AFCRL-66-24 SRRC-RR-65-115

AD630428

PERFORMANCE EVALUATION OF SPEECH PROCESSING DEVICES

II. THE ROLE OF INDIVIDUAL DIFFERENCES

by

William D. Voiers

CLEARINGHOUSE
FOR FEDERAL SCIENTIFIC AND
TECHNICAL INFORMATION

Hardcopy	Microfiche		
$ 3.00	$ 0.75	66 pp	as

ARCHIVE COPY

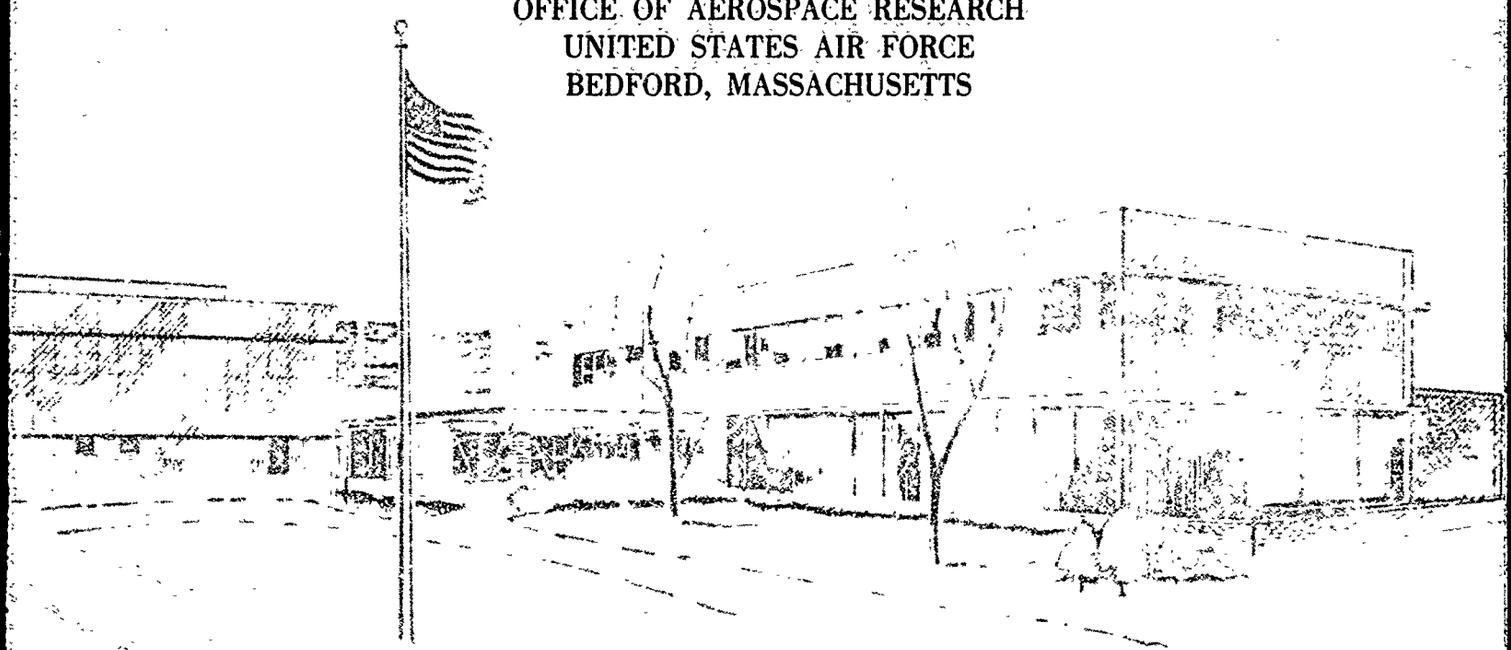
Code 1

Contract No. AF19(628)-4987
Project 4610 Task 461002

Scientific Report: December 1965
Period Covered: 1 March — 30 November 1965

Prepared for

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



SPERRY RAND RESEARCH CENTER

SUDBURY, MASSACHUSETTS

AFCRL-66-24

SRRC-RR-65-115

PERFORMANCE EVALUATION OF SPEECH PROCESSING DEVICES

II. THE ROLE OF INDIVIDUAL DIFFERENCES

by

William D. Voiers

Contract No. AF19(628)-4987

Project 4610 Task 461002

Scientific Report: December 1965

Period Covered: 1 March — 30 November 1965

Prepared for

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

S R R C

ABSTRACT

This report treats three topics related to the phenomena of individual differences in speech: (1) the effect upon speech intelligibility of the listener's familiarity with the speaker's voice, (2) the nature and number of elementary ways, or Perceived Acoustic Traits (PAT's), in terms of which voices are perceived to differ from each other, (3) the physical bases of perceived acoustic traits.

It was found that speech intelligibility, as measured by the Diagnostic Rhyme Test, was unaffected by the degree of the listener's familiarity with the speaker's voice; the basis of familiarity was a maximum of three presentations of a two-minute sample of prose (Gettysburg Address).

A voice-rating experiment, involving both monopolar and bipolar rating scales, failed to reveal any previously unidentified perceived acoustic traits. However, several items were found which offer possibilities for increased precision in the evaluation of previously identified PAT's.

A factor analysis of 23 physical and perceptual voice variables yielded results indicating that physical correlates of some perceived acoustic traits may be qualitatively shifted under certain conditions of stimulus impoverishment. The physical bases of the PAT's, Pitch-Magnitude and Animation-Rate, appear to remain relatively stable under various conditions of stimulus impoverishment, but the physical bases of Loudness-Roughness and Clarity-Beauty may be drastically altered by frequency distortion and vocoderization.

TABLE OF CONTENTS

		Page
INTRODUCTION		1
CHAPTER I	FAMILIARITY WITH THE SPEAKER AS A FACTOR IN SPEECH INTELLIGIBILITY	3
	A. Background	4
	B. Experiments On The Effects Of Familiarity	7
	C. Conclusions	13
CHAPTER II	THE IDENTIFICATION OF PERCEIVED ACOUSTIC TRAITS	14
	A. A Search For Additional Perceived Acoustic Traits	18
	B. Conclusions	29
CHAPTER III	THE PHYSICAL CORRELATES OF PERCEIVED ACOUSTIC TRAITS	30
	A. Theoretical Considerations	30
	B. A Study Of The Physical Basis Of Perceived Acoustical Traits	34
	C. Conclusions	55
REFERENCES		57

LIST OF FIGURES

Figure		Page
1	Effects of Repeated Presentations of Diagnostic Rhyme Test (DRT) Materials (vocoded).	9
2	Design of Experiments I and II. In Experiment I all test materials were vocoded; all familiarization materials were unprocessed speech (continuous). In Experiment II all test and familiarization materials were vocoded.	11
3	Semantic Differential Rating Form (Voiers, 1964).	15
4	Standard Voice-Rating Form.	20
5	Experimental Voice-Rating Form.	21
6	Factor Loadings of Thirty-Two Rating Dimensions in the I-II Plane of the Perceived Acoustic Trait Space.	28
7	Factor Loadings of Twenty-Three Voice Variables in the I-II Plane of the Perceived Acoustic Trait Space.	54

LIST OF TABLES

Table		Page
1	DRT Total Scores for Various Conditions	12
2	Summary of Results of Analysis of Variance of Voice Ratings	23
3	Final Factor Loadings for Thirty-Two Rating Dimensions	26
4	Physical and Perceived Voice Variables Subjected to Factor Analysis	36
5	Final Factor Loadings of Twenty-Three Voice Variables	40
6	Correlations Among Eleven Voice Variables	43
7	Correlations Between Voice Ratings on PAT 3 (Animation-Rate) and Seven Physical Variables Under Four Transmission Conditions	49

ACKNOWLEDGMENTS

Several individuals contributed significantly to the preparation of this report. Mr. Juozas Mickunas provided assistance in data collection and statistical computation. Discussions with Miss Marion F. Cohen on matters of speech production and perception were extremely helpful in the planning and conducting of the program. Miss Virginia Miethe's services in computer programming and analysis of results were indispensable.

INTRODUCTION

Three issues relating to the phenomenon of individual differences in speech and their implications for vocoder performance are treated in this report. The first concerns the influence of the listener's familiarity with individualistic voice characteristics upon his interpretation of the received speech signal. The second concerns the nature of voice recognition by human listeners—specifically the question of what listeners perceive to be the major distinguishing characteristics or traits of individual voices. The third issue concerns the physical acoustical characteristics or traits which distinguish voices from each other — in particular those characteristics by which listeners may identify the voices of individual speakers.

Each of the above issues has been the subject of a series of investigations during the past year. While several of these investigations have yielded results of intrinsic interest, others have served primarily to resolve various methodological issues. In the following chapters, primary emphasis is given to the most significant experiment from each series. Other experiments are cited only as their results bear upon a particular issue under discussion.

Chapter 1 presents results concerning listener-familiarity-with-the-speaker's-voice as a factor in speech intelligibility. Our primary concern here is with the implications of familiarity phenomena for the design of intelligibility tests.

Chapter 2 contains a discussion of previous research on the problem of quantifying the perceptually significant characteristics of individual voices. One experiment concerned with the elementary dimensions of perceived variability among voices (perceived acoustic traits), is described in some detail.

Chapter 3 is devoted to the issue of the physical bases of perceived differences among voices. The results of a factor analysis of 23 perceptual and physical voice variables are presented and their implications for a theory of voice recognition are discussed.

CHAPTER I

FAMILIARITY WITH THE SPEAKER AS A FACTOR IN SPEECH INTELLIGIBILITY

In view of the extent to which individual speech varies, both between and within dialectal groups, it would seem inevitable that a listener's familiarity with the idiosyncracies of a speaker's voice determines, to some degree, his ability to interpret the speech of that speaker. The common experience of individuals who migrate from one dialectal region to another strongly supports this proposition. The experiences of those who have occasion to cope with pathological speech likewise attest to the effects of familiarity with the speaker's idiosyncracies upon the listener's ability to decode such speech. A related phenomenon is reported by individuals who have occasion to deal with vocoded or otherwise-processed speech. Engineers and scientists who work with speech-compression devices commonly report that processed speech becomes more intelligible to them as their experience with it increases. Questions arise, however, as to what generalizations are warranted by these various special instances of the phenomenon.

To what extent, for example, does listener familiarity with a speaker contribute to speech intelligibility in the case of normal speakers and listeners of similar dialectal background? To what extent does such familiarity increase the listener's tolerance for speech which is degraded in one way or another? Does familiarity with the speaker under normal or high-fidelity transmission conditions enhance the listener's capacity to cope with the case of degraded speech? The present chapter bears upon such questions. Specifically, it is concerned with the effects of different types of familiarity upon the intelligibility of vocoded speech as evaluated by the Diagnostic Rhyme Test.

A. BACKGROUND

While the literature pertaining directly to the present problem is not extensive, there are at least two studies which merit notice here.

The first is a study by Peters (1955), in which listeners were tested after various amounts (0, 2, 4 and 8 minutes) of exposure to the speech of a particular speaker. Two transmission conditions were studied: one in which all speech materials were presented in the quiet and a second one in which all materials were presented in approximately speech-shaped noise at a S/N ratio of 0 dB.

Under both conditions, listener reception scores improved with increasing pre-exposure to the test voice although the observed trends were somewhat different for the two conditions. In the quiet, listener reception scores increased significantly for a gain of approximately 5 percentage points (87.0 - 92.4) after eight minutes of exposure to the test voice. However, the greatest improvement (4 percentage points) occurred after only two minutes of familiarization. The total improvement in listener reception scores for the noisy transmission condition was approximately the same as that for the quiet condition. However, substantial improvement occurred only after four minutes of exposure to the test speaker's voice. In any case, these results are consistent with the hypothesis that familiarity with a voice tends to enhance intelligibility for a given listener.

While not concerned specifically with the role of speaker familiarity, an experiment by Ladefoged and Broadbent (1957) has some bearing on this issue. These investigators found that, with experimental manipulation of formant-frequency relationships in a carrier phrase, the

judged identity of a subsequent, unmanipulated vowel sound could be altered in a predictable manner. For example, an artificially-lowered second formant in a carrier phrase caused subsequent unaltered synthetic vowels to be identified as if the listener were taking into account the effects of the formant frequency displacement. Such "errors" are significant primarily because they attest to the listener's capability to detect the idiosyncracies of a speech source from a sample of its output and then to adjust his responses accordingly. Presumably, the sensory-cognitive process involved here would also operate to render the linguistic content from the speech of a speaker whose acoustic "signaling alphabet" is novel for a given listener. The results of these studies leave one important issue unresolved, however — that of where and how the adaptive process in question takes place.

One possibility is that this phenomenon involves a genuine change in sensory-perceptual response to a particular stimulus parameter, that is, a change of "adaptation level" as conceived by Helson and his followers (1961). While not couching their interpretation explicitly in terms of adaptation level theory, Ladefoged and Broadbent seem to espouse this general type of explanation. A second possibility, however, is that the phenomenon in question can find explanation in purely cognitive, "nonsensory," terms. This explanation would maintain that the response changes induced (e.g., by manipulations of the carrier phrase) do not derive from changes in the normal sensory correlates of the test stimulus, but rather that they depend simply upon a modification of the <u>interpretation</u> given a sensory-perceptual event by the listener. In other words, this explanation attributes the results of Ladefoged and

Broadbent to essentially the same cognitive processes as those involved in the learning of a new language.

Experimental resolution of this issue may prove difficult. In any case, the two types of mechanism in question are not incompatible, although the conditions of their occurrence may be somewhat different. In situations where the cognitive aspects of the listener's task predominate, we would expect experimentally-determined response changes to be more pronounced in cases involving simple sensory discriminations.

Thus, while we may expect improvements in the net efficiency with which a listener discriminates the speaker's intent as he gains familiarity with the speaker, with a particular transformation of the speech signal, and perhaps with testing situations in general, the observed effects of these improvements will vary with the manner in which listener performance is evaluated. Where the listener's task is one of simple discrimination, the effects of familiarity with various source characteristics (e.g., speaker idiosyncracies) will tend to be minimal. Where performance depends on more complex cognitive processes, the effects of familiarity will be more pronounced.

While factors other than speaker familiarity are undoubtedly involved, differences among various of the better known intelligibility tests in terms of susceptibility to "practice effects" are in accord with this principle. The PB word lists, which require the listener to make an "absolute recognition response," are notoriously affected by the listener's past experience with the test materials. The Fairbanks Rhyme Test, which requires less complex discriminations of the listener, is alleged to be somewhat less sensitive to listener experience, while several of the more

recent derivatives of the Fairbanks test (e.g., the Modified Rhyme Test of House et al., 1963) are effectively free of such effects.

On the basis of these various considerations, the Diagnostic Rhyme Test (Voiers, Cohen and Mickunas, 1965) could be expected to provide results which are minimally affected by the listener's familiarity with the voice of the speaker or with any other invariants of the testing situation.

The DRT is a two-choice rhyme test in which the stimulus materials are isolated words and the listener's task is to select one of two alternatives which differ (at least nominally) with respect to a single "attribute" of the initial consonant phoneme. Thus, the listener's task is as nearly one of simple sensory discrimination as can be provided in circumstances involving speech stimuli.

B. EXPERIMENTS ON THE EFFECTS OF FAMILIARITY

In the course of more than fifty speech evaluations with the Diagnostic Rhyme Test, a number of opportunities have arisen to test for the effects of various kinds of listener experience upon speech intelligibility. However, except for some indication that scores improve slightly during the first ten minutes or so of the listener's initial exposure to the testing situation, there has been little evidence that familiarity with any aspect of the testing situation (including the speaker) affects the listener's performance in any systematic fashion.

Generally, precautions were taken to minimize any possible effects of familiarity with various aspects of the test situations, but on certain occasions several potentially significant aspects were invariant over the course of a succession of tests with a particular listening crew. Among these aspects were the speaker and the general

form of processing to which test speech materials had been subjected. Even here, the opportunities available to the listener for familiarization with the speaker's voice were limited to those provided by the DRT speech materials (i.e., isolated words). Contextual aids to the detection of speaker characteristics were negligible. The absence of feedback to the listener on the quality of his performance further tended to minimize the effects of learning of speaker idiosyncracies. A test was made, however, of the combined effects of several factors potentially conducive to improvement in DRT scores as a function of listener's experience. This involved presentation of identical DRT materials on six successive occasions, each pair separated by small amounts of other types of speech materials. The results are presented in Fig. 1.

Virtually no trends can be detected to indicate that experiences with the test materials, the speaker's voice, or a particular answer sheet combine to enhance listener performance on the DRT in any manner.

These results are consistent with the proposition that a simple discrimination test, as exemplified by the DRT, does <u>not</u> show improvement as a function of the listener's familiarity with the testing situation and, in particular, the speaker's voice. Their significance, however, is found more in practical than theoretical terms. For, while they provide confirmation of an important aspect of the validity of DRT scores, they do not rule out possible effects of familiarity under circumstances more representative of the actual communications situation. The results of two other experiments bear on this possibility. The Diagnostic Rhyme Test was used in these experiments also, but other steps were taken to provide the listener with opportunities to familiarize himself with the speaker's voice. Listeners were tested before and after exposure to

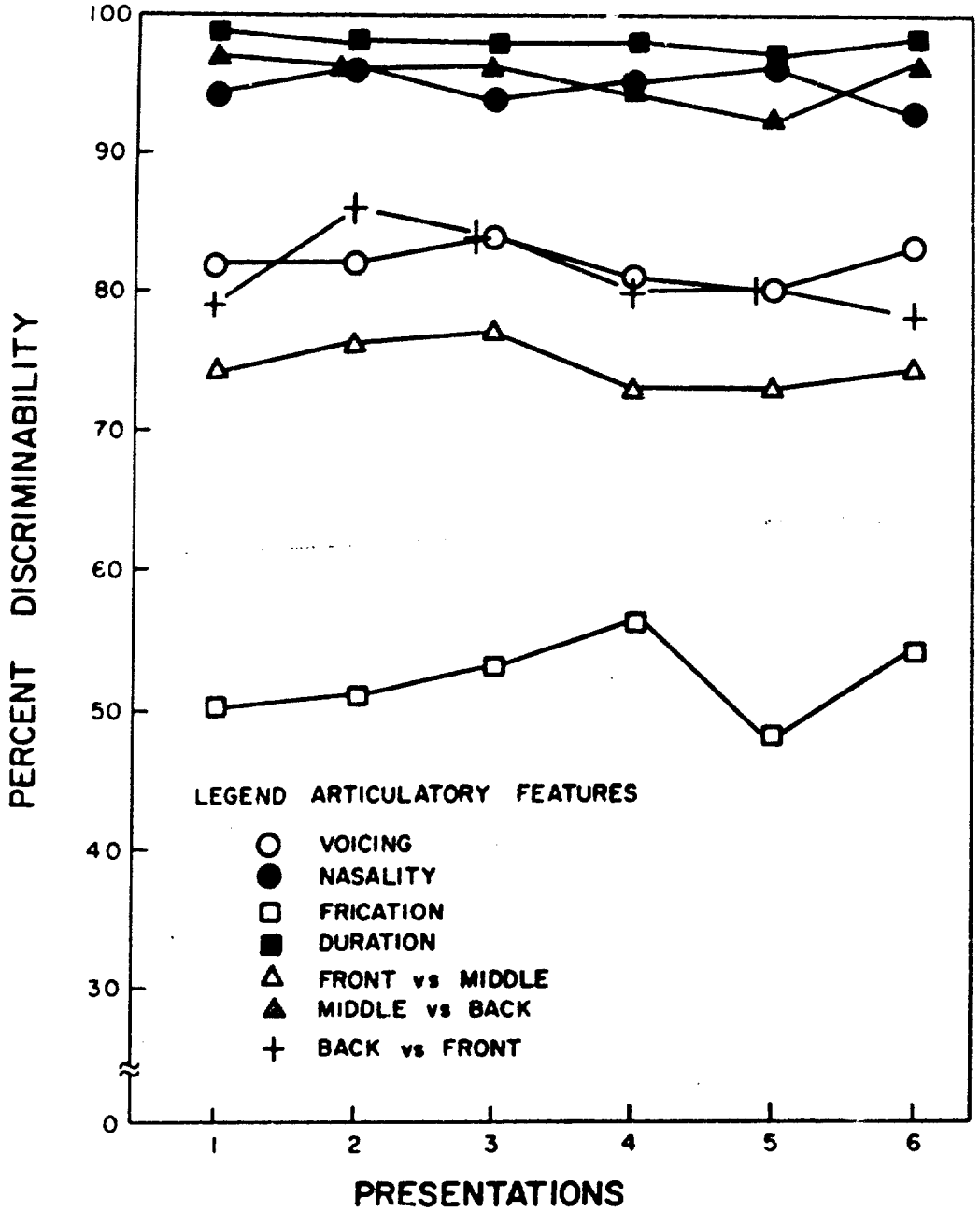


Figure 1. Effects of Repeated Presentations of Diagnostic Rhyme Test (DRT) Materials (vocoded).

samples of continuous speech. Under one set of conditions (experimental) the intervening speech materials were provided by the voice of the test speaker (three presentations of the Gettysburg Address). Under another set of conditions (control) the intervening speech materials were provided by three different, "non-test" speakers. The listeners had had no previous exposure to any of the voices involved.

In both experiments all test materials were processed by a 2400 bit digital vocoder. In the first, however, unprocessed familiarization materials were employed, while in the second both the test and familiarization materials were vocoded. The designs of the two experiments are shown in Fig. 2. The results are summarized in Table 1.

From Table 1 it is apparent that familiarization with the speaker's voice, as provided for in these experiments, does not contribute substantially, if even consistently, to subsequent listener performance on the DRT. Neither familiarization with the speaker's normal speech nor familiarization with his processed speech appears to enhance the intelligibility of the processed speech. There is, however, some indication that certain familiarization conditions affected listener performance by shaping his expectations concerning the character of the test speech: In the first experiment, those listeners who experienced the control condition first exhibited a performance decrement under the subsequently experienced experimental condition. A similar, though less pronounced, effect is also apparent in the second experiment. While these trends are suggestive, however, further research is clearly required on this issue.

EXPERIMENT I

	Test	Familiarization				Test	Test	Familiarization				Test
GROUP I	A	A_0	A_0	A_0	A_0	A	A	W	X	Y	Z	A
GROUP II	A	W	X	Y	Z	A	A	A_0	A_0	A_0	A_0	A
GROUP I	B	B_0	B_0	B_0	B_0	B	B	W	X	Y	Z	B
GROUP II	B	W	X	Y	Z	B	B	B_0	B_0	B_0	B_0	B

EXPERIMENT II

	Test	Familiarization				Test	Test	Familiarization				Test
GROUP I	C	C_0	C_0	C_0	C_0	C	C	W	X	Y	Z	C
GROUP II	C	W	X	Y	Z	C	C	C_0	C_0	C_0	C_0	C
GROUP I	D	D_0	D_0	D_0	D_0	D	D	W	X	Y	Z	D
GROUP II	D	W	X	Y	Z	D	D	D_0	D_0	D_0	D_0	D

Figure 2. Design of Experiments I and II.
In Experiment I all test materials were vocoded; all familiarization materials were unprocessed speech (continuous).
In Experiment II all test and familiarization materials were vocoded.

Table 1. DRT Total Scores for Various Conditions

			Experimental Condition					Control Condition				
			Pre-Test		Post-Test			Pre-Test		Post-Test		
	Listeners	Speaker	$\bar{x}$	σ	$\bar{x}$	σ	Δx	$\bar{x}$	σ	$\bar{x}$	σ	Δx
Unprocessed Familiarization Speech	Group I	A	70.1	1.99	71.2	1.83	1.1	72.8	1.82	74.3	0.67	+1.5
	Group II	A	62.6	2.14	61.9	3.32	-0.7	60.5	3.14	63.2	2.42	+2.7
	Group I	B	77.3	2.53	79.5	0.75	+2.5	78.9	1.95	78.5	1.28	-0.4
	Group II	B	71.1	3.43	68.5	2.63	-2.6	67.5	4.43	70.3	1.59	+2.8
Average							+0.1					1.6
Vocoded Familiarization Speech	Group I	C	63.7	1.91	65.6	0.73	1.9	65.6	0.55	60.7	3.36	-4.9
	Group II	C	58.9	1.83	59.3	1.70	0.4	57.4	0.76	57.3	1.93	-0.1
	Group I	D	64.2	2.92	69.6	0.66	5.4	69.2	1.68	65.7	2.83	-4.5
	Group II	D	60.0	2.58	55.5	3.12	-4.5	52.9	2.11	56.8	3.35	+3.9
Average							+0.8					-1.4

C. CONCLUSIONS

The results of the experiments described here should by no means be construed to rule out the importance of listener-familiarity-with-speaker in actual communications situations. The DRT is designed to test only for the transmission of various speech features essential to the recognition of speech sounds. It does not evaluate the adequacy with which the listener interprets these features. By eliminating one possibility, however, the present research serves to elucidate the manner in which the listener-familiarity-with-the-speaker contributes to speech intelligibility.

CHAPTER II

THE IDENTIFICATION OF PERCEIVED ACOUSTIC TRAITS

The results of a series of investigations by Voiers provided a point of departure for the present line of research. In an exploratory attempt to identify the "major dimensions of perceived variation" among voices, Voiers (1964) had a group of listeners use a 49-item semantic differential rating form (Fig. 3) to describe their perceptions of the distinguishing characteristics of each of the sixteen male voices.

A factor analysis of average ratings received by each speaker revealed that the "speaker components" of variance for the forty-nine items could be accounted for in terms of only four underlying, orthogonal factors or dimensions of perceived variation among voices. It appeared, moreover, that adequate representation of interspeaker variation in these dimensions could be obtained from listener response data for a limited number of judiciously selected items out of the original forty-nine. On the basis of these results, a number of seemingly redundant items, i.e., rating dimensions, were discarded or combined in various ways.

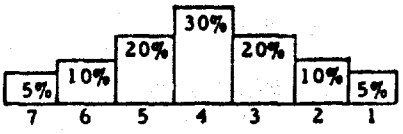
Subsequently, a number of abbreviated rating forms composed of factorially pure "tag items," plus various previously untried items, were tested experimentally. Here, it was expected that the tag items from the original form would continue to measure the factors with which they were previously associated, and that some of the new items would serve to reveal additional dimensions of perceived variation among voices. However, neither of these expectations was consistently realized. Some tag items, for which listener responses were originally uncorrelated, occasionally yielded correlated responses while experimental items failed

Speaker _____

SPEECH RATING FORM IIA

LOUD	7 6 5 4 3 2 1	SOFT
HEAVY	7 6 5 4 3 2 1	LIGHT
BEAUTIFUL	7 6 5 4 3 2 1	UGLY
CLEAR	7 6 5 4 3 2 1	HAZY
FRIENDLY	7 6 5 4 3 2 1	BELLIGERENT
RELAXED	7 6 5 4 3 2 1	TENSE
FAMILIAR	7 6 5 4 3 2 1	STRANGE
COLORFUL	7 6 5 4 3 2 1	COLORLESS
WARM	7 6 5 4 3 2 1	COOL
RISING	7 6 5 4 3 2 1	FALLING
LARGE	7 6 5 4 3 2 1	SMALL
PLEASANT	7 6 5 4 3 2 1	UNPLEASANT
DEFINITE	7 6 5 4 3 2 1	UNCERTAIN
GENTLE	7 6 5 4 3 2 1	VIOLENT
LOOSE	7 6 5 4 3 2 1	TIGHT
WET	7 6 5 4 3 2 1	DRY
RICH	7 6 5 4 3 2 1	THIN
DULL	7 6 5 4 3 2 1	SHARP
MASCULINE	7 6 5 4 3 2 1	FEMININE
RUMBLING	7 6 5 4 3 2 1	WHINING
GOOD	7 6 5 4 3 2 1	BAD
EVEN	7 6 5 4 3 2 1	UNEVEN
CALMING	7 6 5 4 3 2 1	EXCITING
SOFT	7 6 5 4 3 2 1	HARD
ACTIVE	7 6 5 4 3 2 1	PASSIVE
HAPPY	7 6 5 4 3 2 1	SAD
RUGGED	7 6 5 4 3 2 1	DELICATE
FAST	7 6 5 4 3 2 1	SLOW
WIDE	7 6 5 4 3 2 1	NARROW
PLEASING	7 6 5 4 3 2 1	ANNOYING
CONCENTRATED	7 6 5 4 3 2 1	DIFFUSED
REASSURING	7 6 5 4 3 2 1	DISTURBING
SERENE	7 6 5 4 3 2 1	AGITATED
STEADY	7 6 5 4 3 2 1	FLUTTERING
DELIBERATE	7 6 5 4 3 2 1	CARELESS
GLIDING	7 6 5 4 3 2 1	SCRAPING
EASY	7 6 5 4 3 2 1	LABORED
LOW	7 6 5 4 3 2 1	HIGH
SMOOTH	7 6 5 4 3 2 1	ROUGH
OBVIOUS	7 6 5 4 3 2 1	SUBTLE

SIMPLE	7 6 5 4 3 2 1	COMPLEX
MILD	7 6 5 4 3 2 1	INTENSE
NATIVE	7 6 5 4 3 2 1	FOREIGN
FULL	7 6 5 4 3 2 1	EMPTY
POWERFUL	7 6 5 4 3 2 1	WEAK
DEEP	7 6 5 4 3 2 1	SHALLOW
BUSY	7 6 5 4 3 2 1	RESTING
REPEATED	7 6 5 4 3 2 1	VARIED
CLEAN	7 6 5 4 3 2 1	DIRTY



Expected Long-Run Distribution of Ratings

Additional Comments: (Other words or phrases which might be used to characterize the sound of this speaker's voice.)

_____

_____

_____

_____

_____

Do Not Write in This Block

Factor I _____

Factor II _____

Factor III _____

Factor IV _____

Factor V _____

Factor VI _____

Factor VII _____

_____

_____

_____

_____

_____

Figure 3. Semantic Differential Rating Form (Voiers, 1964).

generally to reveal any new dimensions of perceived variation among voices. The results of these experiments did, however, provide several important methodological insights.

Among other things, it appeared that concentration of certain critical items into a short rating form — such that listeners were required to evaluate several different voice characteristics in quick succession — was conducive to the "halo effect." This effect has long been familiar to industrial and military psychologists. It is exemplified by the military commander who, having evaluated a junior officer favorably on "military bearing," also gives him inordinately high ratings on intelligence, competence, etc.

The traditionally accepted method of counteracting halo is to require raters to rate all stimuli (e.g., voices) on one rating dimension at a time rather than one stimulus at a time on all rating dimensions. However, such a procedure is quite impractical in the present application, particularly in view of indications that listeners find it difficult to retain a stable conception of a given rating dimension from one instance to the next.

Moreover, the "one-dimension-at-a-time" procedure is inherently vulnerable to systematic "errors" associated with "adaptation level" changes. Having assigned an extreme value to one stimulus (voice) a rater tends, generally, to alter his perception (or conception) of the "neutral point" such that immediately subsequent stimuli are rated somewhat nearer the opposite extreme than would otherwise be the case. Frequent instances of a significant mean square for "speakers x order of speaker presentation" provide support for this hypothesis. Other things equal, a procedure which provides greatest temporal separation

between successively presented voices will be least conducive to stimulus-induced shifts in adaptation level.

Hence, in the interest of maintaining the stability of the listener's response characteristics, the one-voice-at-a-time procedure was retained and other steps were taken to minimize extraneous influences. These steps included the use of filler items and a format designed to maximize connotative dissimilarity between adjacent items on the rating form. In addition, items found to yield highly correlated responses, e.g., masculine-feminine, heavy-light, were combined. This served effectively to reduce random response variation due to "semantic jitter" (for example, the combining of the item "high-low" with the item "whining-rumbling" operates to minimize ambiguity as to the possible denotations of high-low and also, perhaps, of "whining-rumbling"). Finally, in a further attempt to minimize adaptation level changes, a "neutral voice" was introduced as a standard for evaluation of each experimental voice. The absence of significant interactions of "speakers" with "speaker order" in subsequent studies attests to the efficacy of this innovation. Thus far, however, none of these various modifications of the voice-rating procedure has resulted in the consistent emergence of more than four orthogonal dimensions of the "speaker component" of variance in voice ratings. The research described in the following section represents a recent effort to increase the sensitivity of the voice-rating method and then to identify additional elementary dimensions of perceived variation among voices.

A. A SEARCH FOR ADDITIONAL PERCEIVED ACOUSTIC TRAITS

While previous attempts (Voiers, 1960, 1964; Voiers et al., 1965) to isolate more than four stable, orthogonal dimensions of perceived variability among voices have not been successful, several possibilities have yet to be fully explored. One such possibility is offered by the use of other than bipolar rating scales and is the subject of the present investigation.

Thus far only rating scales defined by antonymous adjective pairs, e.g., "high-low","large-small," have been employed in this series of voice-rating studies. However, this convention, adopted primarily in the interest of uniformity, may conceivably have resulted in the exclusion of certain potentially useful rating dimensions. For example, such adjectives as "tinny," "hissing," and "thumping" would seem applicable to complex sounds in general and to voices in particular, although some ambiguity may exist concerning the antonyms of these and other seemingly applicable terms. In any event, the antonyms of many such terms are not likely to be matters of common knowledge or of general agreement among the listener population at large. The present study was designed to evaluate a number of such "monopolar" terms from the standpoint of potential for the communication of speaker identity information and, at the same time, to determine their antonyms empirically.

Methods and Materials

Listeners. A crew of twelve male, high school seniors served as listeners for this investigation. All members had from ten to twelve hours of experience as listeners for intelligibility tests and speech-quality evaluation, but none had previous experience in voice-rating experiments.

Stimulus Materials. The basic speech-stimulus materials were those used in several previous studies with the voice-rating methods (Voiers et al., 1965) and are described in detail elsewhere. Briefly, they consisted of recordings of speech of 24 adult males whose ages ranged from 23 to 55 years.

Voice-Rating Forms. Two voice-rating forms were used by the listeners to describe their perceptions of the voices. The first of these was Voice-Rating Form IIIA (Fig. 4) used in a number of previous studies. The second was an experimental form (Fig. 5) composed entirely of "monopolar" rating scales.

Experimental Design. The experiment consisted of four voice-rating sessions, each involving the presentation of all twenty-four voices. In the first session, listeners rated the voices as presented in one order (A), using Voice-Rating Form IIIA (Fig. 4). In the second session, the voices were presented in the reverse order (B). The listeners used the Experimental Form (Fig. 5) in this session. In the third session, speakers were again presented in Order B and rated with Form IIIB. In the final session, the speakers were presented in Order A and rated with the experimental rating form. For each of the total of 32 rating dimensions, or items, there were thus two "trials."

Procedure. Following instructions to acquaint them with the general nature of their task, the listeners were allowed to rate one voice (that of the announcer) to get practice with the rating procedure and to stabilize their "images" of the neutral voice. Just prior to the beginning of the experiment proper they were allowed to hear brief samples of each experimental voice in rapid succession. Each speaker identified

Speaker_____

SPEAKER RATING FORM III A

Listener_____

Steady Stable	☐☐☐☐☐☐☐☐☐	Fluttering Unstable
Colorless Monotonous	☐☐☐☐☐☐☐☐☐	Colorful Dynamic
Foreign Rare	☐☐☐☐☐☐☐☐☐	Native Common
Rumbling Low	☐☐☐☐☐☐☐☐☐	Whining High
Unpleasant Annoying	☐☐☐☐☐☐☐☐☐	Pleasant Pleasing
Gradual Rounded	☐☐☐☐☐☐☐☐☐	Abrupt Jagged
Loud Intense	☐☐☐☐☐☐☐☐☐	Soft Mild
Passive Dragging	☐☐☐☐☐☐☐☐☐	Active Brisk
Excited Agitated	☐☐☐☐☐☐☐☐☐	Calm Serene
Gliding Smooth	☐☐☐☐☐☐☐☐☐	Scraping Rough
Fast Busy	☐☐☐☐☐☐☐☐☐	Slow Resting
Beautiful Clean	☐☐☐☐☐☐☐☐☐	Ugly Dirty
Feminine Light	☐☐☐☐☐☐☐☐☐	Masculine Heavy
Familiar Usual	☐☐☐☐☐☐☐☐☐	Strange Unusual
Clear Definite	☐☐☐☐☐☐☐☐☐	Hazy Uncertain
Uneven Irregular	☐☐☐☐☐☐☐☐☐	Even Regular

Figure 4. Standard Voice-Rating Form.

Speaker_____

Listener_____

VOICE RATING FORM XI

				THUMPING-RESONANT						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				BREATHY-HISSING						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				TWANGY-TINNY						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				SOLID-CLOSED						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				CLICKING-SMACKING						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				SQUEEKING-CHIPPING						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				BABBLING-GURGLING						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				SNAPPING-CRACKLING						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				THUDDING-DULL						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				ABRUPT-CLIPPED						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				THROATY-RICH						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				HOOTING-BLEATING						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				RUSHING-GUSHING						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				BUZZING-DRONING						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				HOLLOW-OPEN						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree
				TIGHT-TENSE						
Very low degree	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very high degree

Mark "X" in one cell of each row to show the degree to which each indicated quality is present in the voice you are now hearing as compared to the voice of the announcer.

Figure 5. Experimental Voice-Rating Form.

himself by a number — e.g., "This is speaker number 2." All stimuli were presented over PDR-8 headphones.

Results and Discussion

Analysis of variance was employed to provide an indication of the degree to which listeners effectively discriminated among speakers on each of the thirty-two rating dimensions.

The results of these analyses are summarized in Table 2. Also presented for purposes of comparison are results from a previous experiment (Voiers et al., 1965) in which a different, larger group (32) of listeners used Voice Rating Form III to rate the same sample of speakers. The F-ratios for the previous experiment are adjusted to compensate for differences in the sizes of the listener crews; i.e., the F-ratios shown for the earlier experiment are estimates of the values which would have been obtained with a crew of twelve listeners.[1]

At the outset, it is apparent that listeners in the present experiment did not, in general, respond as discriminately to voices as did the listeners of the earlier experiment. The values of the F-ratios obtained for the speaker effect on the present experiment are, with few exceptions, smaller than their counterparts from the earlier experiment.

Any of several situational factors may have contributed in some degree to the reduced sensitivity of the present experiment. Though

---

[1]Given F(n) as the ratio obtained for a sample of n raters, the ratio, F(m) which would be obtained with a sample of m listeners can be estimated as:

$$F'(m) = \frac{m}{n}\left[F(n)-1 \right] + 1$$

Table 2. Summary of Results of Analysis of Variance of Voice Ratings

	Item	Present Study F-Ratio for Speaker Effect*	Previous Study Est. F-Ratio for Speaker Effect
1	Fluttering, etc.	1.56	4.36
2	Colorful	3.34	5.35
3	Foreign	2.02	3.16
4	Rumbling	6.68	19.12
5	Unpleasant	2.07	4.09
6	Abrupt	2.67	5.93
7	Loud	3.40	6.11
8	Active	5.98	9.80
9	Excited	5.38	11.73
10	Scraping	1.71	3.57
11	Fast	10.55	12.14
12	Beautiful	1.89	2.38
13	Masculine	11.85	20.80
14	Strange	2.21	3.38
15	Clear	3.55	2.72
16	Uneven	1.84	3.48
17	Thumping	3.34	
18	Breathy	2.85	
19	Twangy	2.63	
20	Solid	1.71	
21	Clicking	3.59	
22	Squeaking	2.12	
23	Babbling	——	
24	Snapping	1.03	
25	Thudding	2.31	
26	Abrupt	2.61	
27	Threaty	3.10	
28	Hooting	2.14	
29	Rushing	1.05	
30	Buzzing	——	
31	Hollow	——	
32	Tight	1.72	

*$F = \frac{\text{M.S. Speakers}}{\text{M.S. Speakers \& Listeners}}$

contributions of these various factors cannot be isolated at this point, we may at least note some of the possibilities.

One factor may have been the educational level of the listeners. In the present experiment, the listeners were high school students with little experience in speech evaluation experiments and no experience in voice-rating experiments, specifically. In the earlier experiment, the listeners were college students with 15-20 hours of experience as listeners in various types of speech-evaluation experiments, including voice-rating experiments. In the present experiment, listeners were not individually isolated, but were seated in chairs grouped around the tape recorder in the center of a large room. In the earlier experiment, listeners were tested in specially constructed booths, each of which housed two listeners in visual and acoustical isolation from the remaining listeners and from other possible sources of distractions. Whatever the factors involved, the depressed sensitivity of the present experiment serves to stress the need for extremely rigid control of situational parameters in experiments utilizing the voice-rating method.

A further examination of Table 2 reveals a general tendency for items from Form XI (Fig. 5) to yield smaller F-ratios than items from Form IIIA (Fig. 4). While this possibly implies a superiority of bipolar to monopolar scales, a more likely explanation is found in the select character of the items comprising Form IIIA. All had survived a series of screenings on the basis of demonstrated validity while, at the outset, the items comprising Form XI were of undetermined validity. In any case, nine of the sixteen experimental items yielded F-ratios for the speaker effect which are significant at $P \geq .01$, ($F = 1.89$), which is to say that

a mean of 12 listeners' ratings of a voice on one of these items carries a statistically significant amount of speaker identity information. Statistical significance, per se, does not, however, ensure practical validity.

Given that

$$"C" = 1/2 \log_2 F(n)$$

is a measure of the capacity of an average rating (i.e., of a n ratings) for speaker identity information, it becomes apparent that none of the items (i.e., as averaged for 12 listeners) served to "transmit" as much as one bit of speaker identity information in the present experiment.

Assuming that the informational capacities of the various items can be increased by more stringent control of situational variables, there remains, in any case, the issue of whether any of them tap unique aspects of listeners' perceptions of voices. To resolve this issue, speaker averages for the set of 32 items were subjected to factor analysis by means of the method of principal components.

By several criteria it was apparent that no more than three orthogonal dimensions were required to account for the systematic component (i.e., speaker component) of variance for the entire set of thirty-two items. The results obtained following rotation of the factor frame to an arbitrary criterion of simple structure are shown in Table 3.

Appropriate labels for the three factors are apparent from the configuration of factor loadings. All three, moreover, find precedence in the results of previous voice-rating experiments. Factor I evidently represents the perceptual dimension, Pitch-Magnitude, which has appeared in all previous voice-rating studies. It is identified by high loadings on such items as:

TABLE 3. FINAL FACTOR LOADINGS FOR THIRTY-TWO RATING DIMENSIONS

	I	II	III	h^2
1 Fluttering etc.	.571	.768	.432	.626
2 Colorful	.403	.440	.111	.337
3 Foreign	-.044	-.657	.194	.471
4 Rumbling	-.950	-.054	-.038	.910
5 Unpleasant	-.076	-.630	.495	.648
6 Abrupt	.235	-.173	.823	.763
7 Loud	.118	.349	.820	.810
8 Active	.647	.472	.427	.825
9 Excited	.606	.320	.656	.901
10 Scraping	.121	-.152	.758	.707
11 Fast	.691	.303	.624	.848
12 Beautiful	.293	.654	-.167	.543
13 Masculine	-.984	.060	-.030	.972
14 Strange	-.120	-.757	.184	.621
15 Clear	.081	.840	.257	.780
16 Uneven	.608	-.439	.303	.656
17 Thumping	-.862	-.065	.321	.850
18 Breathy	.708	-.448	.043	.706
19 Twangy	.884	-.060	.165	.812
20 Solid	.844	.232	.554	.768
21 Clicking	.583	-.266	.349	.532
22 Squeaking	.770	-.195	.207	.676
23 Babbling	.427	-.079	.253	.252
24 Snapping	.430	.044	.292	.272
25 Thudding	-.875	-.111	-.239	.835
26 Abrupt	.543	.386	.336	.558
27 Threaty	-.870	.019	.026	.758
28 Hooting	.834	-.045	.093	.705
29 Rushing	.392	-.043	.382	.302
30 Buzzing	-.432	-.227	.155	.346
31 Hollow	.208	.106	-.067	.059
32 Tight	.551	.063	.377	.457

Rumbling ----------Whining
Low ----------High

or

Masculine -------- Feminine
Heavy -------- Light

Factor II represents the Loudness-Roughness dimension of perceived variability among voices, while Factor III is the Clarity-Beauty dimension. Both of these factors have their counterparts in all previous voice-rating studies.

Conspicuously absent is the Animation-Rate dimension which has emerged on several previous occasions. Here, as in the past, experiments, judgments of "fast," "active," etc., are highly correlated with judgments of "whining-high" and "loud-rough." Thus, rating variance which on some occasions requires a separate dimension, Animation-Rate, can be accounted for here in terms of the Pitch-Magnitude and Loudness-Roughness dimensions. Also absent is the Normality dimension which has emerged on several occasions.

While none of the experimental items treated in this study appears to be of value in relation to new dimensions of perceived variability among voices, several offer possibilities for improved measurement with respect to previously identified dimensions. This can perhaps be more easily shown by means of graphic presentation.

Figure 6 shows the projections of the various rating dimensions in the I-II plane of the three-factor space. From the figure it is apparent that several of the experimental items (17 through 32) can provide measures of high "factorial purity" in the Pitch-Magnitude dimension. It is also apparent how various of the monopolar items of Form XI can be combined to define a bipolar rating dimension. For example, items 17 and 19 appear to represent opposite extremes of a semantic continuum

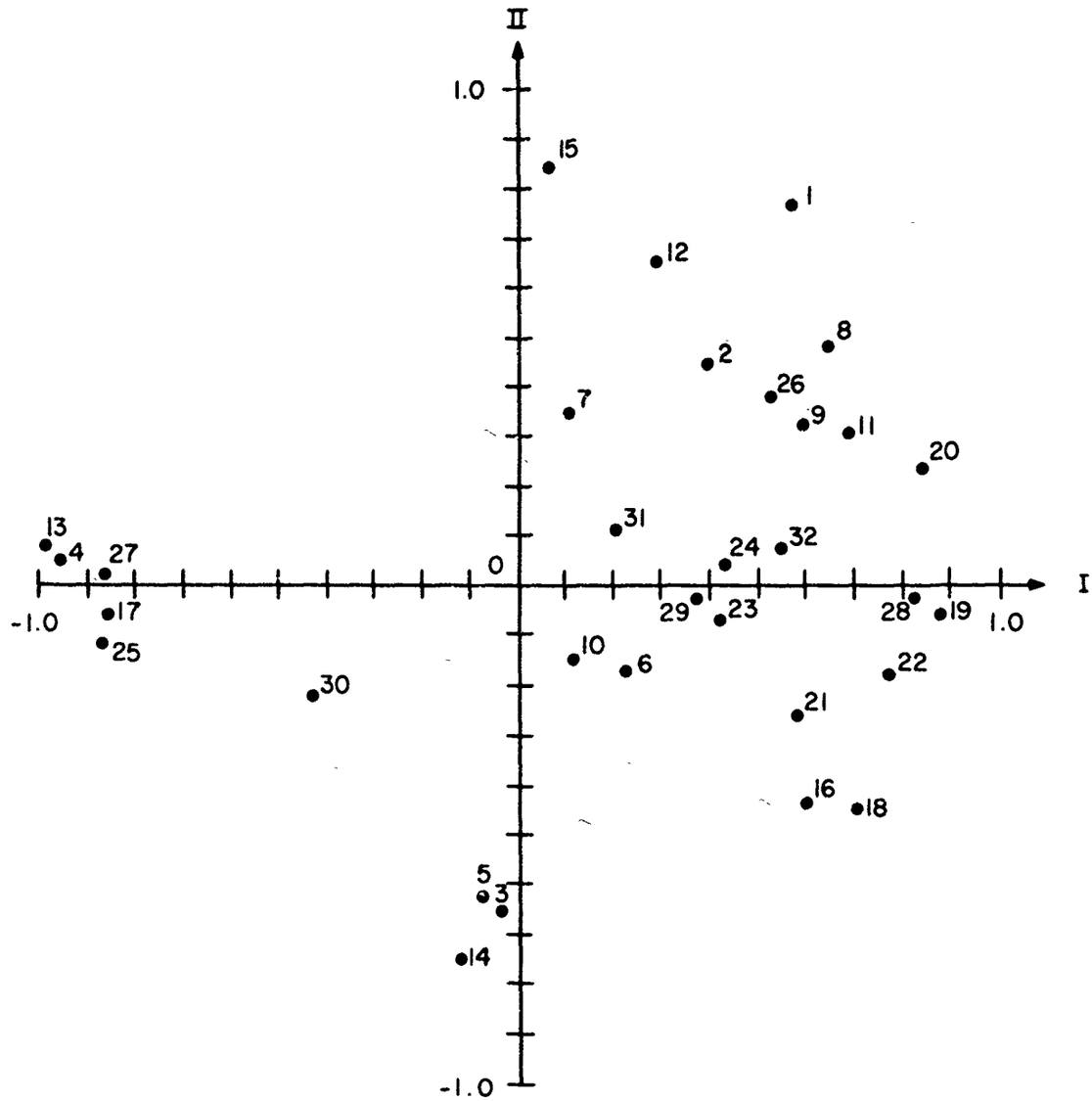


Figure 6. Factor Loadings of Thirty-Two Rating Dimensions in the I-II Plane of the Perceived Acoustic Trait Space.

which happens, incidentally, to correspond quite closely to the Pitch-Magnitude dimension. These and similar findings with respect to other experimental items provide a tentative basis for the refinement of the presently used (Form IIIA) voice-rating method. However, confirmation of these findings by additional research is clearly desirable before any modification of the present voice-rating procedures and materials is undertaken.

B. CONCLUSIONS

From the results of research conducted thus far we are led to conclude that the voice-rating method is presently adequate to evaluate listener reception of four perceived acoustic traits. One possibility is that refinement of the method, as a research tool, will permit the discovery of additional PAT's. Another possibility, however, is that the limited number of traits thus far identified is a function not of the limitations of the method but rather a fundamental limitation of human taxonomic capacity. Only further research can resolve this issue.

CHAPTER III

THE PHYSICAL CORRELATES OF PERCEIVED ACOUSTIC TRAITS

A. THEORETICAL CONSIDERATIONS

The fact that listeners are able to order voices with some degree of consistency on such continua as "high-low," "large-small," or even "red-green" has a number of potentially significant implications concerning the physical acoustic bases of individual differences in speech. Even where direct evidence regarding the nature of these bases is lacking, behavioral data can yield valuable insights regarding several aspects of the problem. For example, the fact that listeners are able to rate voices on the continuum "fast-slow" in a consistent manner suggests that the independently-measurable variable, speech rate, is among the more important physical voice parameters. Similarly, the fact that listeners are able to rate voices on the continuum "high-low" with a high degree of consistency strongly suggests that some function of glottal pulse rate is a major parameter of interindividual variation. As it turns out, neither of these two psychophysical correlations is perfect — judged rate of speech is dependent upon other physical variables in addition to physical speech rate, and judgments with respect to "high-low" are not perfectly predictable from measures of "average pitch frequency"— but, the results of voice-rating experiments can serve, in any case, to narrow the field of search for the elementary physical parameters of individual differences in speech.

Even in instances where the physical implications of listener response are perhaps trivial (it is admittedly unlikely that pitch frequency would have escaped attention but for the results of voice-rating studies)

or quite obscure (what, for example, are the possible physical bases of listener consistency in rating voices on the continuum "rare-common"?), voice-rating data may provide other kinds of information as to the physical basis of acoustic individuality.

Given knowledge of the number of independent "ways" in which listeners can classify voices (i.e., the number of orthogonal dimensions required to specify completely the average listener's perception of a voice), we may in turn infer the number of independent perceptually-significant physical parameters, since for each perceptual dimension there necessarily exists a corresponding physical acoustic dimension. Thus, the knowledge that listeners are able to discriminate voices with respect to the semantic continuum "clear-unclear," may serve to narrow the field of search for one physical parameter of interindividual variation in speech. The fact that listener ratings on the continuum "clear-unclear" are uncorrelated with ratings on the continuum "fast-slow" may serve even further to narrow the field of search inasmuch as it implies that the physical acoustic correlate of perceived clarity is uncorrelated with the physical correlate of perceived speech rate. Alternatively, the fact that voice ratings on the continuum "fast-slow" tend generally to be correlated with judgments of "loud-soft" is prima facie suggestive of a theoretically significant relation between two physical voice variables, i.e., speech rate and speech intensity.

In light of the foregoing it would appear that research on the perception of individual differences in speech can contribute to the solution of numerous practical problems of voice communications, such as the problem of automatic voice recognition and that of system evaluation from the stand-

point of speaker recognizability. Before these problems can be adequately solved, however, several fundamental issues concerning the nature of the perceptual processes underlying voice recognition in general and voice-rating behavior in particular must be resolved.

Assuming that the perceived characteristics of voices are related in a fixed, immutable manner to some underlying set of physical acoustic traits,-i.e., that changes in the physical structure of the speech signal will affect its perceived characteristics only according as they obscure, enhance or distort the "fixed" physical correlates of these characteristics — problems of automatic recognition and system evaluation can be approached in a relatively straightforward manner.

There are, however, some grounds for hypothesizing that the manner in which a perceived acoustic trait relates to a physical voice parameter is "relationally determined" — that it depends upon the statistical structure of the voice population or source, more particularly, upon the listener's perception of that structure. Specifically, we might hypothesize that:

> Given one or more dimensions of response to characterize his perceptions of the distinguishing characteristics of individual voices, a listener will tend to allocate the available response dimensions to various voice characteristics in a manner which tends to maximize the speaker-identity-information content of his responses.

Various aspects of this hypothesis find support in the literature on human information processing phenomena; Quastler (1956) has formalized several of them in his "daisy" model of the human sensory information processing system. This model involves manifold channels (each potentially assignable to a single sensory-perceptual dimension) converging on a common

channel which has an information capacity less than the sum of the capacities of the "feeder" channels. The common channel space allocated to any one feeder channel is limited in any case, but depends in a given instance upon the distribution of information loads upon the remaining channels. Thus, stimulus impoverishment in a particular dimension may act to decrease the total stimulus information transmitted to the observer but to increase the information transmitted concerning one or more of the remaining stimulus dimensions. The results described by Pollack (1963) for cases involving separate and combined pitch and loudness judgments are typical of the experimental support for the "daisy" model. A priori, analogous results might be expected in the case of voice-rating experiments (i.e., where some form of speech degradation effectively reduces the dimensionality of the speech stimulus) and some results, described by Voiers et al. (1965) for vocoded speech, appear consistent with these expectations. However, the voice-rating situation is distinguished in several important ways from the situations typical of the "classical" studies of human channel capacity, and some elaboration of the "daisy" model may be required to cover this case.

In most of the early experiments on human information processing, there has existed little or no ambiguity concerning the stimulus dimension to which a particular response dimension is most appropriately associated. Experiments involving pitch ratings of pure tones are typical in this respect. In the case of voice ratings or judgments, however, the critical stimulus parameters are, per se, likely to be relatively obscure or ill-defined. In turn, there may be little basis a priori for listener consensus regarding which response dimension is most appropriate to each parameter. Accordingly, we might hypothesize that:

> Lacking effective instruction as to the most appropriate allocation of his response repertoire to a particular set of physical voice parameters, a listener will tend, generally, to infer a solution on the basis of his experience in the experimental situation and, in particular, his conception of the stimulus population or source involved.

For example, given stimuli which vary substantially with respect to the acoustical correlates of the sensory quality, pitch, the typical listener's "normal" reaction will be to utilize such rating dimensions as "low," "high," and "masculine-feminine" to indicate his perceptions of this sensory quality. Where, however, interindividual variation in the stimulus correlates of pitch are reduced or obscured in one way or another, the listener's reaction — according to hypothesis — will be to utilize the "low-masculine — high-feminine" channel to carry information with regard to other stimulus properties where this can be accomplished in any "reasonable" manner. To the extent that such stimulus-response realignments — "qualitative shifts" in the psychophysics of voice perception — can occur, changes in the information content of voice ratings under various transmission conditions must necessarily be interpreted with some caution. More generally, statements concerning the physical correlate of the various perceived acoustic traits may require some qualification.

The experiment described below was designed to provide some clarification of this issue and, more generally, the issue of the physical basis of voice recognition.

B. A STUDY OF THE PHYSICAL BASIS OF PERCEIVED ACOUSTICAL TRAITS

A factor analysis of twenty-three voice variables was performed in an attempt to elucidate the psychophysics of voice perception. Sixteen of

these are perceptual variables derived from the listener's ratings of voices. However, they involve only four rating dimensions, each representing one of four elementary perceived acoustic traits: I. Pitch-Magnitude, II. Loudness-Roughness, III. Animation-Rate, and IV. Clarity-Beauty. Data relating to each dimension are included for four different experimental conditions, each involving a different form of speech processing. These were unprocessed speech, digitally vocoded speech, severely low-passed speech and severely high-passed speech.

The seven physical voice variables treated in the analysis include an indicant of the speaker's natural or average pitch frequency, a measure of his characteristic rate of speech and of his "spectral characteristics," a measure of the level at which his speech was presented to the listeners and a measure of his inherent speech level. These variables are described in detail in Table 4.

Methods and Materials

The voice-rating methods, stimulus materials, etc. are described in detail elsewhere (Voiers et al., 1965).[1] A "relative rating" procedure was employed to collect all of the voice-rating data used here. The basic stimulus materials were provided by the tape-recorded voice samples used in previous research involving the relative rating method.[2] In all instances listeners used Voice Rating Form IIIA to describe their perceptions of the 24 voices.

Data for the variables described in Table 4 were subjected to a factor analysis by the method of principal components. The results of this analysis are presented below.

---

[1] Voice-rating data for the case of unprocessed speech were provided by the "First Normative Study" of Voiers et al. (1965).

[2] Ibid.

Table 4. Physical and Perceived Voice Variables Subjected to Factor Analysis

1. (PAT 1_u) Speaker values on PAT 1 (Pitch-Magnitude) based on averages of 32 listeners' ratings for two items:

Rumbling---------Whining
Low----------High

Masculine--------Feminine
Heavy----------Light

for the case of unprocessed speech.

2. (PAT 2_u) Speaker values on PAT 2 (Loudness-Roughness) based on averages of 32 listeners' ratings for two items:

Loud----------Soft
Intense----------Mild

Scraping---------Gliding
Rough----------Smooth

for the case of unprocessed speech.

3. (PAT 3_u) Speaker values on PAT 3 (Animation-Rate) based on averages of 32 listeners' ratings for two items:

Active----------Passive
Brisk----------Dragging

Fast----------Slow
Busy----------Resting

for the case of unprocessed speech.

4. (PAT 4_u) Speaker values on PAT 4 (Clarity-Beauty) based on averages of 32 listeners' ratings for two items:

Clear----------Hazy
Definite---------Uncertain

Beautiful--------Ugly
Clean----------Dirty

for the case of unprocessed speech.

5. (PAT 1_v) Speaker values on PAT 1 based on averages of 8 listeners' ratings of voice samples processed by a 2400-bit digital vocoder.

Table 4 (cont.)

6.	(PAT 2_v)	Speaker values on PAT 2 based on averages of 8 listeners' ratings of voice samples processed by a 2400-bit digital vocoder.
7.	(PAT 3_v)	Speaker values on PAT 3 based on averages of 8 listeners' ratings of voice samples processed by a 2400-bit digital vocoder.
8.	(PAT 4_v)	Speaker values on PAT 4 based on averages of 8 listeners' ratings of voice samples processed by a 2400-bit digital vocoder.
9.	(PAT 1_{lp})	Speaker values on PAT 1 (8 listeners) for the case of speech low-passed at 500 Hz.
10.	(PAT 2_{lp})	Speaker values on PAT 2 (8 listeners) for the case of speech low-passed at 500 Hz.
11.	(PAT 3_{lp})	Speaker values on PAT 3 (8 listeners) for the case of speech low-passed at 500 Hz.
12.	(PAT 4_{lp})	Speaker values on PAT 4 (8 listeners) for the case of speech low-passed at 500 Hz.
13.	(PAT 1_{hp})	Speaker values on PAT 1 (8 listeners) for the case of speech high-passed at 2000 Hz.
14.	(PAT 2_{hp})	Speaker values on PAT 2 (6 listeners) for the case of speech high-passed at 2000 Hz.
15.	(PAT 3_{hp})	Speaker values on PAT 3 (8 listeners) for the case of speech high-passed at 2000 Hz.
16.	(PAT 4_{hp})	Speaker values on PAT 4 (8 listeners) for the case of speech high-passed at 2000 Hz.
17.	(LST)	Frequency of speaker's lowest singing tone. (Frequency scale inverted for this analysis.)
18.	(ST)	Time required (exclusive of gaps) for speaker to enunciate the 16 sentences comprising voice-rating stimulus materials.
19.	(SS)	Indicant of "characteristic spectral slope" for speaker obtained by $SS = (L_{0-.5\ kc} + L_{.5-1\ kc}) - (L_{1-2\ kc} + L_{2-4\ kc})$, where L is the mean (dB re arbitrary level) for averaged "three highest peaks" under indicated frequency-pass condition.
20.	(SC)	Indicant of "characteristic spectral convexity" obtained by $SC = (L_{.5-1\ kc} + L_{1-2\ kc}) - (L_{0-.5\ kc} + L_{2-4\ kc})$.
21.	(J)	Indicant of "characteristic spectral jaggedness" obtained by $J = (OAL_{0-.5\ kc} + OAL_{1-2\ kc}) - (OAL_{.5-1\ kc} + OAL_{2-4\ kc})$.

Table 4 (cont.)

22.	(L)	Mean (dB re arbitrary level) for speaker of averaged (visual) three highest peaks as determined from General Radio Graphic Level Recorder.
23.	(IL)	Indicant of inherent speech level, obtained by IL (L - noise level) as determined with graphic level recorder.

Results and Discussion

Five orthogonal factors suffice to describe the systematic variation in the 23 variables under consideration, where the criterion of sufficiency is that the sum of the five largest eigenvalues is approximately equal to the summed coefficients of reliability for the 23 variables.[1]

The initial factorial axes were rotated to satisfy a varimax criterion of simple structure for the "tag" variables, PAT 1_u, PAT 2_u, PAT 3_u, and PAT 4_u.

Inspection of the resulting set of factor loadings revealed that an additional rotation (-26^0) in the IV - V plane would permit identification of factors IV and V with the physical variables inherent speech level and lowest singing tone. The pattern of loadings yielded by the rotation of axes is shown in Table 5.

Consider first the question of factor identification. Little difficulty is encountered in this connection. Factors I, II, and III are defined by perceptual variables, while factors IV and V are defined primarily by physical variables.

Factor I clearly represents the perceived acoustic trait, Pitch-Magnitude. All variables involving this PAT have loadings of .90 or higher. Factor II appears to represent the perceived acoustic trait, Loudness-Roughness, in light of the high loadings exhibited by all variables involving this PAT. It is noteworthy, however, that variables involving

---

[1]Coefficients of reliability are intraclass correlations in the case of the 16 perceptual variables. "Split-half" correlations adjusted by means of the Spearman Brown formula provided estimates of the reliabilities of the physical voice variables except LST. The communality of this variable obtained from a previous factor analysis was used as an estimate of its reliability.

Table 5. Final Factor Loadings of Twenty-Three Voice Variables

	I	II	III	IV	V		h^2	Reliability
1 (PAT 1_u)	0.95	0.09	-0.06	-0.02	0.02		0.91	0.98
2 (PAT 2_u)	0.10	0.94	-0.05	0.02	-0.01		0.91	0.93
3 (PAT 3_u)	-0.66	0.65	-0.08	-0.20	0.10		0.92	0.97
4 (PAT 4_u)	-0.02	-0.05	0.86	0.00	-0.00		0.75	0.84
5 (PAT 1_v)	0.92	0.18	-0.04	-0.11	0.10		0.88	0.94
6 (PAT 2_v)	-0.52	0.51	-0.12	0.15	0.36		0.70	0.62
7 (PAT 3_v)	-0.74	0.44	-0.15	-0.16	0.35		0.90	0.88
8 (PAT 4_v)	0.41	0.08	-0.44	-0.37	0.01		0.50	0.41
9 (PAT 1_{lp})	0.93	0.07	0.29	-0.14	-0.12		0.98	0.96
10 (PAT 2_{lp})	0.58	0.58	0.05	-0.10	0.03		0.69	0.74
11 (PAT 3_{lp})	-0.61	0.52	-0.23	-0.25	0.31		0.86	0.86
12 (PAT 4_{lp})	0.39	-0.27	0.63	0.17	-0.22		0.70	0.71
13 (PAT 1_{hp})	0.89	0.19	0.10	-0.23	0.00		0.89	0.93
14 (PAT 2_{hp})	-0.15	0.78	-0.13	-0.06	-0.08		0.66	0.70
15 (PAT 3_{hp})	-0.73	0.59	-0.14	-0.19	0.04		0.94	0.87
16 (PAT 4_{hp})	0.58	-0.48	0.42	-0.13	-0.01		0.79	0.71
17 (LST)	0.70	0.03	-0.13	0.02	0.52		0.78	0.76
18 (ST)	0.47	-0.14	-0.43	0.42	0.02		0.65	0.87
19 (SS)	0.37	-0.66	0.32	0.42	0.26		0.91	0.90
20 (SC)	-0.24	0.34	-0.50	0.40	-0.19		0.62	0.63
21 (SJ)	0.48	-0.49	0.15	-0.16	0.39		0.66	0.69
22 (L)	0.26	-0.30	0.20	0.45	0.20		0.44	0.69
23 (IL)	-0.27	0.10	-0.45	0.73	0.04		0.83	0.83
Σx^2	7.88	4.64	2.57	1.75	1.02	Σ	17.87	18.42

ratings of this PAT under conditions of degraded speech tend to have somewhat reduced loadings.

The high loadings obtained for the perceptual variables PAT 4_u and PAT 4_{1p} serve to identify factor III with the perceived acoustic trait Clarity-Beauty, though other variables associated with that PAT do not exhibit exceptionally high loadings on the factor in question.

No factor appears to be identified with the PAT, Animation-Rate, as has been the case on several previous occasions. Rather, values on this PAT — at least as estimated by the items "fast-slow," "busy-resting," etc. — appear to be predictable from ratings on PAT 1 and PAT 2. The loadings of PAT 3_u on factors I and II indicate a tendency for voices which are judged "high" and "loud" to be judged "fast" and "active" as well. The implications of this result will be discussed more fully at another point.

Factor IV is defined entirely by physical voice variables, and in particular IL (inherent voice level). No perceptual variable loads highly on this factor, though several of the physical variables, with high loadings here, also have relatively high loadings on factors defined by perceptual variables. Factor IV appears, in other words, to be a repository for the component of intensive variation among voices, which is of negligible perceptual significance.

Factor V is defined by a single physical variable, lowest singing tone, and appears to represent the perceptually insignificant component of variance in this physical voice parameter.

Psychophysics of Voice Perception

The results in Table 5 have various implications for the psychophysics of voice perception. Consider first the question of the "normal" physical correlates of the various perceived voice characteristics.

Pitch-Magnitude. In Table 5 it can be seen that the physical correlates of Pitch-Magnitude are quite well-defined and stable over a range of speech-transmission conditions. All variables involving this PAT have high loadings on factor I. Several physical variables also have substantial loadings on the factor associated with this PAT. Predictably, perhaps, the highest of these (.70) is found in the case of lowest singing tone (LST). Other variables relating to "spectral shape," SC and SJ, appear to contribute to the perception of Pitch-Magnitude, but it is of particular interest that speech time (ST) also appears to be a correlate of this PAT.

A more detailed examination of the correlations among Pitch-Magnitude and its various physical correlates provides some further insight as to the physical basis of this PAT. Table 6 shows, among others, the intercorrelations among the PAT 1 and the four physical voice variables which have the highest loadings on factor I. As is to be expected from the results in Table 5, LST, lowest singing tone, is the primary physical correlate of Pitch-Magnitude. Although the remaining physical variables exhibit some correlation with this perceptual variable, they tend to be intercorrelated in varying degrees.

In this connection it is of particular interest to know the basis of the somewhat unexpected correlation between Pitch-Magnitude and the physical variable, speech time. One might hypothesize that a slow rate of speech is per se conducive to the perception of "low pitch" and "large magnitude." However, a theoretical basis for such a hypothesis is somewhat difficult to find.

Table 6. Correlations Among Eleven Voice Variables*

	1	2	3	4	17	18	19	20	21	22	23
1 (PAT 1_u)		0.17	-0.59	-0.09	0.71	0.34	0.27	-0.15	0.36	0.22	-0.21
2 (PAT 2_u)			0.57	-0.08	0.07	0.02	-0.60	0.33	-0.41	-0.29	0.10
3 (PAT 3_u)				-0.09	-0.40	-0.41	-0.71	0.32	-0.59	-0.47	0.13
4 (PAT 4_u)					-0.07	-0.37	0.23	-0.38	0.10	0.22	-0.40
17 (LST)						0.42	0.31	-0.20	0.47	0.25	-0.11
18 (ST)							0.33	0.22	0.17	0.19	0.35
19 (SS)								-0.38	0.58	0.65	0.02
20 (SL)									-0.51	-0.15	0.67
21 (SJ)										0.30	-0.40
22 (L)											0.26
23 (IL)											

*See Table 4 for description of variables.

An alternative hypothesis is that speech rate, per se, is of no direct perceptual consequence here, but that it is correlated with other physical voice variables on which the perception of Pitch-Magnitude is directly dependent. In simpler terms, this is the hypothesis that the observed correlation between PAT 1 and speech time derives from a tendency for speakers with "low" voices to speak somewhat slower than average, rather than from any direct significance of speech time for the perception of Pitch-Magnitude.

Available data permit a test of this inference. This involves estimating the correlation which would exist between PAT 1 and speech time if LST and ST were uncorrelated. A "part correlation" of .06 is, in fact, consistent with the hypothesis that speech time, per se, has no implications for perception of the trait Pitch-Magnitude: that the observed correlation between speech time and PAT 1 is attributable primarily to the correlation between LST and ST. Further verification of this hypothesis, by experimentation is, of course, to be desired.

Loudness-Roughness. From Table 5 it is evident that this perceived acoustic trait has a negligible correlation with any gross aspect of intensive variation among voices. Neither the variable L, an indicant of voice intensity as heard by the listeners, nor IL, a measure of voice level relative to recorded noise level, exhibits significant correlation with judged Loudness-Roughness. However, two physical variables, themselves correlated, exhibit substantial correlation with this PAT. SS, an indicant of low-frequency energy relative to high-frequency energy, has a negative loading of .65 on factor II while SJ exhibits a negative loading of -.49 on the factor representing Loudness-Roughness. The implication of these results is that voices

having relatively large amounts of high-frequency energy tend to be judged loud and rough while voices for which low-frequency energy predominates tend to be judged soft and smooth.

It is also interesting to note that, while PAT 2_v and PAT 2_{lp} have substantially reduced loadings on factor II, PAT 2_{hp} (based on ratings of high-passed speech) has a relatively high loading on factor II. This would seem to provide further support of the hypothesis that the physical correlate of Loudness-Roughness resides in the upper region of the speech spectrum. In this connection, however, we might raise the question of whether the loading of SJ on factor II implies that spectral jaggedness *per se* is a correlate of Loudness-Roughness or whether the observed correlation of these two variables derives from the correlation between SJ and SS. Part correlation techniques can serve again to throw some light on the issue. If the communality of SS and SJ were non-existent, a correlation of only -.06 would be expected between PAT 2_u and SJ. This is consistent with the hypothesis that, as such, spectral jaggedness is not a condition of the perception, Loudness-Roughness.

Clarity-Beauty. Three physical voice variables have substantial loadings on the factor representing this perceived acoustic trait. SC, an indicant of spectral "convexity". has a loading of -.50 indicating that, *ceteris paribus*, voices having relatively high concentrations of energy outside of the 500-2000 Hz tend to be judged Clear-Beautiful.

A loading of 0.43 exhibited by ST, speech time, on this factor suggests that the voices of "slow talkers" tend to be judged more "Clear-Beautiful" than those of fast talkers.

A loading of -.45 exhibited by the variable, IL, implies that voices of inherently low intensity tend to be judged as more "Clear-Beautiful" than those of inherently high intensity. Since attempts were made to equate the levels (VU readings) of the voices as recorded and presented to the listeners, the significance of this result is somewhat difficult to assess. A priori it is conceivable that the experimenter over- or under-compensated for low-intensity voices in adjusting recording gain and thus introduced a spurious physical correlate of Clarity-Beauty. A correlation of .25 between L and IL indicates that, in fact, differences in inherent speech levels were somewhat inadequately compensated for in recording the voice samples. In any case, we may ask what the correlation between the PAT, Clarity-Beauty, and inherent speech level would be if recording gains were the same for all voices. The part correlation coefficient obtained by removing the effects of the correlation between IL and L is of interest here. The observed zero-order correlation between PAT 4_u and IL is -.40. Adjustment of this coefficient for the communality of IL and L increases its absolute value to -.47, which suggests that variations in recording gain tended to mask the "true" correlation (negative) between PAT 4 and the inherent speech level. It follows, in turn, that the correlation (positive, at least over the range of levels involved here) between PAT 4 and presentation level (L) can be masked to some extent by differences in inherent level. These conclusions might at first seem somewhat incompatible but some theoretical basis for them can be found.

Intuitively, it is reasonable that over some part of the range below the distortion point of the recording-playback system (and/or the

human auditory system) speech presentation level is positively related to its perceived clarity. Although the basis of a negative correlation between inherent speech level and the perception of clarity is not so apparent, the fact that inherent level, as evaluated here, is effectively inherent vowel level, raises the possibility that IL is tantamount to a negative measure of the ratio of consonant/vowel intensities which characterize a given voice. It is quite plausible, in turn, that voices characterized by a high consonant/vowel intensity ratio tend to be more intelligible and, in turn, to be perceived as more Clear-Beautiful than other voices. Needless to say, several points in this line of reasoning could bear experimental verification.

Animation-Rate. While PAT 3, Animation-Rate, does not emerge here as an independent dimension of perceived variability among voices, some interest attaches, nevertheless, to the issue of its physical acoustic basis. An examination of Table 6 reveals that this PAT exhibits substantial correlation with all but one of the physical variables treated here. Interestingly enough, speech time is not among the more important of these. Rather, spectral slope and spectral jaggedness appear to be the major correlates of the PAT, Animation-Rate, while several other physical variables correlate with this PAT as highly as speech time.

From Table 6 it can be seen that speech time has relatively low correlation with the other physical variables under consideration. This would suggest that various correlations between PAT 3 and these physical variables are not altogether artifacts of the sort we have encountered before.

This is to say that several of the physical variables may contribute directly and independently to the perception of Animation-Rate. Where the correlation between a physical variable and a perceptual variable

can be experimentally altered, independently of other physical variables, we may infer that the first variable has direct perceptual consequence for a given PAT.

The results in Table 7 reveal that such is the case for at least two of the physical variables under consideration. Speech degradation tends, in several instances, to reduce the correlation between PAT 3 and one physical variable while not affecting or increasing the correlation of other variables with this trait. In particular, we are led to conclude, if somewhat predictably, that speech time (and perhaps inherent level) contribute to the perception of Animation-Rate independently of the remaining physical variables and possibly of each other. Still other, less pronounced, examples of this type can be found in the table. However, further research is clearly required in this connection.

Effects of Speech Degradation

Our primary concern thus far has been with the "normal" psychophysics of voice perception — with relations between the perceived characteristics of undegraded voice samples and selected physical voice variables.

A number of important psychophysical relations, as well as relations among physical voice variables, have been at least tentatively established. On the assumption, however, that they will find verification in the results of further research, the question arises as to what generalizations are warranted by them. Can they, for example, serve as a basis for predicting the perceptual consequences of various experimental transformations of the speech signal?

Table 7. Correlations Between Voice Ratings on PAT 3 (Animation-Rate) And Seven Physical Variables Under Four Transmission Conditions

	Physical Variables						
Condition	LST	ST	SS	SC	SJ	L	IL
Unprocessed (PAT 3_u)	-0.40	-0.41	-0.70	0.32	-0.59	-0.47	0.12
Vocoded (PAT 3_v)	-0.32	-0.45	-0.58	0.27	-0.37	-0.32	0.23
Low-Passed (PAT 3_{lp})	-0.19	-0.43	-0.63	0.30	-0.50	-0.42	0.20
High-Passed (PAT 3_{hp})	-0.47	-0.52	-0.76	0.38	-0.65	-0.44	0.20

Before attempting to answer this question, let us examine the implications of various hypotheses concerning the psychophysics of voice perception. Consider first some specific implications of the hypothesis of a "fixed" psychophysics for cases involving degraded speech.

If the effect of degradation is simply to obscure the physical correlate of a given PAT in some degree, the most immediate consequence will be a reduction in the correlation between listener ratings on the trait under normal conditions and ratings of the same trait under degraded speech conditions. Thus, to the extent that vocoderization, for example, serves simply to obscure the normal physical correlates of Pitch-Magnitude, the primary consequence would be a reduction in the loading of PAT 1_V on factor I. To the extent that vocoderization distorts the normal physical correlates of Pitch-Magnitude, a reduction in the loading of PAT 1_V on factor I would be again expected. Additionally, however, extreme distortion of the normal physical correlate of PAT 1 may result in the emergence of a new factor representing the "distortion component of variance" in PAT 1_V. Recall again that in both of these cases the listener's perceptual response to a particular physical parameter of the speech signal, as presented to him, is assumed to be unchanged.

The alternative of the hypothesis of "fixed" psychophysics is the modified "daisy" hypothesis. This hypothesis holds that, when deprived of the normal physical correlate of a given PAT, a listener allocates the "perceptual channel" associated with that PAT to a different physical voice variable. Conceivably, the new variable may be one of previously negligible perceptual consequence. On the other hand, it may be the physical correlate of another perceived acoustic trait.

Evidence of the latter type of "psychophysical shift" is provided by instances where ratings of a given PAT under a degraded speech condition are correlated with ratings of a different PAT than in the case of unprocessed speech. If, for example, PAT 1_v were found to have a high loading on factor II Loudness-Roughness, we would infer that the normal stimulus correlate of Pitch-Magnitude was obscured by vocoderization and that the listener had responded by allocating his "Pitch-Magnitude" channel to the stimulus correlate of Loudness-Roughness. At present, however, a third possibility must be considered in evaluating such results. This is the possibility that a particular form of degradation operates to induce a correlation between normally uncorrelated physical voice variables and, in turn, to induce a spurious psychophysical shift.

The kinds of data presently available do not permit rigorous distinction between true and spurious psychophysical shifts. Additional physical acoustic measurements will be required for this. There are, however, some present findings which are sufficiently suggestive of each possibility to warrant their discussion here. While it is easy to conceive of how the drastic transformation of the speech signal effected by vocoderization can provide conditions conducive to spurious psychophysical shifts, it is more difficult to conceive of how simple frequency distortion can do this. Subject to the results of additional research, therefore, we are inclined to attribute shifts observed in cases of frequency filtered speech to true changes in the physical bases of certain perceived voice characteristics. Let us now re-examine some of the results in Table 5.

In the case of PAT 1, it would appear that none of the effects discussed above has occurred. The loadings on factor I of variables involving PAT 1 are essentially unchanged over the range of transmission conditions represented here. Subject to the results of future experimentation, we are therefore led to conclude that the present results for Pitch-Magnitude are not inconsistent with the hypothesis of a fixed psychophysics.

PAT 2 presents a different picture. For the case of vocoded speech (PAT 2_v), there is a substantially reduced loading on factor II, which is normally identified with Loudness-Roughness. This is accompanied, moreover, by an increased loading on factor I, indicating the possibility of a psychophysical shift. For the case of low-passed speech, there is again a suggestion of a shift in the physical correlate of Loudness-Roughness. The depressed loading of PAT 2_{lp} on factor II (.52) is accompanied by an elevated loading (.58) on factor I, suggesting psychophysical shift, but in the opposite direction of the shift found with vocoded speech. While vocoding "induced" a negative loading of PAT 2 ratings on factor I, low passing induced a positive loading on the same factor.

For the case of high-passed speech, there is no indication of a substantial shift in the physical correlate of Loudness-Roughness. Rather, the depressed loading of PAT 2_{hp} on factor II (.78) can find an explanation consistent with the hypothesis of a fixed psychophysics. The results for PAT 2_{hp} can, in other words, be accounted for simply in terms of a tendency for high passing to obscure the normal physical correlates of PAT 2.

The results for PAT 4 (identified with factor III) provide additional indications of possible psychophysical shifts, whether true or

spurious. The results for vocoded speech are particularly interesting in this connection. PAT 4_v has a negative loading on factor III, indicating that judgments of Clarity-Beauty under this condition are inversely related to judgments of the same trait under normal conditions. At the same time, there is a possible indication that, with vocoded speech, judgments of this trait are determined by the normal physical correlate of Pitch-Magnitude. For both high-passed and low-passed speech, the variables associated with PAT 4_v (PAT 4_{lp} and PAT 4_{hp}) have depressed loadings on factor III, accompanied by elevated loadings on factor I.

Since the PAT Animation-Rate is not uniquely associated with any of the orthogonal factors revealed by the present analysis, the effects of the various experimental conditions upon this variable are not so clearly apparent in Table 5. In graphic presentation, however, they are more easily apprehended.

Figure 7 shows a plot of all 23 voice variables in the I-II (Pitch-Magnitude — Loudness-Roughness) plane of the perceived acoustic trait space. To facilitate interpretation, the points for all variables involving a given PAT are joined by lines.

From the figure it is evident, first, that no psychophysical shifts of any consequence occurred in the case of PAT 3, at least within the I-II plane of the trait space. Moreover, other projections of the trait space fail to reveal any evidence of shifts in the physical correlation of this PAT.

Figure 7 also serves to provide graphic demonstration of the results described above for other PAT's. Thus, the "tight grouping" observed in the case of PAT I illustrates the psychophysical stability of Pitch-Magnitude, while the dispersion of points involving each of the re-

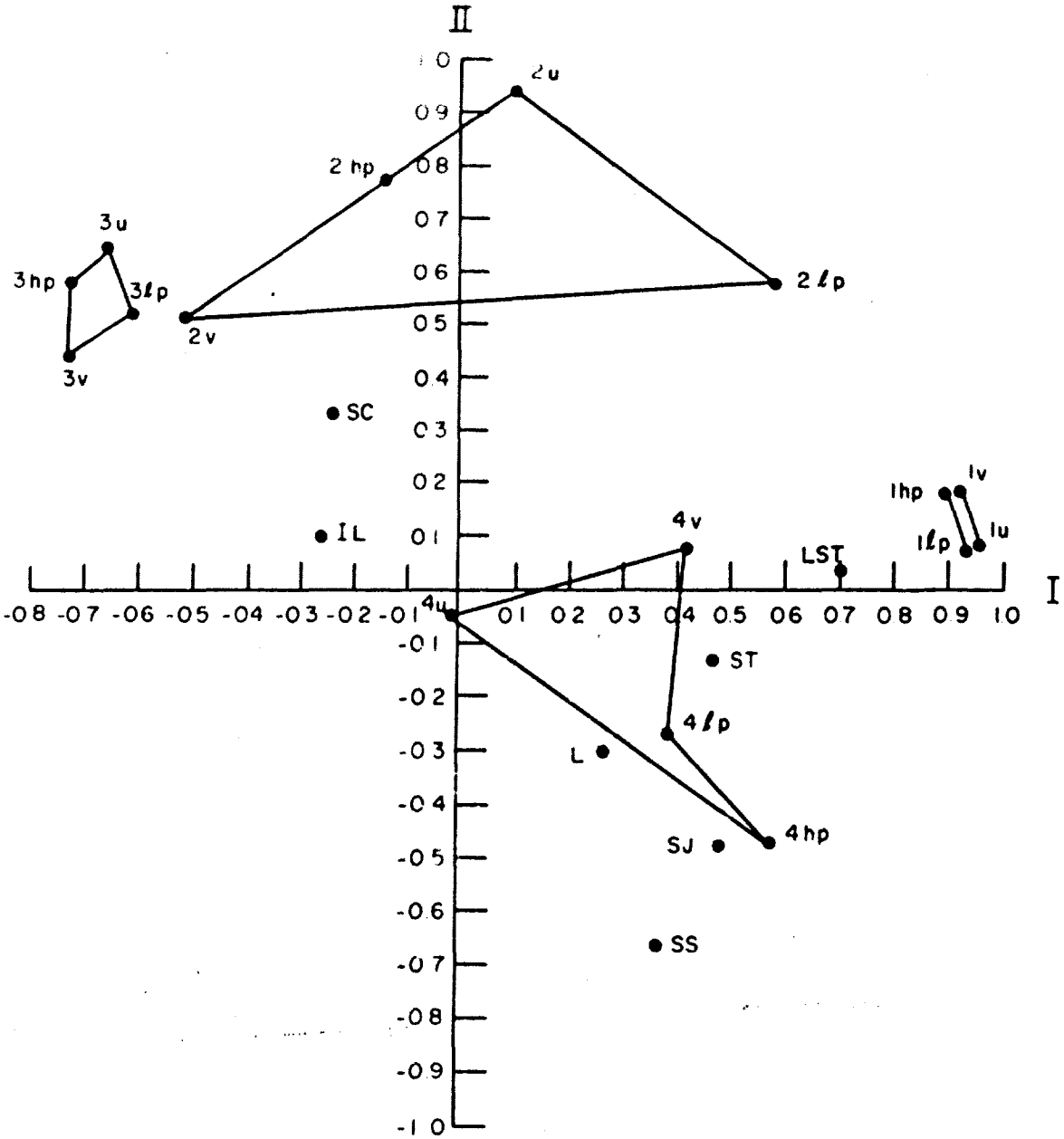


Figure 7. Factor Loadings of Twenty-Three Voice Variables in the I-II Plane of the Perceived Acoustic Trait Space.

maining PATS dramatically portrays some of the evidence as to their psychophysical instability.

C. CONCLUSIONS

The results presented in this chapter indicate that while the physical correlates of some perceived acoustic traits are quite stable over various forms of speech degradation, the physical correlates of others may be altered by degradation of the speech stimulus. The normal physical correlates of Pitch-Magnitude and Animation-Rate, identified thus far, exhibit little change with stimulus impoverishment, but the stimulus correlates of Loudness-Roughness and Clarity-Beauty are drastically altered by frequency filtering and vocoderization.

The latter results are consistent with a modified "daisy" hypothesis which would attribute them to qualitative psychophysical shifts — true changes in the stimulus correlates of various perceived voice characteristics. Conceivably, however, some or all of these results could also find explanation in terms of spurious shifts brought about by experimentally induced correlations between the normal stimulus correlates of two or more perceived acoustic traits.

Available data do not suffice to resolve the ambiguity of the present results. Additional correlated data on the physical acoustic implications of speech degradation will be required for this purpose. In any case, however, the present results raise some significant issues concerning the uses of voice-rating data for purposes of system evaluation.

Where the assumption of a fixed psychophysics can be justified, voice-rating data are potentially of value for purposes of indirect evaluation of the fidelity with which specific physical voice variables are transmitted.

Spurious psychophysical shifts are not, of course, inconsistent with the concept of a rigid psychophysics. By definition they arise simply from experimentally-induced changes in the interrelations among physical voice variables. Where the possibility of true psychophysical shifts can be eliminated, voice-rating data are of potential value simply in detecting system deficiencies in the transmission of specific physical parameters of interindividual differences in speech. Given the possibility of true psychophysical shifts, the diagnostic value of voice-rating data is necessarily limited. Under these circumstances, however, such data may yield more valid measures of gross system performance with respect to speaker recognizability than perhaps any practical method based on physical measurements of transmitted speech.

In light of these considerations, it is evident, at least, that further research to distinguish between the physical and psychological bases of psychophysical shifts will contribute significantly to the technology of system evaluation from the standpoint of speaker recognizability.

REFERENCES

1. Asher, J. W., T. D. Hanley, and M. D. Steer, "A Factor Analysis of Twelve Physical Measures of Voice," NAVTRADEVCEN Tech. Report, No. 104-2-48 (1957).

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

3. Helson, Harry, "Adaptation-Level Theory," New York, Harper and Row, 1961.

4. Holmgren, G. L., "Physical and Psychological Correlates of Speaker Recognition," Unpublished Ph.D. dissertation, Texas Christian University, 1964.

5. House, A. S., C. Williams, M. H. L. Hecker, and K. D. Kryter, "Psychoacoustic Speech Tests: A Modified Rhyme Test," Decision Sciences Lab. Tech. Report, No. ESD-TDR-63-403 (1963).

6. Joos, Martin, "Acoustic Phonetics," Language Monograph No. 23, 1948.

7. Ladefoged and Broadbent, D. E., "Information Conveyed by Vowels," J. Acoust. Soc. Am., 29 (1957), 98-104.

8. Miller, G. A., and P. A. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," J. Acoust. Soc. Am. 27, 338 (1955).

9. Peters, Robert W., "The Effect of Length of Exposure to Speaker's Voice Upon Listener Reception," Joint Project Report No. 44, The Ohio State University Research Foundation and U. S. Naval School of Aviation Medicine (1955).

10. Pollack, I., "Information of Elementary Auditory Displays," J. Acoust. Soc. Am., 24, 745 (1952).

11. Pollack, I., "The Information of Elementary Auditory Displays, II," J. Acoust. Soc. Am., 25, 765 (1953).

12. Quastler, Henry, "Studies of Human Channel Capacity" in "Information Theory," Third London Symposium, New York, Academic Press, 1956.

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

14. Voiers, W. D., "Further Studies of the Dimensions of Perceived Variation Among Voices," Unpublished research, Sperry Rand Research Center (1963-64).

15. Voiers, W. D., Cohen, M. W. and Mickunas, J., "Performance Evaluation of Speech Processing Devices, I. Intelligibility, Quality, Speaker Recognizability," Final Report, Contract No. AF19(628)-4195, AFCRL 1965.

DOCUMENT CONTROL DATA - R&D

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

Sperry Rand Research Center Sudbury, Massachusetts 01776	2a REPORT SECURITY CLASSIFICATION unclassified 2b GROUP ----------

3 REPORT TITLE

Performance Evaluation of Speech Processing Devices
II. The Role of Individual Differences

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Scientific Report (1 March - 30 November 1965)

5 AUTHOR(S) (Last name, first name, initial)

Voiers, William D.

6 REPORT DATE	7a TOTAL NO OF PAGES	7b NO OF REFS
December 1965	57	15

8a CONTRACT OR GRANT NO	9a ORIGINATOR'S REPORT NUMBER(S)
AF19(628)-4987 b PROJECT NO 4610 Task No. 461002 c DOD Element No. 62405304 d DOD Subelement No. 674610	SRRC-RR-65-115 9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AFCRL-66-24

10 AVAILABILITY LIMITATION NOTICES

Distribution of this document is unlimited.

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY
	AF Cambridge Research Laboratories (CRBS) Office of Aerospace Research, USAF Bedford, Massachusetts

13 ABSTRACT This report treats three topics related to the phenomenon of individual differences in speech: (1) the effect upon speech intelligibility of the listener's familiarity with the speaker's voice: (2) the nature and number of elementary ways (perceived acoustic traits) in which voices are perceived to differ from each other; (3) the physical bases of perceived acoustic traits.

It was found that speech intelligibility as measured by the Diagnostic Rhyme Test was unaffected by the degree of the listener's familiarity with the speaker's voice, where the basis of familiarity was a maximum of three presentations of a two-minute sample of prose (Gettysburg Address).

A voice rating experiment, involving both mono-polar and bi-polar ratings scales failed to reveal any previously unidentified perceived acoustic traits, though several items were found which offer possibilities for increased precision in the evaluation of previously identified PAT's.

A factor analysis of 22 physical and perceptual voice variables yielded results to indicate that the physical correlates of some perceived acoustic traits may be qualitatively shifted under certain conditions of stimulus impoverishment. The physical bases of the PAT's, Pitch-Magnitude and Animation-Rate appear to be relatively stable under various conditions of stimulus impoverishment, but the physical bases of Loudness-Roughness and Clarity-Beauty may be drastically altered by frequency distortion and vocoderization.

DD FORM 1473

Security Classification

Security Classification

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Speech Evaluation						
System Evaluation						
Intelligibility						
Speaker Recognizability						
Speech Quality						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Security Classification