	4.7	4.3	3.5	4.1	.9	.5	.1	.7	.1	47.0
18	ARMY VOC>CVSD 16K	3.6	4.8	3.6	5.7	2.5	3.5	1.9	3.0	3.6	3.5	2.8	2.0	.4	1.7	.5	.4	2.0	39.9
19	CVSD.16K>ARMY VOC	3.4	4.6	3.3	4.5	2.2	3.7	1.9	3.5	4.8	3.8	2.6	3.8	.7	.5	.3	.8	.2	37.2
20	CONUS.MED VOICE G	5.6	4.1	3.8	4.1	5.0	3.8	5.6	4.2	3.8	4.8	5.7	3.9	.2	.1	.2	.2	.1	76.6
21	HIGH ANCHOR	6.0	4.0	3.9	4.0	6.3	3.8	6.5	4.2	3.8	4.0	6.6	4.0	.1	.0	.2	.2	.0	88.8
22	LOW ANCHOR	2.3	4.6	4.0	4.3	1.7	4.2	1.4	2.2	5.0	2.9	1.5	2.9	.0	.3	.2	1.0	1.1	20.1
23	9DB S/N (4KHZ LP)	5.0	4.0	2.8	4.1	3.9	2.9	2.7	3.6	2.3	4.2	4.4	2.7	1.2	.1	1.1	1.7	1.3	51.6
24	CLIPPED SPEECH	3.9	4.2	2.8	4.5	2.5	3.0	2.2	4.3	2.6	4.2	3.3	2.2	1.2	.5	1.0	1.4	1.8	45.5
25	INTERRPTD SPEECH	4.3	3.9	3.3	4.2	3.9	3.9	3.6	3.6	4.7	3.7	4.5	4.4	.7	.2	.1	.7	.4	60.7
26	2 KHZ LP	5.7	3.9	3.8	4.0	5.6	3.7	5.5	3.5	3.8	3.9	5.9	4.0	.2	.0	.3	.2	.0	78.4
		INTRP INTRM	THUMP THUD	BUZZ DRONE	SQISH PLOP	UNATR MECHN	SIMMR SEETH	DIRTY CLUTR	DULL MUFLD	RUSH GUSH	GUTRL THICK	UNINT GARBL	SCRAT SCRAP	NUTRL	NUTRL	NUTRL	NUTRL	NUTRL	UNACP

TABLE 5.2 Means, Standard Deviations and F-ratios for QUART Scales

	SCALE			MEAN	S.D.	F-ratio for System-Condition*
1.	CONTN SUSTN	VS	INTRP INTRM	3.9	1.12	51.8
2.	CLICK TICK	VS	THUMP THUD	4.2	.36	7.3
3.	CLATR PATTR	VS	BUZZ DRONE	3.9	.67	16.7
4.	CRAKL CLATR	VS	SQUISH PLOP	4.3	.57	11.4
5.	NATRL HUMAN	VS	UNATR MECHN	3.2	1.23	107.5
6.	CHIRP CHEEP	VS	SIMMR SEETH	4.0	.91	41.1
7.	CLEAR UNCLU	VS	DIRTY CLUTR	3.0	1.37	133.3
8.	SHARP PIERC	VS	DULL MUFLD	3.5	.44	9.5
9.	BABBL GURGL	VS	RUSH GUSH	4.3	1.20	71.3
10.	NASAL THIN	VS	GUTRL THICK	3.9	.41	8.7
11.	INTLG DISTC	VS	UNINT GARBL	3.7	1.31	138.4
12.	FLUTR TWITR	VS	SCRAT SCRAP	3.8	1.33	87.4
13. (3D)**	BUZZ CLATR	VS	NUTRL	.5	.39	13.9
14. (4D)**	SQUISH CRAKL	VS	NUTRL	.4	.48	20.9
15. (6D)**	SIMMR CHIRP	VS	NUTRL	.7	.60	28.5
16. (9D)**	RUSH BABBL	VS	NUTRL	1.0	.74	43.3
17. (12D)**	SCRAT FLUTR	VS	NUTRL	1.0	.85	51.0
18.	ACCPT	VA	UNACP	50.7	17.29	230.8

*F = M.S. Conditions/M.S. Conditions x Listeners

With 25 and 850 degrees of freedom,

$P < .01$ for $F \geq 1.18$

**Derived variables

5.3.4.1 Dimensionality of Listener Response to System-Conditions - By design, the semantic differential approach provides a redundant characterization of the listener's perception of the individual system-condition. This is evident from Table 5.3, which shows the intercorrelations among the 18 primary and derived variables. Clearly, fewer than 18 dimensions are required to characterize listener response to a system-condition. The nature and number of the underlying dimensions of listener response thus become issues in need of resolution. Factor analysis was used for this purpose.

The correlation matrix in Table 5.3 was subjected to factor analysis by the principle components method. Five orthogonal factors were found to account for the systematic or reliable component of listener response to the 26 conditions. Following rotation of axes to a Varimax criterion of simple structure, further minor rotations were made in order to obtain the psychologically most meaningful set of factors. The matrix of factor loadings yielded by these procedures is shown in Table 5.4.

The pattern of factor loadings in Table 5.4 provides an adequate basis for identifying the five factors in psychological or subjective terms. However, some additional insights are to be gained from an examination of the configuration of the system-conditions in the data space defined by the five factors, i.e., a hyperspace whose primary axes are factors rather than explicit variables. Table 5.5 contains the coordinates of the 26 laboratory and system-conditions in the factorial data space, where the origin and scale have been transformed such that the means of all five distributions of factor scores fall at 50 and the standard deviations reflect the reliabilities of scores in each dimension. The effect of these transformations

TABLE 5.3 Intercorrelations of QUART Ratings of System-Conditions by the Professional Listener Sample

	CONTN SUSTN	CLICK TICK	CLATR PATTR	CRAKL CLATR	NATRL HUMAN	CHIRP CHEEP	CLEAR UNCLU	SHARP PIERC	BABBL GURGL	NASAL THIN	INTLG DSTNC	FLUTR TWITR	BUZZ CLATR	SQISH CRAKL	SIMMR CHIRP	RUSH BABBL	SCRAT FLUTR	ACCPT	
CONTN-SUSTN	1.00	-.22	-.58	.08	.93	-.65	.85	.60	-.69	.54	.94	-.45	-.40	-.10	-.57	-.70	-.53	.93	INTRP-INTRM
CLICK- TICK	-.22	1.00	-.02	.69	-.34	-.25	-.44	-.06	-.13	.00	-.36	-.40	.35	.59	.11	.13	.39	-.36	THUMP- THUD
CLATR-PATTR	-.58	-.02	1.00	-.45	-.35	.89	-.19	-.32	.86	-.42	-.35	.83	.10	-.36	.35	.45	.06	-.33	BUZZ-DRONE
CRAKL-CLATR	.08	.69	-.45	1.00	-.08	-.66	-.29	.01	-.59	.18	-.13	-.77	.14	.86	.03	-.15	.43	-.12	SQISH- PLOP
NATRL-HUMAN	.93	-.34	-.35	-.08	1.00	-.42	.95	.54	-.49	.40	.98	-.22	-.47	-.32	-.59	-.70	-.63	.97	UNATR-MECHN
CHIRP-CHEEP	-.65	-.25	.89	-.66	-.42	1.00	-.22	-.31	.96	-.48	-.41	.94	.14	-.48	.40	.54	.08	-.39	SIMMR-SEETH
CLEAR-UNCLU	.85	-.44	-.19	-.29	.95	-.22	1.00	.52	-.28	.37	.95	-.02	-.56	-.45	-.62	-.69	-.73	.96	DIRTY-CLUTR
SHARP-PIERC	.60	-.06	-.32	.01	.54	-.31	.52	1.00	-.40	.77	.62	-.15	.24	-.10	-.06	-.19	-.13	.61	DULL-MUFLD
BABBL-GURGL	-.69	-.13	.86	-.59	-.49	.96	-.28	-.40	1.00	-.57	-.48	.91	.10	-.42	.29	.49	.01	-.45	RUSH- GUSH
NASAL- THIN	.54	.00	-.42	.16	.40	-.48	.37	.77	-.57	1.00	.49	-.37	.14	.15	-.10	-.29	-.08	.47	GUTRL-THICK
INTLG-DSTNC	.94	-.36	-.35	-.13	.98	-.41	.95	.62	-.48	.49	1.00	-.19	-.44	-.32	-.60	-.70	-.66	.99	UNINT-GARBL
FLUTR-TWITR	-.45	-.40	.83	-.77	-.22	.94	-.02	-.15	.91	-.37	-.19	1.00	.13	-.63	.30	.45	-.12	-.18	SCRAT-SCRAP
BUZZ-CLATR	-.40	.35	.10	.14	-.47	.14	-.56	.24	.10	.14	-.44	.13	1.00	.14	.71	.75	.63	-.47	NUTRL-
SQISH-CRAKL	-.10	.59	-.36	.86	-.32	-.48	-.45	-.10	-.42	.15	-.32	-.63	.14	1.00	.22	.03	.58	-.31	NUTRL-
SIMMR-CHIRP	-.57	.11	.35	.03	-.59	.40	-.62	-.06	.29	-.10	-.60	.30	.71	.22	1.00	.90	.85	-.63	NUTRL-
RUSH-BABBL	-.70	.13	.45	-.15	-.70	.54	-.69	-.19	.49	-.29	-.70	.45	.75	.03	.90	1.00	.72	-.73	NUTRL-
SCRAT-FLUTR	-.53	.39	.06	.43	-.63	.08	-.73	-.13	.01	-.08	-.66	-.12	.63	.58	.85	.72	1.00	-.67	NUTRL-
-ACCPT	.93	-.36	-.33	-.12	.97	-.39	.96	.61	-.45	.47	.99	-.18	-.47	-.31	-.63	-.73	-.67	1.00	UNACP-
	INTRP INTRM	THUMP THUD	BUZZ DRONE	SQISH PLOP	UNATR MECHN	SIMMR SEETH	DIRTY CLUTR	DULL MUFLD	RUSH GUSH	GUTRL THICK	UNINT GARBL	SCRAT SCRAP	NUTRL	NUTRL	NUTRL	NUTRL	NUTRL	UNACP	

TABLE 5.4 Factorial Structure of QUART Ratings - Professional Listener Sample (N = 35)

	FACTOR LOADINGS				
FACTOR	1	2	3	4	5
1 CONTIN	.9018	-.3713	-.0713	.0104	.0508
2 CLICK	-.3568	-.0963	-.1290	.8587	.1777
3 CLATTR	-.2337	.9229	.1449	.0502	-.0829
4 CRAKLE	-.1317	-.5914	.1500	.7448	-.0257
5 NATURL	.9676	-.1165	-.0577	-.0391	-.0415
6 CHIRP	-.3178	.8798	.1447	-.2713	-.0473
7 CLEAN	.9706	.0761	-.1246	-.1244	-.0672
8 SHARP	.6070	-.1071	.0385	.0518	.7313
9 BABBLE	-.3940	.8737	-.0130	-.1755	-.1168
10 NASAL	.4634	-.3429	-.0524	.1067	.6489
11 INTELL	.9775	-.1083	-.1043	-.0679	.0513
12 FLUTTR	-.1205	.8935	.0837	-.3878	.0453
13 D3	-.4828	.0319	.2493	.0627	.7655
14 D4	-.3000	-.5421	.2916	.6317	-.0678
15 D6	-.5364	.1549	.7294	-.0556	.3603
16 D9	-.6770	.2912	.4042	-.1553	.3562
17 D12	-.6113	-.1802	.6684	.2207	.2338
18 ACPT	.9812	-.0806	-.1238	-.0402	.0197

TABLE 5.5 Factorial Coordinates of Laboratory Conditions and System Conditions

#	CONDITION	FACTOR 1	2	3	4	5
1	CVSD-32KB-0%	76.54	53.26	54.54	60.50	44.72
2	CVSD-16KB-0%	57.77	41.40	39.57	48.96	65.81
3	CVSD-9.6KN-0%	47.21	37.95	35.53	57.66	66.60
4	CVSD-32KB-5%	49.93	29.02	70.34	70.77	58.28
5	CVSD-16KB-5%	44.47	25.38	73.79	65.30	50.95
6	CVSD-9.6KB-5%	39.74	24.95	77.54	75.57	42.18
7	LPC-4.8KB-0%	63.17	52.88	43.95	39.54	27.72
8	LPC-3.6KB-0%	56.56	52.17	37.01	40.11	32.31
9	LPC-2.4KB-0%	53.54	46.59	38.02	30.60	20.00
10	LPC-4.8KB-5%	33.55	86.26	74.95	57.26	72.41
11	LPC-3.6KB-5%	33.29	76.18	60.44	48.14	58.93
12	LPC-2.4KB-5%	34.24	76.79	63.98	53.61	57.97
13	HY2-2.4KB-0%	50.31	53.92	42.40	34.88	47.09
14	HY2-2.4KB-5%	34.55	73.62	62.85	45.75	52.96
15	APC--0%	61.59	45.63	43.38	31.87	33.09
16	APC--5%	22.02	57.32	78.71	10.11	39.72
17	PARKHILL-20DB S/N	43.88	70.38	8.47	83.27	73.98
18	ARMY VOC>CVSD 16K	39.96	37.75	62.07	91.58	18.66
19	CVSD,16K>ARMY VOC	30.08	46.20	9.91	56.91	59.67
20	CONUS,MED VOICE G	85.06	56.69	40.87	52.76	69.75
21	HIGH ANCHOR	100.46	63.14	63.17	56.71	46.44
22	LOW ANCHOR	14.87	47.20	23.24	52.72	4.23
23	9DB S/N (4KHZ LP)	47.87	21.22	54.81	18.90	78.47
24	CLIPPED SPEECH	39.08	21.83	46.95	35.75	89.54
25	INTERRPTD SPEECH	55.45	47.61	35.63	35.83	53.46
26	2 KHZ LP	84.79	54.65	57.68	44.90	35.06
27	PERCEPTUAL ORIGIN*	60.41	54.35	24.18	44.74	49.94
	MEAN	50.00	50.00	50.00	50.00	50.00
	SIGMA	20.00	17.60	19.17	18.60	20.20

* Represented by hypothetical condition having an acceptability rating of 50 and neutral (4) ratings on all primary semantic scales.

is to preserve psychological distance relationships among system-conditions with some degree of accuracy. Also shown are the coordinates of a hypothetical subjectively neutral system-condition for both the professional listener sample and for the "target sample" (see Chapter 6). Projections of these coordinates on selected planes of the factorial data space are shown in Figures 5.3.1 - 5.3.4.

Factor I - Overall Acceptability - A factor loading of .98 in the case of the acceptability scale coupled with high loadings on other evaluative scales identifies Factor I as the affective or evaluative component of the listener's reactions to the 26 conditions. Table 5.5 and Figures 5.3.1 - 5.3.4 show the various system-conditions to be ordered in a manner which is consistent with this interpretation.

Further examination of the pattern of loadings on Factor 1 provides some insights concerning the antecedents or correlates of acceptability in the present instance. Particularly noteworthy is the high loading of Scale 1. Evidently, perceived temporal continuity of the speech signal was a major consideration in the relative acceptabilities of the 26 conditions involved here. Conditions that were perceived to preserve the temporal continuity of the speech signal were generally regarded with greater favor than those for which the signal was perceived as interrupted or intermittent. Also noteworthy is the high loading of the intelligibility scale on this factor, indicating that perceived intelligibility is a major condition of overall acceptability.

Listeners placed a high premium on naturalness, cleaness, sharpness, and nasality (as opposed to gutturality). High negative loadings on the derived variables D3, D4, D6, D9, and D12 suggest that they looked on all forms of degradation with some disfavor. Forced to choose, however, they favored conditions involving noise-like degradation over conditions involving various types of distortion. More specifically, negative loadings in the cases of Scales 2, 3, 4, 6, 9, and 12 indicate that listeners preferred:

System-conditions characterized as:		System-conditions characterized as:
Thumping Thudding		Clicking Ticking
Buzzing Droning		Rattling Pattering
Squishing Plopping	TO	Crackling Clattering
Simmering Seething		Chirping Cheeping
Rushing Gushing		Babbling Gurgling
Scratching Scraping		Fluttering Twittering

It must be stressed, however, that the relative preferences indicated, with respect to these qualities, are undoubtedly determined to a significant degree by the composition of the limited sample of conditions available for this investigation. <u>Extreme caution should be exercised in extrapolating or generalizing these results beyond the present sample of system-conditions</u>.

<u>Factor II - Babbling-Chirping</u> - This factor is defined by a number of scales, all of which would appear to describe a time-varying form of degradation as opposed to a temporally continuous, or noise-like form of degradation.

Support for this interpretation is provided by the configuration of data points in Figure 5.3.1. From the listener's

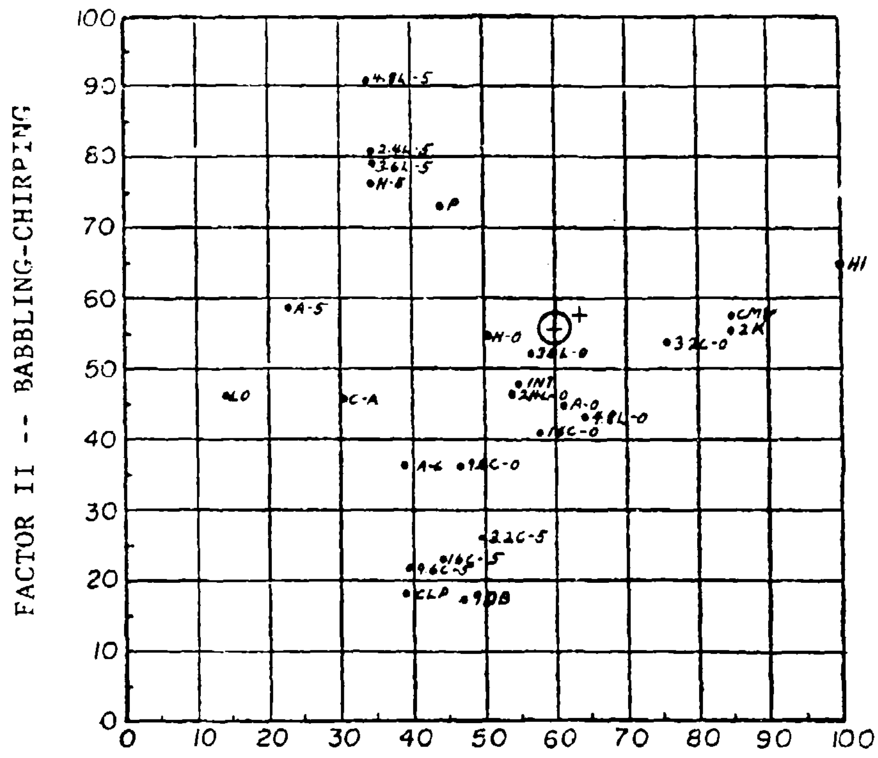

Fig. 5.3.1 Configuration of System-conditions in the I x II Plane of the Factor Space

⊕ Point of subjective neutrality for professional listener sample.

+ Point of subjective neutrality for target sample standpoint, it is this non-evaluative, perceptual quality that most conspicuously distinguishes the delta modulation systems from the narrowband analysis-synthesis systems.

Factor III - General Degradation - This factor is defined entirely by derived rating items. To the extent that a system-condition has a non-neutral status with respect to such perceptual continua as chirping-simmering and fluttering-scratching it is characterized by this factor. Figure 5.3.2 shows the configuration of system-conditions in this dimension of the factor space. Conditions involving digital transmission errors tend to rank highly on this dimension but other forms of degradation are also condusive to high rankings in this dimension.

Factor IV - Clicking-Clattering - This factor in combination with Factor III, effectively segregates system conditions in which bit errors occur (as shown in Figure 5.3.3), though the two factors are defined by different rating scales. The seemingly redundant functions of these two factors is probably due to the fact that bit errors provide the predominant form of degradation in the sample of system-conditions used in this investigation. The low standing of the 9 dB S/N on this factor suggests that it represents a noise versus distortion opposition. However, further research involving more diverse forms of degradation will be required to clarify this issue.

Factor V - Sharpness-Nasality - This factor is defined by two scales which were conceived in an attempt to capture the perceptual characteristics that distinguish vocoders from other narrowband systems. The attempt was not successful, but the factor evidently discriminates among systems on the basis of other characteristics, as shown in Figure 5.3.4.

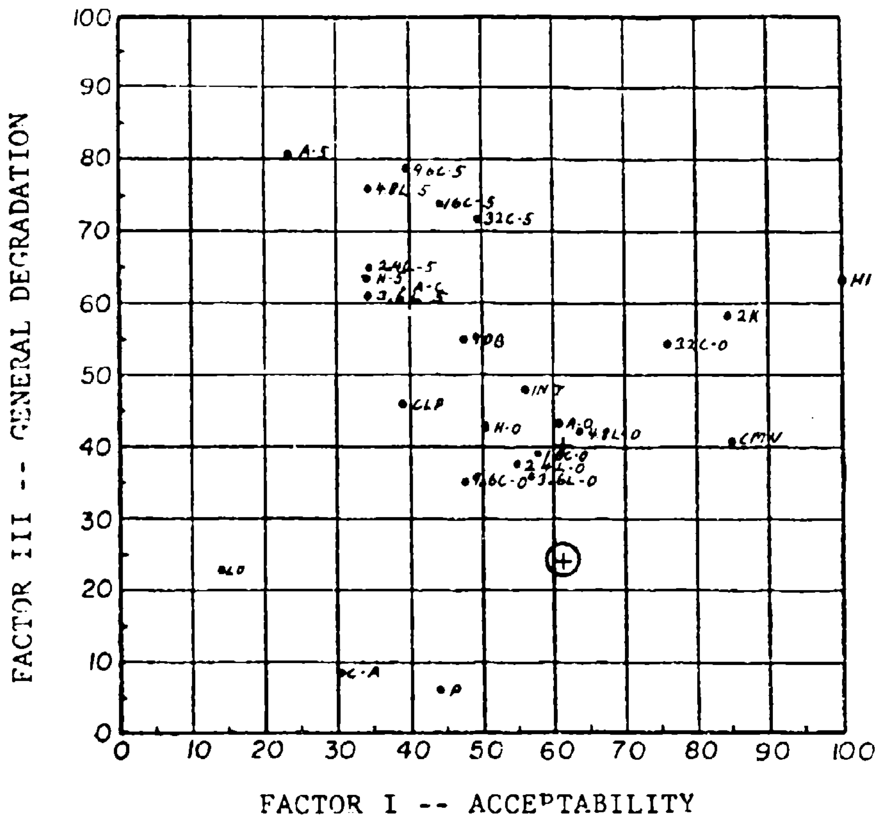

Fig. 5.3.2 Configuration of System-conditions in the I x III Plane of the Factorial Data Space

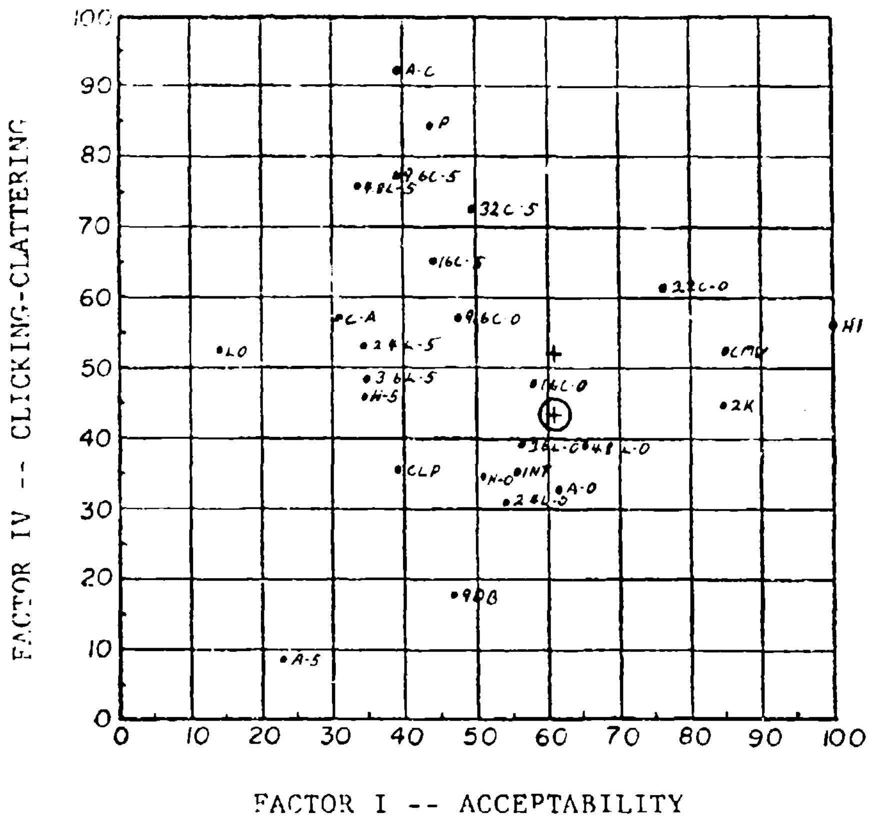

Fig. 5.3.3 Configuration of System-conditions in the I x IV Plane of the Factorial Data Space.

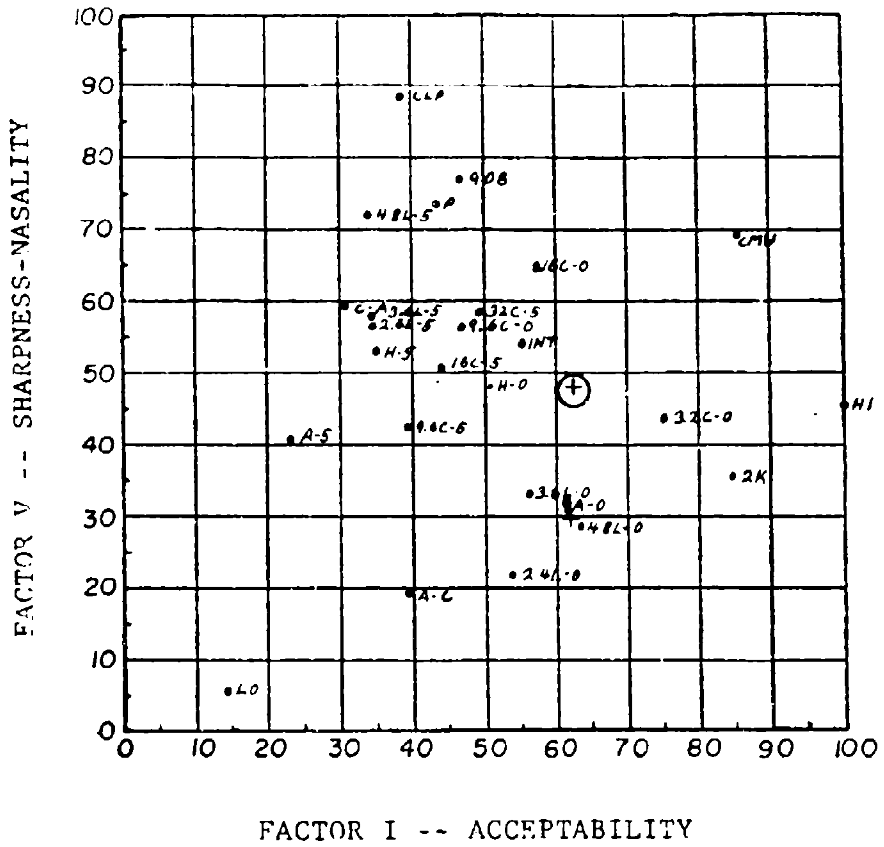

Fig. 5.3.4 Configuration of System-conditions in the I x V Plane of the Factorial Data Space.

It is evident that the precise nature and number of the perceptual parameters of degraded speech have yet to be conclusively defined. To do so will require further research involving a greater diversity of system-conditions than was available for this investigation. Examinations of the factor loading of the QUART scales and the configuration of factor scores for system-conditions strongly suggests that several potentially independent perceptual parameters tended to covary in this limited sample of system-conditions, but are potentially independently variable. More generally, the problem of identifying factorial dimensions is complicated by the relatively restricted sample of system-conditions used in this investigation: the bulk of this sample falls within a relatively circumscribed region of the perceptual space defined by the five factors. In Figure 5.3 it may be seen that the centroid of the configuration of systems in the factor space does not lie at the point of subjective neutrality i.e., the point representing a hypothetical system-condition that would receive an acceptability rating of 50 and neutral ratings on the twelve primary semantic rating scales.

In view of the foregoing considerations, judgment as to the exact nature and number of the elementary perceptual parameters of speech quality must be reserved at this time. But whatever the factorial structure of listeners' perceptions of system-conditions, the rating data yielded by QUART have some immediate practical value.

5.3.4.2 Predictive Validity of QUART - Individual rating scales, both evaluative and non-evaluative, have substantial potential for predicting system acceptance by the user population. Evidence of this is provided by Table 5.6 which shows the correlations between average semantic ratings of system-conditions and average acceptability ratings by the target sample. Also

TABLE 5.6 Correlations Between Semantic Differential Ratings and Target Sample Acceptability Ratings.

	Rating Scale	Correlation with Target Sample Acceptability Rating	
		Prof. List. Sample	Target Sample
1.	Cont-Sustained	.93	.98
2.	Click-Tick	-.36	-.25
3.	Clatter-Patter	-.33	-.35
4.	Crackle-Clatter	-.12	.11
5.	Natural-Human	.97	.97
6.	Chirping-Cheeping	-.39	-.70
7.	Clean-Uncluttered	.96	.95
8.	Sharp-Piercing	.61	.87
9.	Babbling-Gurgling	-.45	-.68
10.	Nasal-Thin	.47	.78
11.	Intelligible-Distinct	.99	.99
12.	Fluttering-Twittering	-.18	-.49
13.	Clattering-Buzzing	-.47	-.37
14.	Crackling-Squishing	-.31	-.78
15.	Chirping-Simmering	-.63	-.86
16.	Babbling-Rushing	-.73	-.75
17.	Fluttering-Scratching	-.67	-.35
18.	Acceptability	.98	----

shown for comparative purposes are correlations between average semantic ratings of system-conditions by the professional listener sample and acceptability ratings by the target sample. A correlation of .98 between acceptability ratings by the target sample and acceptability ratings by the professional listener sample implies that the two groups strongly agree on the <u>relative</u> merits of the various system-conditions. This implication is borne out by the pattern of correlations between acceptability ratings by the target sample and semantic ratings by both groups. The target sample's ratings of <u>continuity</u>, <u>naturalness</u>, <u>clarity</u>, and <u>intelligibility</u> are highly correlated with its ratings of acceptability. Corresponding semantic ratings by the professional listener sample are only slightly less correlated with the target sample's acceptability ratings. The latter results provide a strong indication of the feasibility of predicting user acceptance from QUART data yielded by laboratory listeners. Further indications are provided by a comparison of samples from these two populations in terms of how they perceive the differences among representative system-conditions. To this end, semantic differential rating data obtained from the target sample were subjected to factor analysis. As in the case of the professional listener sample, five interpretable factors were obtained.

The axes of the original factor space for the target sample were rotated to maximize their congruence with the axes on the factor space of the professional listening crew (Veldman, 1967). The resulting factor matrix is presented in Table 5.7. Also shown, for purposes of comparison, is the matrix yielded by the professional listening crew. Virtually perfect congruence of the corresponding axes was achieved. Shown in Table 5.8 are cosines between individual scale vectors (i.e., coefficients of correlation between ratings by professional and target samples).

TABLE 5.7 Factorial Structures of QUART Ratings - Professional Listener Sample and Target Sample

FACTOR LOADINGS

Professional Listener Sample (35)

Scale	Factor 1	2	3	4	5
1 CONTIN	.9018	-.3713	-.0713	.0104	.0508
2 CLICK	-.3568	-.0963	-.1290	.8587	.1777
3 CLATTR	-.2337	.9229	.1449	.0502	-.0829
4 CHAKLE	-.1317	-.5914	.1500	.7448	-.0257
5 NATURL	.9676	-.1165	-.0577	-.0391	-.0415
6 CHIRP	-.3178	.8798	.1447	-.2713	-.0473
7 CLEAN	.9706	.0761	-.1246	-.1244	-.0672
8 SHARP	.6070	-.1071	.0385	.0518	.7313
9 BAUBLE	-.3940	.8737	-.0130	-.1755	-.1168
10 NASAL	.4634	-.3429	-.0524	.1067	.6489
11 INTELL	.9775	-.1083	-.1043	-.0679	.0513
12 FLUTTR	-.1205	.8935	.0837	-.3878	.0453
13 D3	-.4828	.0319	.2493	.0627	.7655
14 D4	-.3000	-.5421	.2916	.6317	-.0678
15 D6	-.5364	.1549	.7294	-.0556	.3603
16 D9	-.6770	.2912	.4842	-.1553	.3562
17 D12	-.6113	[illegible]	.6604	.2207	.2338
18 ACPT	.9812	-.0806	-.1238	-.0402	.0197

Target Sample (90)

Scale	Factor 1	2	3	4	5
1 CONTIN	.9779	-.0198	.0967	-.0746	-.0024
2 CLICK	-.1243	-.3516	-.0336	.8752	.2465
3 CLATTR	-.4658	.8120	.0421	.2966	.0688
4 CHAKLE	.2586	[illegible]	-.0246	.7272	.0188
5 NATURL	.9571	[illegible]	.1369	.0261	-.0439
6 CHIRP	-.8026	.5527	-.0607	-.0963	.0684
7 CLEAN	.8990	.3057	.0813	-.1778	-.1557
8 SHARP	.8699	.1341	.2916	.0789	.2753
9 BAUBLE	-.7867	.5229	-.1989	-.1329	-.0945
10 NASAL	.8210	-.0377	.3577	.1234	.3053
11 INTELL	.9762	.0768	.1357	-.0562	-.0163
12 FLUTTR	-.6125	.6339	-.1470	-.4052	-.0009
13 D3	-.2853	-.0843	-.3444	.1055	.8654
14 D4	-.3270	-.5890	.4091	.5352	-.1157
15 D6	-.8371	.2814	.1355	-.0330	.3760
16 D9	-.8959	.1400	.0343	-.1101	.3522
17 D12	-.7490	-.2568	.4012	.2343	.3441
18 ACPT	.9719	.1459	.0943	-.0669	-.0423

TABLE 5.8 Correlations between System Ratings by the Professional Listener Sample and by the Target Sample

(+) SCALE	COEFFICIENTS OF CORRELATION	(-) SCALE
1 CONTN/SUSTN	.98	INTRP/INTRM
2 CLICK/TICK	.98	THUMP/THUD
3 RATTL/PATTR	.96	BUZZ/DRONE
4 CRAKL/CLATR	.94	SQISH/PLOP
5 NATRL/HUMAN	.97	UNATR/MECHN
6 CHIRP/CHEEP	.96	SIMMR/SEETH
7 CLEAR/UNCLU	.99	DIRTY/CLUTR
8 SHARP/PIERC	.90	DULL/MUFLD
9 BABBL/GURGL	.97	RUSH/GUSH
10 NASAL/THIN	.89	GUTRL/THICK
11 INTLG/DSTNC	.98	UNINT/GARBL
12 FLUTR/TWITR	.96	SCRAT/SCRAP
13 BUZZ/RATTL	.93	NEUTRAL
14 SQISH/CRAKL	.97	NEUTRAL
15 SIMMR/CHIRP	.90	NEUTRAL
16 RUSH/BABBL	.97	NEUTRAL
17 SCRAT/FLUTR	.99	NEUTRAL
18 ACCEPTABLE	.98	UNACCEPTABLE

From the foregoing results it is clear that the two samples discriminated systems with respect to essentially the same perceptual parameters, although there are minor indications that they value some perceptual qualities somewhat differently. For example, the loadings of Scale 18 (acceptability) on Factors II and III, though small, are somewhat higher for the target sample than for the professional listener sample. The practical and theoretical implications of these differences would appear to be rather trivial, particularly when it is recalled that the professional listener sample had undoubtedly had more extensive exposure to modern digital voice communication systems than the typical member of the target sample. Given a more broadly experienced target sample, or a less experienced professional sample, less pronounced differences might be expected. Further examination revealed that the two samples also differed in terms of their subjective neutral points, or adaptation levels, for the various perceptual qualities, as is shown in Table 5.5 and Figures 5.3.1 - 5.3.4. In general, the target sample tended to be more lenient than the professional listener sample in its ratings of the various conditions. The most likely explanation of this discrepancy is that the target sample had a different conception than the professional sample of what is implied by "routine communications." Undoubtedly there were also individual differences in this respect within both the professional and target samples. Pre-exposure of listeners to a standard, simulated communications situation might, thus, serve to significantly improve the reliability of QUART results.

5.3.4.3 Practical Uses of QUART for the Prediction of User Acceptance of Communication Systems - The results described above support the hypothesis that professional listeners and potential system users base their evaluative reactions to communication systems on essentially the same perceptual qualities and place similar values on each of these qualities. In any case, there is a high correlation between professional listeners' perceptions and users' affective or evaluative reactions to processed speech. Several approaches to the practical prediction of user acceptance thus merit consideration.

Extremely good prediction of user acceptance reactions can be obtained using only the acceptability ratings of a professional listener sample. The correlation between these variables is shown, graphically, in Figure 5.4. However, the high correlations between the perceptual reactions (via semantic ratings) of professional listeners and acceptability ratings by the target sample, suggest that even better prediction of user acceptance reactions can ultimately be obtained by the use of multiple prediction techniques.

Unfortunately, the sample of system-conditions (20), for which ratings by both the professional listener and target samples are available, is far too small to permit a valid test of the feasibility of multiple prediction procedures (or in any case, to yield a generally applicable set of regression coefficients). Rating data from a sample of system users for a large, representative sample of speech processing and communication systems would be very desirable, but in the absence of such data, a further step toward the validation of the multiple prediction approach is possible. This step requires the assumption that the professional listener population and population of system users do in fact value the various relevant perceptual qualities of processed speech in essentially the same way, which assumption finds support from results described above. The results of a study conducted after the formal termination of this project are then of interest.

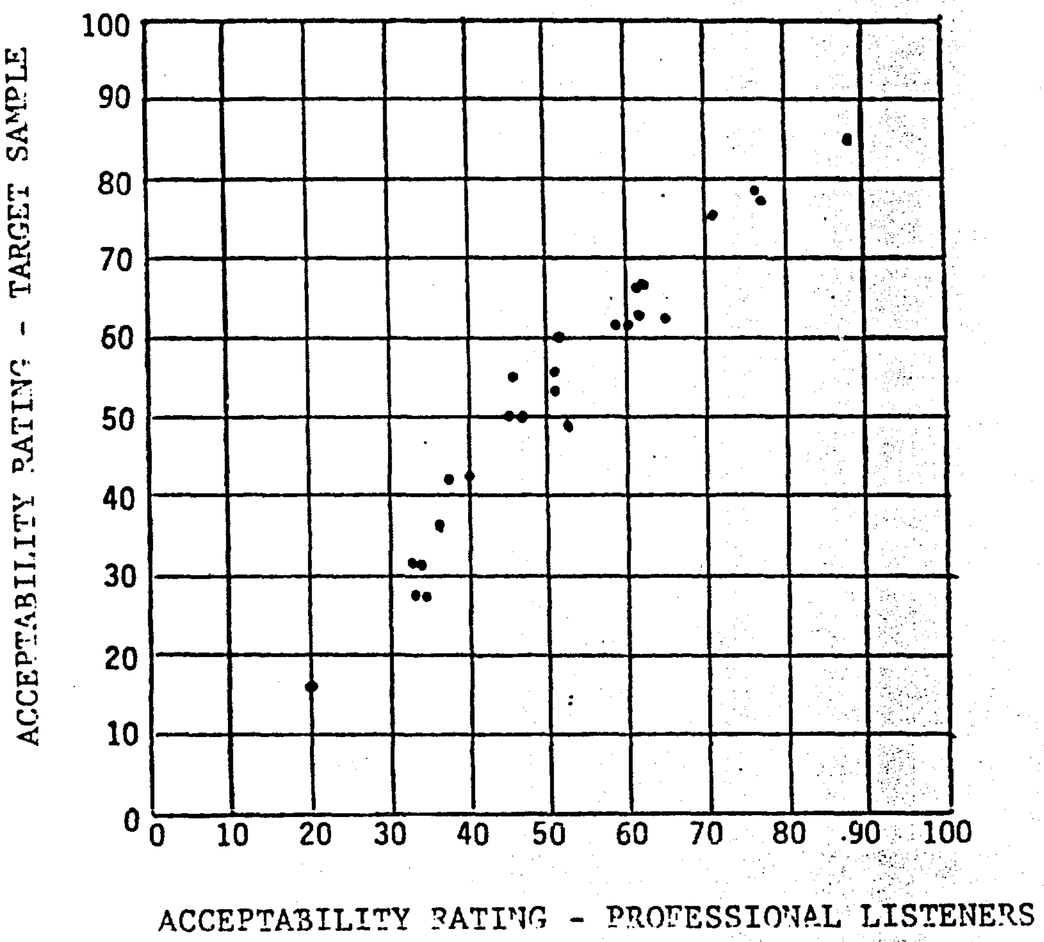

Fig. 5.4 Correlation between acceptability ratings of the target sample and professional listener sample

These results were yielded by QUARTs conducted on a large sample of system-conditions using Dynastat's professional listener sample, only. A total of 182 conditions, including 3 bit error rates for each of 37 system-conditions and six probes (each of which was rated nine times) were rated by 17 professional listeners, using System Rating Form III (Figure 5.2).

The multiple correlation between the average acceptability rating of a condition and its ratings on the twelve semantic scales was .99. The correlations between individual semantic scale and rated acceptability are shown in Table 5.9 which also shows the normalized regression coefficients (betas) for each semantic scale. These results demonstrate the feasibility of predicting acceptability from non-evaluative rating data or of supplementary results of acceptability ratings with semantic rating data. They have a number of potentially significant implications for the methodology of speech acceptability evaluation.

Although present evidence does not support the hypothesis of qualitative differences between the value system of professional listeners and system users--the two samples discriminated systems with respect to the same perceptual qualities and valued these qualities similarly--the possibility remains that other populations of system users will be found to apply a different system of values in evaluating communication systems. (None of the members of the present target sample held positions at the command and staff level.) Given individuals with different communications needs and purposes, one may expect to find different criteria of acceptability employed. Isometric methods of acceptability evaluation will fail in such circumstances, but parametric methods, as exemplified above, can be adapted to them. There is some basis, moreover, for predicting that the parametric approach will prove less susceptible to the effects of attitudinal and mood changes in the professional listener. It is not difficult

TABLE 5.9 Correlations between Semantic Ratings and Acceptability Ratings of 182 System-Conditions by the Professional Listener Sample

(+) SCALE	COEFFICIENTS OF CORRELATION	NORMALIZED REGRESSION COEFFICIENTS	(-) SCALE
CONTINUOUS SUSTAINED	.95	.18	INTERRUPTED INTERMITTENT
CLICKING TICKING	-.47	-.03	THUMPING THUDDING
RATTLING PATTERING	-.37	.03	BUZZING DRONING
CRACKLING CLATTERING	.14	.00	SQUISHING PLOPPING
NATURAL HUMAN	.95	.22	UNNATURAL MECHANICAL
CHIRPING CHEEPING	-.29	-.06	SIMMERING SEETHING
CLEAN UNCLUTTERED	.96	.12	DIRTY CLUTTERED
SHAPR PIERCING	.40	.01	DULL MUFFLED
BABBLING GURGLING	-.46	.03	RUSHING GUSHING
NASAL THIN	.31	.00	GUTTURAL THICK
INTELLIGIBLE DISTINCT	.99	.48	UNINTELLIGIBLE GARBLED
FLUTTERING TWITTERING	-.22	.04	SCRATCHING SCRAPING

to imagine that a listener will tend to rate systems less favorably when depressed, more favorably when elated; but is more difficult to conceive of how his mood would affect his judgments of "continuous vs. interrupted," "natural vs. unnatural" or "rushing vs. babbling."

In summary, the validity of QUART whether employed isometrically, parametrically or with a combination of the two approaches, is attested to by a variety of evidence. What remains to be accomplished is the implementation of standard procedures for its use.

In the above connection it would be highly desirable to have normative data for a more diverse sample of the types of degradation imposed on the speech signal by modern speech processing and communication systems. Although a large number of conditions have been treated in the course of QUART research to date, they nevertheless represent a relatively circumscribed class. The majority of these were narrow band digital voice systems involving a limited number of speech processing and coding algorithms. Poorly represented in this sample were the various forms of noise and distortion typical of analog communication systems operating in various environments. Before QUART is standardized--particularly with respect to the regression coefficients used for parametric evaluation, and even with respect to the semantic rating scales comprising the QUART rating form--QUART data for such conditions must become available. In this connection it should be emphasized again that the set of semantic ratings scales used in Systems Rating Form III was optimized for discrimination within the particular sample system-conditions available at the time. A different set will undoubtedly be required to render QUART more generally applicable. However, the manner in which this issue is resolved is unlikely to affect the validity and reliability of QUART acceptability ratings, so long as the listener is required to attend closely to a variety of perceptually relevant system characteristics before making an acceptability rating.

Secondly, it would be very desirable to obtain normative QUART data from other segments of the population of military communication system users, for example, from users in command and staff position. In the meantime, QUART, used only in the isometric mode with properly selected probes and anchors, can provide a highly reliable, valid and cost effective means of practical system evaluation from the standpoint of overall acceptability.

6.0 FURTHER VALIDATION OF PARM AND QUART

A factor which complicated the task of validating PARM and QUART within the term, proper, of this project was the unavailability of a sufficient amount of correlated PARM and QUART data. Part of the problem was that acceptability ratings by the target sample could be obtained only for a small and questionably representative sub-sample of the total sample of system-conditions ultimately evaluated with PARM. QUART data for the remaining system-conditions were not available for either the target sample or professional listener sample. Fortunately, however, taped materials in QUART format for a sample of 101 system-conditions were made available to Dynastat after the formal completion of work on the project.

Dynastat undertook the performance of QUART evaluations of these 101 conditions on its own volition. This made available a set of correlated QUART and PARM data subject to identification by DCA of the systems for which PARM evaluations had been conducted under Contract No. DCA100-75-C-0034. Completion of these QUART evaluations, under Dynastat's auspices made it possible to test more fully the cross predictability of PARM and QUART rating. For this set of system-conditions the coefficient of correlation was found to be .94. Figure 6.1 shows this correlation in graphic form. The correlation appears to be somewhat lower than that previously obtained for a sample of system-conditions with no bit errors and with 5% bit errors. In this connection it should be recalled that all PARM data were corrected for long term adaptation level drift on the basis of an empirically derived algorithm. There is little question but what this algorithm was less than totally efficacious. But for this complication a higher correlation would undoubtedly have been obtained.

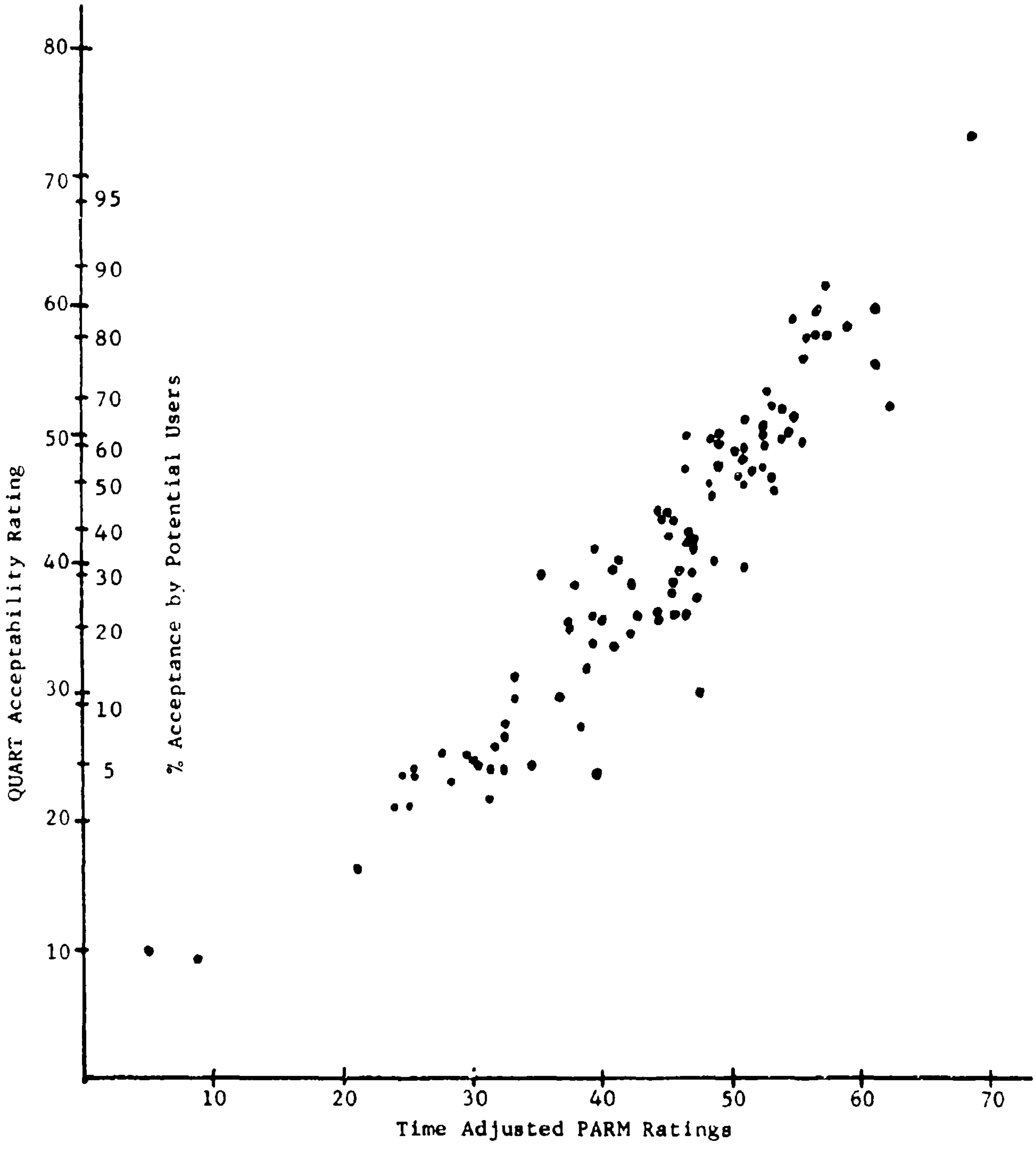

Figure 6.1. Correlation Between QUART and PARM Ratings

It is clear, in any event, that PARM and QUART measure essentially the same aspects of listener reaction to processed speech. With adequate control of listener factor, both can provide highly reliable and valid indicants of the acceptability of voice communications equipment.

APPENDIX

I. PRODUCTION OF MASTER TAPES

In accordance with contract specifications, Dynastat prepared master tape recordings of both DRT and acceptability test materials

Description of Speech Materials

The Diagnostic Rhyme Test (DRT) is a two-choice test of consonant discriminability or, more accurately, a test of the apprehensibility of the speaker's intent with respect to the states of six elementary attributes of consonant phonemes (Voiers, et al, 1973). It yields a gross indicant of speech intelligibility and additional scores relating to specific aspects of the performance of the speaker, listener or system under test and it utilizes a corpus of 192 words (96 rhyming pairs). In a given instance, the listener's task is to indicate which member of the pair has actually been spoken. A correct choice indicates that the listener has, in effect, apprehended the speaker's intent as to the state of one of six essentially binary perceptual attributes of English consonant phonemes. An incorrect choice indicates that the speaker, listener or system under test has failed to distinguish the source state of the attribute. Depending on the word pair involved, each item tests for the apprehensibility of one of the following elementary phonemic attributes:

Voicing
Nasality
Sustention
Sibilation
Graveness
Compactness

The DRT contains sixteen items, or word pairs, to test the apprehensibility of each attribute, and the two states of each attribute are given equal representation in the test. Table 1 shows the corpus of stimulus words used in the present version (Form IV) of the Diagnostic Rhyme Test.

The speech materials for acceptability test recordings consisted of 900 six-syllable sentences, 600 declarative sentences and 300 interrogative. Sentences were constructed to meet the following criteria: at least one of the six-syllables contained a vowel from each of the categories shown in Table 2 and each sentence contained at least one consonant from each of the categories shown in Table 3.

Recording Master Tapes

The speaker was seated in a Tracoustics single wall sound room 10' x 10' 8". Scotch 206 half-inch, magnetic recording tape was used with an Ampex 440B 4-track tape recorder, which was located outside of the sound room.

Tapes were recorded at a speed of 15 ips. with peak recording levels not exceeding a 0.5% harmonic distortion threshold and an overall signal-to-noise ratio of at least 55 dB. National Association of Broadcasters equalization standards were observed for recording and playback.

Quiet Environment Recordings

In the quiet environment two full list (384 words) DRTs and a set of 90 acceptability sentences were recorded for each speaker shown in Table 4. The microphones used and their respective channels were as follows:

TABLE 1. CORPUS OF STIMULUS ITEMS USED IN THE DRT (Form IV)

VOICING	NASALITY	SUSTENTION
DAUNT-TAUNT	MOOT-BOOT	SHEET-CHEAT
ZED-SAID	GNAW-DAW	SHOES-CHOOSE
DINT-TINT	NECK-DECK	THONG-TONG
VOLE-FOAL	NIP-DIP	FENCE-PENCE
BOND-POND	MOAN-BONE	VILL-BILL
VAST-FAST	KNOCK-DOCK	THOSE-DOZE
BEAN-PEEN	MAD-BAD	VOX-BOX
ZOO-SUE	NEED-DEED	THAN-DAN
VAULT-FAULT	NEWS-DUES	VEE-BEE
DENSE-TENSE	MOSS-BOSS	FOO-POOH
GIN-CHIN	MEND-BEND	SHAW-CHAW
GOAT-COAT	MITT-BIT	THEN-DEN
JOCK-CHOCK	NOTE-DOTE	THICK-TICK
GAFF-CALF	MOM-BOMB	THOUGH-DOUGH
VEAL-FEEL	NAB-DAB	VON-BON
DUNE-TUNE	MEAT-BEAT	SHAD-CHAD

SIBILATION	GRAVENESS	COMPACTNESS
JAB-GAB	POT-TOT	GHOST-BOAST
CHEEP-KEEP	BANK-DANK	GOT-DOT
CHEW-COO	WEED REED	SHAG-SAG
SAW-THAW	POOL-TOOL	YIELD-WIELD
JEST-GUEST	FOUGHT-THOUGHT	COOP-POOP
SING-THING	MET-NET	CAUGHT-TAUGHT
JOE-GO	BID-DID	YEN-WREN
CHOP-COP	FORE-THOR	HIT-FIT
SANK-THANK	WAD-ROD	SHOW-SO
ZEE-THEE	FAD-THAD	HOP-FOP
JUICE-GOOSE	PEAK-TEAK	GAT-BAT
JAWS-GAUZE	MOON-NOON	KEY-TEA
CHAIR-CARE	BONG-DONG	YOU-RUE
JILT-GILT	PENT-TENT	YAWL-WALL
SOLE-THOLE	FIN-THIN	KEG-PEG
JOT-GOT	BOWL-DOLE	GILL-DILL

TABLE 2. VOWEL CATEGORIES

	Front	Mid	Back
High	team - i tip - I		tool - u took - ʊ tone - o
Mid		ton - ʌ bird - ɜ	
Low	ten - ɛ tap - æ		talk - ɔ top - a

TABLE 3. CONSONANT CATEGORIES

Sibilants	Stops	Fricatives
zip - z	pat - p	vat - v
sit - s	top - t	for - f
chat - ʃ̂	bat - b	thin - θ
shot - ʃ	dot - d	that - ð
jot - ʒ̂	get - g	
	kit - k	

TABLE 4. FUNDAMENTAL FREQUENCY OF SPEAKERS

Low Pitch	CH - 102 Hz	BV - 103 Hz	MP - 200 Hz
Average Pitch	RH - 115 Hz	JE - 118 Hz	JS - 236 Hz
High Pitch	PK - 126 Hz	LL - 133 Hz	LS - 260 Hz

Channel	Microphones
1	Altec Dynamic, Model # 659A, Serial # 1431
2	Western Electric, Model # T1
3	Grason Stadler Throat, Model # E7300M
4	General Radio Ceramic Studio, Model # 1560-P5, Serial # 2180

The Altec microphone was placed approximately two inches to the right of the speaker's lips; the Western Electric microphone to the left of the lips at the same distance. The throat microphone was taped to the speaker just below the frontal projection of the larynx; and the General Radio microphone was suspended 20 cm. from the front of the speaker's lips, in grazing position. Figures 1 and 2 show the microphone placements from two views.

Noise Environment Recordings

Three male speakers (CH, JE, and RH) recorded one full list DRT and 90 acceptability test sentences in each of the following noise conditions:

1. Air Borne Command Post (ABCP) - 85 dB*
2. Helicopter - 115 dB
3. Shipboard - 82 dB
4. Office - 63 dB

One female speaker (JS) recorded one full list DRT and 90 sentences in the office noise condition only. A General Radio Sound Level Meter, Model 1551C, was used for measuring the noise level in each condition (C-weighted). Figure 3 shows block diagrams of the equipment and the sound room used in the recording of the noise environment conditions.

*SPL (C-weighted)

Fig. 1 Microphone Placement in Quiet Environment - View 1.

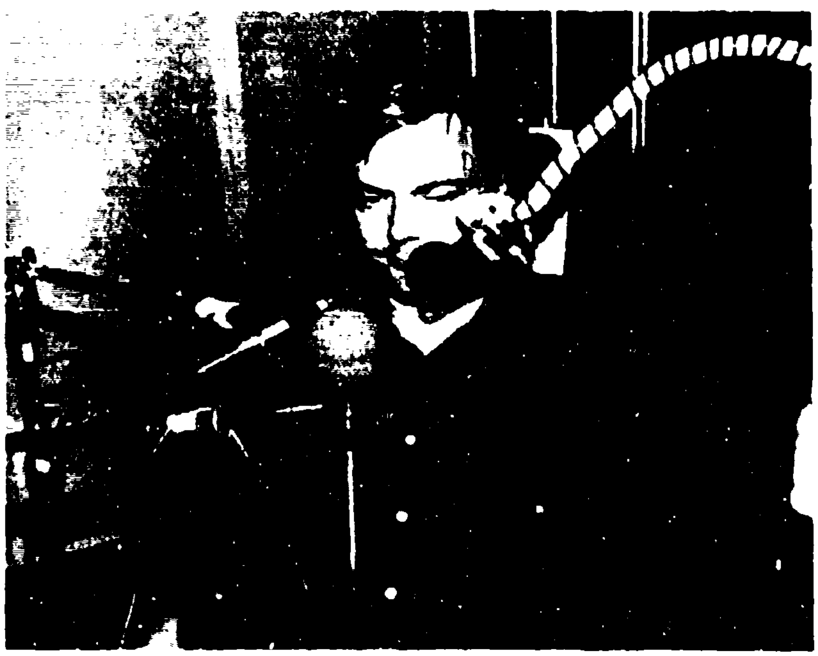

Fig. 2 Microphone Placement in Quiet Environment - View 2.

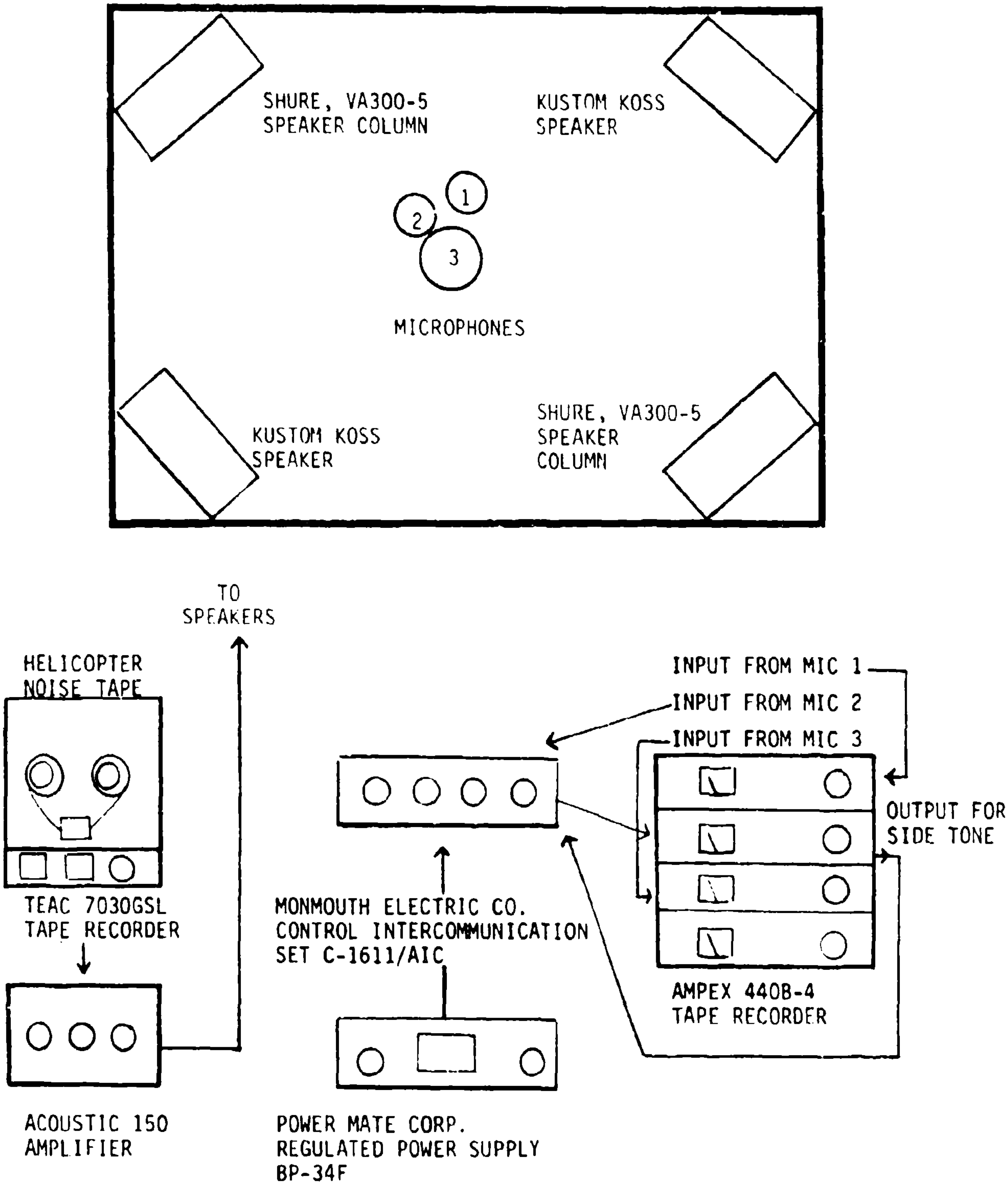

Figure 3. DIAGRAM OF AUDIO EQUIPMENT AND ROOM USED IN RECORDING OF SPEECH MATERIAL IN VARIOUS NOISE ENVIRONMENTS.

In the ABCP, shipboard, and office noise environments the following microphones were used:

Channel	Microphones
1	Altec Dynamic, Model # 659A, Serial # 1431
2	Roanwell Noise Cancelling
3	Grason Stadler Throat, Model # E7300M

The microphone placements, shown in Figures 4 and 5, were the same as in the quiet environment with the exception that the Roanwell microphone was within one-half inch of the lips. Rudmose headphones, RA-125 with TDH-39 elements were used for ear protection, as well as for carrying a feedback signal to the speaker.

For the helicopter noise environment an Electrovoice M-78/AIC Dynamic microphone replaced the Roanwell. The helicopter microphone, the Gentrex helicopter helmet Model SPH-4, was used to protect the speaker's ears and provide a feedback signal in the 115 dB environment. Microphone placement for the helicopter noise condition is shown in Figure 6.

Editing and Quality Control

After recording the full list DRTs and acceptability test materials, tapes were edited and assembled for evaluation by the listening crew. Full test DRTs were presented to the crew, scored, and the results carefully analyzed. Tapes were re-edited and evaluated again by the listening crew. Three-speaker test modules were then assembled into their final format.

Fig. 4 Microphone Placement in Noise Condition (Front View)

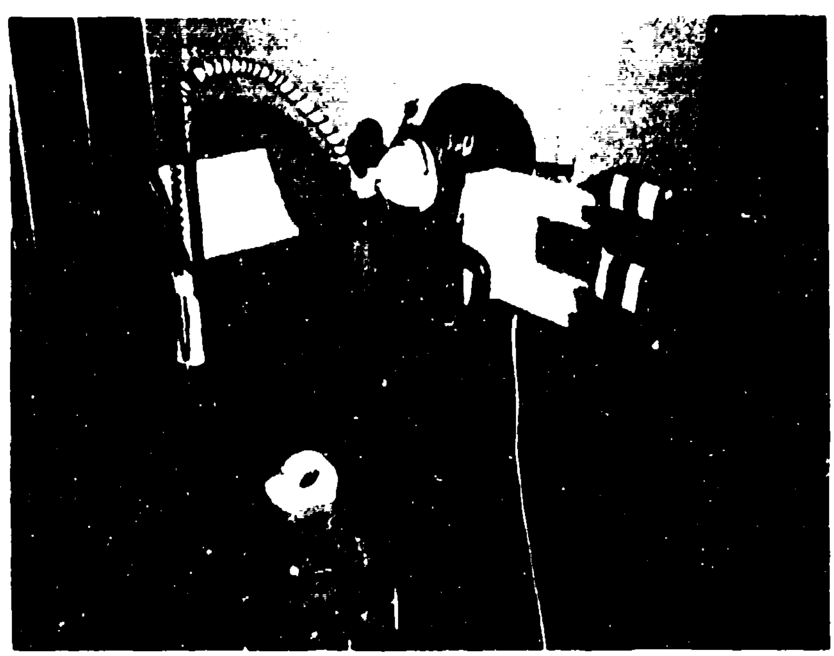

Fig. 5 Microphone Placement in Noise Condition (Back View)

Fig. 6 Microphone Placement in
Helicopter Noise Environment

Acceptability test materials were presented to a listening crew to verify the correctness and quality of the sentence recordings. Nine-speaker master sentence tapes were then assembled.

II. ANALOG COPIES

All copies of analog tape recordings required by the contract were delivered. Tape recorders used in making the recordings were two Ampex 440B 4-Track recorders, one TEAC 7030 GSL 2-Track recorder, and one Ampex 602.2 2-Track recorder Scotch 208 magnetic recording tape was used. Tables 4 and 5 provide a summary of analog tapes delivered.

Tape ID	DRT TAPES Speaker / List	Environ-ment	ANALOG ¼" - 7½ ips 1	2	3	4	5	¼" - 15 ips 1	2	3	4	5	½"-15ips 6	7	8	DIGITAL Seven Track DM	CM	TM	CSM	DNCM	HEL	Nine Track DM	CM	TM	CSM	DNCM	HEL
E-1-A	LL 302A,CH 308B,RH 310A	Quiet	X	X	X			X	X	X			X				X		X			X					
E-1-B	JE 306A,BV 303A,PK 309A	"	X	X	X			X	X	X			X				X		X			X					
E-2-A	LL 302B,CH 307A,RH 310B	"	X	X	X			X	X	X			X				X		X			X					
E-2-B	JE 306B,BV 303B,PK 312B	"	X	X	X			X	X	X			X				X		X			X					
E-3-A	LL 301A,CH 308A,RH 311A	"	X	X	X			X	X	X			X				X		X			X					
E-3-B	JE 305A,BV 304A,PK 312A	"	X	X	X			X	X	X			X				X		X			X					
E-4-A	LL 301B,CH 307B,RH 311B	"	X	X	X			X	X	X			X				X		X			X					
E-4-B	JE 305B,BV 304B,PK 309B	"	X	X	X			X	X	X			X				X		X			X					
E-5-A	JS 317A,LS 315A,MP 314A	"	X	X	X			X	X	X			X				X		X			X					
E-5-B	JS 317B,LS 315B,MP 314B	"	X	X	X			X	X	X			X				X		X			X					
E-6-A	JS 318A,LS 316A,MP 313A	"	X	X	X			X	X	X			X				X		X			X					
E-6-B	JS 318B,LS 316B,MP 313B	"	X	X	X			X	X	X			X				X		X			X					
G-1-A	RH 318A,JE 310A,CH 314A	ABCP	X			X		X			X			X		X										X	
G-1-B	RH 318B,JE 310B,CH 314B	"	X			X		X			X			X		X										X	
G-2-A	RH 317A,JE 309A,CH 313B	Helic					X					X			X												X
G-2-B	RH 317B,JE 309B,CH 313A	"					X					X			X												X
G-3-A	RH 303A,JE 311A,CH 315A	Ship	X			X		X			X			X		X										X	
G-3-B	RH 303B,JE 311B,CH 315B	"	X			X		X			X			X		X										X	
G-4-A	RH 304A,JE 312A,CH 316A,JS 305A	Office	X			X		X			X			X						X		X					
G-4-B	RH 304B,JE 312B,CH 316B,JS 305B	"	X			X		X			X			X						X		X					

Microphone Codes:

1- DM/TM
2- CM/TM
3- CSM/TM
4- DNCM/TM
5- HEL/TM
6- DM/CM/TM/CSM
7- DM/DNCM/TM
8- DM/HEL/TM

MICROPHONE ABBREVIATIONS

DM - Altec 659A Dynamic Mic.
CM - Western Electric T1 Carbon Mic.
TM - Grason-Stadler E7300M Throat Mic.
CSM - General Radio 1560-P5 Ceramic Studio Mic.
DNCM - Roanwell Dynamic Noise Cancelling Mic.
HEL - Electrovoice M-78/A1C Dynamic Helicopter Mic.

TABLE 4

Tape ID	VQAT TAPES Speaker / List	Environ-ment	ANALOG													DIGITAL											
			¼" - 7½ ips					¼" - 15 ips					½"-15ips			Seven Track						Nine Track					
			1	2	3	4	5	1	2	3	4	5	6	7	8	DM	CM	TM	CSM	DNCM	HEL	DM	CM	TM	CSM	DNCM	HEL
F-1	CH 4A,LL 1A,RH 6A,JE 3A,PK 7A, BV 2A,JS 5A,MP 8A,LS 9A	Quiet	X	X				X	X	X			X									X	X		X		
F-2	CH 4B,LL 1B,RH 6B,JE 3B,PK 7B, BV 2B,JS 5B,MP 8B,LS 9B	"	X	X				X	X	X			X									X	X		X		
F-3	CH 4C,LL 1C,RH 6C,JE 3C,PK 7C, BV 2C,JS 5C,MP 8C,LS 9C	"	X	X				X	X	X			X									X	X		X		
H-1	RH 6A,JE 3A,CH 4A,RH 6B,JE 3B, CH 4B,RH 6C,JE 3C,CH 4C	ABCP	X			X		X			X			X								X				X	
H-2	RH 6A,JE 3A,CH 4A,RH 6B,JE 3B, CH 4B,RH 6C,JE 3C,CH 4C	Helic					X					X			X												X
H-3	RH 6A,JE 3A,CH 4A,RH 6B,JE 3B, CH 4B,RH 6C,JE 3C,CH 4C	Ship	X			X		X			X			X								X				X	
H-4 (I)	RH 6A,JE 3A,CH 4A,JS 5A, RH 6B,JE 3B,CH 4B,JS 5B	Office	X			X		X			X			X								X				X	
H-4 (II)	RH 6C,JE 3C,CH 4C,JS 5C	"	X			X		X			X			X								X				X	

Microphone Codes:

1- DM/TM
2- CM/TM
3- CSM/TM
4- DNCM/TM
5- HEL/TM
6- DM/CM/TM/CSM
7- DM/DNCM/TM
8- DM/HEL/TM

MICROPHONE ABBREVIATIONS

DM - Altec 659A Dynamic Mic.
CM - Western Electric T1 Carbon Mic.
TM Grason-Stadler E7300M Throat Mic.
CSM General Radio 1560-P5 Ceramic Studio Mic.
DNCM - Roanwell Dynamic Noise Cancelling Mic.
HEL - Electrovoice M-78/A1C Dynamic Helicopter Mic.

TABLE 5

III. ANALOG TO SEVEN TRACK DIGITAL CONVERSION

As an intermediate step in producing nine-track digital versions of the master tapes a seven-track digital tape was recorded. Seven-track tapes were recorded on one half inch digital tape at 800 bytes per inch NRZI in ASCII code and format. Digital sampling was at 12,000 Hz, with each sample digitally represented in two's compliment format by at least 11 bits plus a sign bit. The speech signal amplitude range was set at + 5 volts peak. Figure 7 shows a block diagram of the equipment used in the analog to digital conversion. Table 6 provides a summary of seven-track tapes delivered.

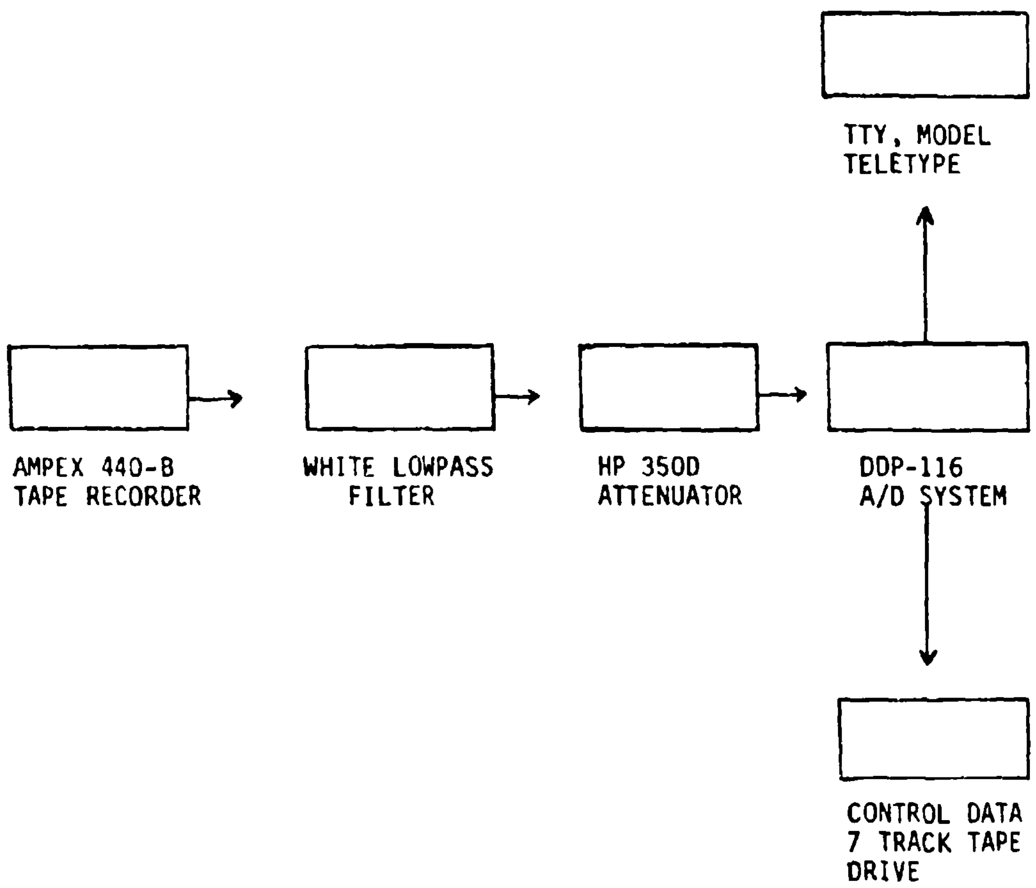

Figure 7. EQUIPMENT SET UP FOR ANALOG TO SEVEN TRACK CONVERSION.

TABLE 6 (1) SEVEN TRACK DIGITAL TAPES

Tape	Speaker	Sex	List	Date	Mic.	Environment	Place
E1A1/E1A2	LL	M	302A	8/24/74	GR	Quiet	Dynastat
	CH	M	308B	8/29/74	"	"	"
E1A3/E1B1	RH	M	310A	9/04/74	"	"	"
	JE	M	306A	9/05/74	"	"	"
E1B2/E1B3	BV	M	303A	9/24/74	"	"	"
	PK	M	309A	9/23/74	"	"	"
E2A1/E2A2	LL	M	302B	8/24/74	"	"	"
	CH	M	307A	8/29/74	"	"	"
E2A3	RH	M	310B	9/04/74	"	"	"
E2B1	JE	M	306B	9/05/74	"	"	"
E2B2/E2B3	BV	M	303B	9/24/74	"	"	"
	PK	M	312B	9/23/74	"	"	"
E3A1/E3A2	LL	M	301A	8/25/74	"	"	"
	CH	M	308A	8/29/74	"	"	"
E3A3/E3B1	RH	M	311A	9/04/74	"	"	"
	JE	M	305A	8/28/74	"	"	"
E3B2/E3B3	BV	M	304A	9/24/74	"	"	"
	PK	M	312A	9/23/74	"	"	"
E4A1/E4A2	LL	M	301B	8/25/74	"	"	"
	CH	M	307B	8/29/74	"	"	"
E4A3/E4B1	RH	M	311B	9/04/74	"	"	"
	JE	M	305B	8/24/74	"	"	"
E4B2/E4B3	BV	M	304B	9/24/74	"	"	"
	PK	M	309B	9/23/74	"	"	"
E5A1	JS	F	317A	8/30/74	"	"	"
E5A2	LS	F	315A	9/20/74	"	"	"
E5A3/E5B1	MP	F	314A	9/21/74	"	"	"
	JS	F	317B	8/30/74	"	"	"
E5B2	LS	F	315B	9/20/74	"	"	"
E5B3	MP	F	314B	9/21/74	"	"	"

TABLE 6 (2) SEVEN TRACK DIGITAL TAPES

Tape	Speaker	Sex	List	Date	Mic.	Environment	Place
E6A1	JS	F	318A	8/30/74	GR	Quiet	Dynastat
E6A2	LS	F	316A	9/05/74	"	"	"
E6A3/E6B1	MP	F	313A	9/21/74	"	"	"
	JS	F	318B	8/30/74	"	"	"
E6B2/E6B3	LS	F	316B	9/05/74	"	"	"
	MP	F	313B	9/21/74	"	"	"

TABLE 6 (3) SEVEN TRACK DIGITAL TAPES

Tape	Speaker	Sex	List	Date	Mic.	Environment	Place
E1A1/E1A2	LL	M	302A	8/24/74	Carbon	Quiet	Dynastat
	CH	M	308B	8/29/74	"	"	"
E1A3/E1B1	RH	M	310A	9/04/74	"	"	"
	JE	M	306A	9/05/74	"	"	"
E1B2/E1B3	BV	M	303A	9/24/74	"	"	"
	PK	M	309A	9/23/74	"	"	"
E2A1/E2A2	LL	M	302B	8/24/74	"	"	"
	CH	M	307A	8/29/74	"	"	"
E2A3/E2B1	RH	M	310B	9/04/74	"	"	"
	JE	M	306B	9/05/74	"	"	"
E2B2/E2B3	BV	M	303B	9/24/74	"	"	"
	PK	M	312B	9/23/74	"	"	"
E3A1/E3A2	LL	M	301A	8/25/74	"	"	"
	CH	M	308A	8/29/74	"	"	"
E3A3/E3B1	RH	M	311A	9/04/74	"	"	"
	JE	M	305A	8/28/74	"	"	"
E3B2/E3B3	BV	M	304A	9/24/74	"	"	"
	PK	M	312A	9/23/74	"	"	"
E4A1/E4A2	LL	M	301B	8/25/74	"	"	"
	CH	M	307B	8/29/74	"	"	"
E4A3/E4B1	RH	M	311B	9/04/74	"	"	"
	JE	M	305B	8/24/74	"	"	"
E4B2/E4B3	BV	M	304B	9/24/74	"	"	"
	PK	M	309B	9/23/74	"	"	"
E5A1	JS	F	317A	8/30/74	"	"	"
E5A2	LS	F	315A	9/20/74	"	"	"
E5A3/E5B1	MP	F	314A	9/21/74	"	"	"
	JS	F	317B	8/30/74	"	"	"
E5B2	LS	F	315B	9/20/74	"	"	"
E5B3	MP	F	314B	9/21/74	"	"	"
E6A1	JS	F	318A	8/30/74	"	"	"
E6A2	LS	F	316A	9/05/74	"	"	"

TABLE 6 (4) SEVEN TRACK DIGITAL TAPES

Tape	Speaker	Sex	List	Date	Mic.	Environment	Place
E6A3/E6B1	MP	F	313A	9/21/74	Carbon	Quiet	Dynastat
	JS	F	318B	8/30/74	"	"	"
E6B2/E6B3	LS	F	316B	9/05/74	"	"	"
	MP	F	313B	9/21/74	"	"	"
G1A1/G1A2	RH	M	318A	9/07/74	Altec	ABCP	"
	JE	M	310A	9/14/74	"	"	"
G1A3/G1B1	CH	M	314A	9/07/74	"	"	"
	RH	M	318B	9/07/74	"	"	"
G1B2/G1B3	JE	M	310B	9/14/74	"	"	"
	CH	M	314B	9/07/74	"	"	"
G3A1	RH	M	303A	9/11/74	"	Shipboard	"
G3A2	JE	M	311A	9/15/74	"	"	"
G3A3	CH	M	315A	9/12/74	"	"	"
G3B1	RH	M	303B	9/11/74	"	"	"
G3B2	JE	M	311B	9/15/74	"	"	"
G3B3	CH	M	315B	9/12/74	"	"	"
G4A1	RH	M	304A	9/15/74	Roanwell	Office	"
G4A2	JE	M	312A	9/15/74	"	"	"
G4A3/G4A4	CH	M	316A	9/15/74	"	"	"
	JS	F	305A	9/16/74	"	"	"
G4B1	RH	M	304B	9/15/74	"	"	"
G4B2	JE	M	312B	9/15/74	"	"	"
G4B3/G4B4	CH	M	316B	9/15/74	"	"	"
	JS	F	305B	9/16/74	"	"	"

IV. CONVERSION OF SEVEN-TRACK DIGITAL TAPES TO NINE-TRACK FORMAT

Seven-track digital tapes were converted to nine-track digital format via a Dynastat written FORTRAN program on a Data General NOVA 2/10 computer system. Sixteen bit data words were constructed to include a twelve bit sample plus four sync bits as specified in the Statement of Work. Records were 1000 words each (2000 bytes). Nine-track tapes were written in even parity at 800 bytes per inch. Each tape file is prefaced by a header record which specifies various analog recording data including: type of analog material (i.e., DRT scrambling, acceptability test sentence, tape announcement, speaker announcement, or calibration tone) microphone information, speaker identification, recording dates and other data as outlined in subject Statement of Work. A summary of the nine-track digital tapes delivered by Dynastat is shown in Table 7.

TABLE 7 (1) NINE TRACK DIGITAL TAPES

Tape	Speaker	Sex	List	Date	Mic.	Environment	Place
E1A1/E1A2	LL	M	302A	8/24/74	Altec	Quiet	Dynastat
	CH	M	308B	8/29/74	"	"	"
E1A3/E1B1	RH	M	310A	9/04/74	"	"	"
	JE	M	306A	9/05/74	"	"	"
E1B2/E1B3	BV	M	303A	9/24/74	"	"	"
	PK	M	309A	9/23/74	"	"	"
E2A1/E2A2	LL	M	302B	8/24/74	"	"	"
	CH	M	307A	8/29/74	"	"	"
E2A3/E2B1	RH	M	310B	9/04/74	"	"	"
	JE	M	306B	9/05/74	"	"	"
E2B2/E2B3	BV	M	303B	9/24/74	"	"	"
	PK	M	312B	9/23/74	"	"	"
[illegible]/E3A2	LL	M	301A	8/25/74	"	"	"
	CH	M	308A	9/29/74	"	"	"
E3A3/E3B1	RH	M	311A	9/04/74	"	"	"
	JE	M	305A	8/28/74	"	"	"
E3B2/E3B3	BV	M	304A	9/24/74	"	"	"
	PK	M	312A	9/23/74	"	"	"
E4A1/E4A2	LL	M	301B	8/25/74	"	"	"
	CH	M	307B	8/29/74	"	"	"
E4A3/E4B1	RH	M	311B	9/04/74	"	"	"
	JE	M	305B	8/24/74	"	"	"
E4B2/E4B3	BV	M	304B	9/24/74	"	"	"
	PK	M	309B	9/23/74	"	"	"
E5A1	JS	F	317A	8/30/74	"	"	"
E5A2	LS	F	315A	9/20/74	"	"	"
E5A3	MP	F	314A	9/21/74	"	"	"
E5B1	JS	F	317B	8/30/74	"	"	"
E5B2	LS	F	315B	9/20/74	"	"	"
E5B3	MP	F	314B	9/21/74	"	"	"
E6A1	JS	F	318A	8/30/74	"	"	"
E6A2	LS	F	316A	9/05/74	"	"	"
E6A3	MP	F	313A	9/21/74	"	"	"
E6B1	JS	F	318B	8/30/74	"	"	"
E6B2	LS	F	316B	9/05/74	"	"	"
E6B3	MP	F	313B	9/21/74	"	"	"

TABLE 7 (2) NINE TRACK DIGITAL TAPES

Tape	Speaker	Sex	List	Date	Mic.	Environment	Place
G1A1	RH	M	318A	9/07/74	Roanwell	ABCP	Dynastat
G1A2	JE	M	310A	9/14/74	"	"	"
G1A3	CH	M	314A	9/07/74	"	"	"
G1B1	RH	M	318B	9/07/74	"	"	"
G1B2/G1B3	JE CH	M M	310B 314B	9/14/74 9/07/74	" "	" "	" "
G2A1	RH	M	317A	9/11/74	Helicopter	Helicopter	"
G2A2	JE	M	309A	9/14/74	"	"	"
G2A3	CH	M	313B	9/12/74	"	"	"
G2B1	RH	M	317B	9/11/74	"	"	"
G2B2	JE	M	309B	9/14/74	"	"	"
G2B3	CH	M	313A	9/12/74	"	"	"
G3A1	RH	M	303A	9/11/74	Roanwell	Shipboard	"
G3A2	JE	M	311A	9/15/74	"	"	"
G3A3	CH	M	315A	9/12/74	"	"	"
G3B1	RH	M	303B	9/11/74	"	"	"
G3B2	JE	M	311B	9/15/74	"	"	"
G3B3	CH	M	315B	9/12/74	"	"	"
G4A1	RH	M	304A	9/15/74	Altec	Office	"
G4A2	JE	M	312A	9/15/74	"	"	"
G4A3	CH	M	316A	9/15/74	"	"	"
G4A4	JS	F	305A	9/16/74	"	"	"
G4B1	RH	M	304B	9/15/74	"	"	"
G4B2	JE	M	312B	915/74	"	"	"
G4B3	CH	M	316B	9/15/74	"	"	"
G4B4	JS	F	305B	9/16/74	"	"	"

REFERENCES

Coulter, D. C. Quality and Acceptability Testing of Voice Processors for Military Applications, NRL Report No. 7773 (November, 1974).

Fairbanks, G. "Test of Phonemic Differentiation: The Rhyme Test," J. Acoust. Soc. Am. 30 (1958) 596-600.

Grether, C. B. and R. W. Stroh. "Subjective Evaluation of Differential Pulse Code Modulation Using the Speech 'Goodness' Rating Scale," 1972 Conference on Speech Communication and Processing, AFCRL-72-0120, (1972), 175-178.

Guilford, J. P. Psychometric Methods, New York: McGraw-Hill, 1954.

Hecker, H. L. and C. E. Williams. "Choice of Reference Conditions for Speech Preference Tests," J. Acoust. Soc. Am. 39 (1966) 946-952.

Helson, Harry. Adaptation Level Theory, in Sigmund Koch, ed., Psychology: A Study of Science, Vol. 1., New York: McGraw-Hill, 1959.

House, A. S.; Williams, C. E.; Hecker, H. L.; and Kryter, K. D. "Articulation Testing Methods: Consonantal Differentiation with a Closed Response Set," J. Acoust. Soc. Am. 37 (965) 158-166.

McDermott, B. J. "Multidimensional Analyses of Circuit Quality Judgments," J. Acoust. Soc. Am. 45 (1969) 774-781.

Munson, W. A., and J. E. Karlin. "Isopreference Method for Evaluating Speech Transmission Circuits," J. Acoust. Soc. Am. 34 (1962) 762-774.

Osgood, C. E. "The Nature and Measurement of Meaning," Psychol. Bull. 49 (1952) 197-237.

Parducci, A. "Sequential Effects in Judgment," Psychol. Bull. 61 (1964) 163-167.

Richards, D. L., and J. Swaffield. "Assessment of Speech Communications Links, Proc. I.E.E. 106 (1959) 77-92.

Rothauser, E. H.; Urbanek, G. E.; and Pach, W. P. Speech Quality Measurements, Institut Fur Niederfrequenztechnik der Technische Hochschule Wien, Final Scientific Report (1967).

Solomon, L. N. "Semantic Approach to the Perception of Complex Sounds," J. Acoust. Soc. Am. 30 (1958) 421-425.

Veldman, D. J. Fortran Programming for the Behavioral Sciences, Holt, 1967.

Voiers, W. D. "Perceptual Bases of Speaker Identity," J. Acoust. Soc. Am. 36 (1964) 1065-1073.

Voiers, W. D. "Present Status of the Diagnostic Rhyme Test," Presented at the Meetings of the Groupement des Acousticiens de Langue Francaise, Groupe, "Communication Parlee," Aix-en-Provence, France (1971)

Voiers, W. D. "Interim Quality Evaluation Plans and Procedures," Interim Report, Contract No. DCA100-74-C-0056 (November 1974).

Voiers, W. D.; Cohen, M. F.; and Mickunas, J. "Evaluation of Speech Processing Devices, I. Intelligibility, Quality, Speaker Recognizability," AFCRL-65-926 (1965).

Voiers, W. D., and Smith, C. P. "Diagnostic Evaluation of Intelligibility in Present-Day Digital Vocoders," AFCRL-72-0120 (1972).

Voiers, W. D.; Sharpley, A. D.; and Hehmsoth, C. J. "Research on Diagnostic Evaluation of Speech Intelligibility," Final Report, Contract No. AF 19(628)-70-C-0182, AFCRL-72-0694 (January 1973).

Williams, C. E.; Hecker, H. L.; Stevens, K. N.; and Woods, B. "Intelligibility Test Methods for the Evaluation of Communication Systems," Final Report, Contract No. AF 19(628)-5659, ESD-TR-66-677, (December, 1966).

Winer, B. J. Statistical Principles in Experimental Design, Second Edition, New York: McGraw-Hill, 1972.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER D-76-001-U	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Methods of Predicting User Acceptance of Voice Communication Systems		5. TYPE OF REPORT & PERIOD COVERED Scientific: FINAL 10 June 1974-30 June 1976
		6. PERFORMING ORG. REPORT NUMBER D-76-001-U
7. AUTHOR(s) William D. Voiers		8. CONTRACT OR GRANT NUMBER(s) DCA100-74-C-0056
9. PERFORMING ORGANIZATION NAME AND ADDRESS Dynastat, Inc. 2704 Rio Grande Austin, Texas 78705		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Communications Agency Washington, D.C. 20305		12. REPORT DATE 15 July 1976
		13. NUMBER OF PAGES 139
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Defense Communications Engineering Center 1860 Wiehle Avenue Reston, Virginia 22090		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT (of this Report)

[redacted]

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

SPEECH ACCEPTABILITY
SPEECH QUALITY
NARROWBAND COMMUNICATION SYSTEMS

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A program of research was undertaken to develop improved methods of predicting user acceptance of voice communication systems. Two methods were developed and standardized: The Paired Acceptability Rating Method (PARM) and the Quality Acceptance Rating Test (QUART).

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)