AFCRL-67-0101

PERFORMANCE EVALUATION OF SPEECH PROCESSING DEVICES
III. DIAGNOSTIC EVALUATION OF SPEECH INTELLIGIBILITY

by

William D. Voiers

AD 650158

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

Final Report: March 1967

Period Covered: 1 March 1965 – 30 November 1966

Distribution of this document is unlimited.

Prepared for

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

D D C
APR 17 1967
B

SPERRY RAND RESEARCH CENTER
SUDBURY, MASSACHUSETTS

ARCHIVE COPY

The SPERRY RAND Corporation

AFCRL-67-0101 SRRC-CR-67-6

PERFORMANCE EVALUATION OF SPEECH PROCESSING DEVICES
III. DIAGNOSTIC EVALUATION OF SPEECH INTELLIGIBILITY

by

William D. Voiers

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

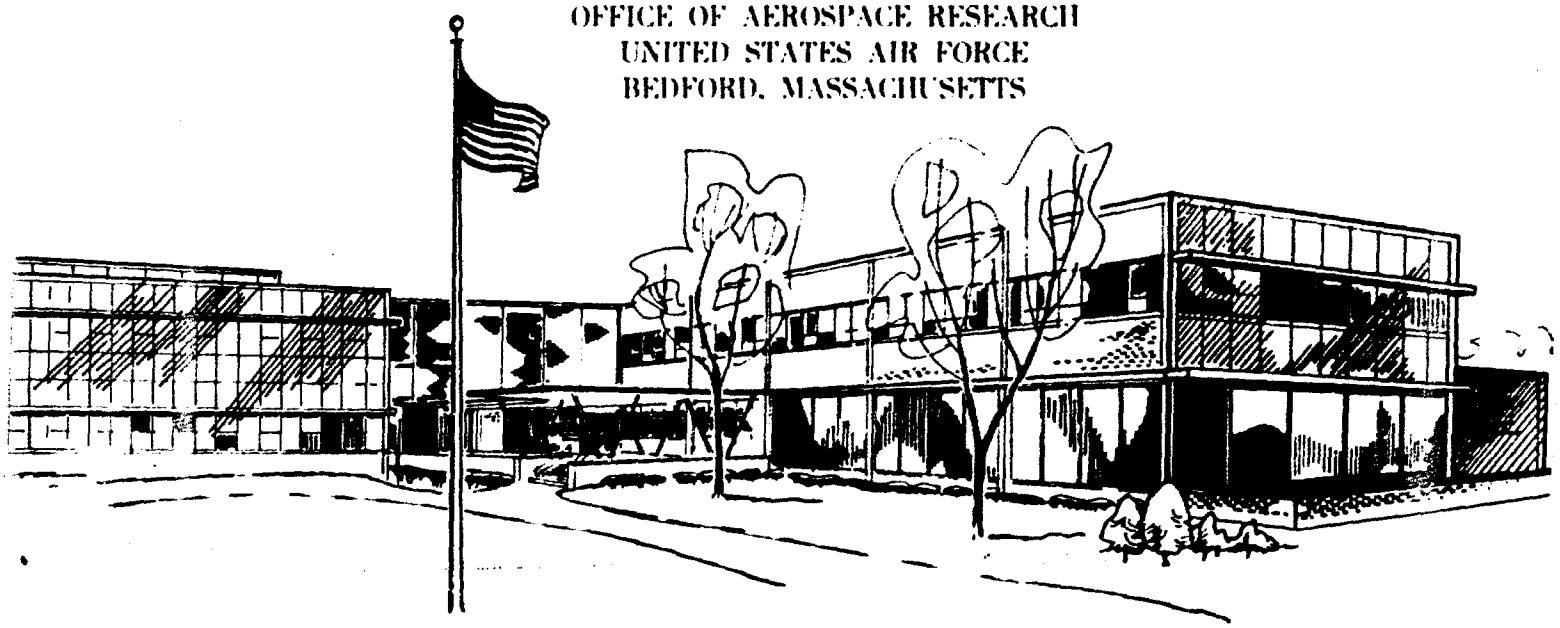
Final Report: March 1967

Period Covered: 1 March 1965 – 30 November 1966

Distribution of this document is unlimited .

Prepared for

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



S R R C

ABSTRACT

The report summarizes the results of a program of research on communication system evaluation from the standpoint of speech intelligibility and speaker recognizability.

The history and present status of the Diagnostic Rhyme Test (DRT) Form III are described along with the results of research relating to the validity of the DRT in various applications.

TABLE OF CONTENTS

	Page
INTRODUCTION	
Publications	1
Papers Presented	1
Reports	1
Conferences	2
Summary of Major Research Activities	3
APPENDIX - Diagnostic Evaluation of Speech Intelligibility	

INTRODUCTION

Summarized below are the major accomplishments of the Speech Research Group of the Sperry Rand Research Center under Air Force Contract AF19(628)-4987 from May 1965 through November 1966.

Publications

No publications based on research conducted under this contract have yet appeared in any of the relevant technical journals. However, one major report, "Diagnostic Evaluation of Speech Intelligibility," is presented here in essentially the form in which it will be submitted for publication in the Journal of the Acoustical Society of America. Eight other reports, based on research conducted under Contract No. AF19(628)-4987, are in various stages of preparation for submission to the appropriate technical journals. They are tentatively titled as follows:

"The Physical Bases of Perceived Voice Characteristics," W. D. Voiers and J. Mickunas

"Fairbanks Rhyme Test as a Multiple-Choice Test," J. Mickunas and W. D. Voiers.

"Cross Validation of Two Methods for the Psychological Scaling of Voices," J. Miller and W. D. Voiers.

"Elementary Dimensions of Phonemic Confusion in Short-Term Memory: A Re-examination of the Taxonomic Implications of Wicklegren's Results," W. D. Voiers.

"The Elementary Dimensions of Malperformance in Channel Scaling of Voices," J. Miller and W. D. Voiers.

"Elementary Dimensions of Malperformance in Channel Vocoders: A Factor Analytic Study with the Diagnostic Rhyme Test," W. D. Voiers, J. Mickunas and V. Miethe.

"An Item Analysis of the Diagnostic Rhyme Test," W. D. Voiers and V. Miethe.

Papers Presented

Three papers were presented at the fall 1965 meeting of the Acoustical Society of America in St. Louis, Missouri. These were:

"Comparative Evaluation of Conventional Whispering and Monotone Vocoder Speech," W. D. Voiers.

"Effect of Multiple Vocoderization and Consonant Intelligibility," M. F. Cohen and W. D. Voiers.

"Further Developments in Unit-Variance Scaling Method "
J. Mickunas

Dr. Voiers presented a paper at the fall 1966 meeting of the Acoustical Society of America. It was entitled:

"The Physical Bases of Perceived Voice Characteristics "

Reports

In accordance with the terms of the contract, a technical report, "Performance Evaluation of Speech Processing Devices II, The Role of Individual Differences," AFCRL-66-24, was submitted in December 1965

Conferences

Dr. Voiers visited the Stanford Research Institute, Palo Alto, California in April 1966, to confer with Dr. Frank Clarke on matters of common interest in the area of voice recognition and intelligibility testing.

Dr. Voiers accompanied Mr. Caldwell Smith on a visit to Texas Instruments, Inc., Dallas, Texas in April 1966. The purpose of the visit was to discuss problems related to the evaluation of the AFCRL Polymodal Vocoder.

Summary of Major Research Activities

The research activities conducted under this contract were directed toward two major goals: (1) further refinement and validation of the techniques developed under AF19(628)-4195 for the diagnostic evaluation of intelligibility in speech processing systems and; (2) identification of the physical bases of perceived voice characteristics and the application of this information to the solution of practical problems of system evaluation from the standpoint of speaker recognizability. Substantial steps toward the achievement of these goals were accomplished during the period covered by this contract. They are briefly reviewed below.

Diagnostic Evaluation of Intelligibility

Early in the course of the present contract steps were taken to apply the insights gained in the course of AF19(628)-4195 to the development of more refined and reliable techniques for the diagnostic evaluation of voice communications equipment.

The most immediate result of this effort was a modified version of the original Diagnostic Rhyme Test. This version of the test while still

providing for the measurement of consonant discriminability in only four vowel contexts, utilized a different set of vowels than the original DRT. While in the original DRT [i], [ɛ], [u] and [ɔ] were used, [ae] was substituted for [ɛ] in Form II, [ae] was substituted for [ɛ]. Otherwise it employed only those items of the original DRT which were of empirically demonstrated validity and reliability. Additional items were provided to replace those found to be unsatisfactory in the original DRT. Unlike the original DRT, however, Form II did not involve the repeated use of certain test words or word pairs. It also utilized a slightly different temporal pattern of stimulus word presentation. This modification (designed to simplify the task of the speaker) was subsequently abandoned, however.

Extensive tests with Form II of the DRT revealed its validity to be equal if not superior to that of the original DRT. However, these tests also revealed various means for further improving the test. As a consequence, Form II was shortly abandoned in favor of Form III. Form III is the current version of the DRT and embodies all of the insights gained in the course of over two years of research on the problems of diagnostic evaluation of voice communications equipment. It is described in detail at another point in this report.

In the course of refining the Diagnostic Rhyme Test it became necessary to undertake a fairly comprehensive program of research on various basic issues concerning the nature of speech perception, in particular concerning the nature of the elementary perception attributes of consonant phonemes. The results of this effort have led to some potentially fundamental discoveries concerning the optimal taxonomy for consonant phonemes. While this taxonomy appears to differ somewhat from the taxonomy on which the DRT Form III was based, no changes thus far indicated (e.g., a change in the classification of the liquids and glides from "positive" to "neutral" with respect to the attribute, sustension), have operated to invalidate the design of the DRT. However, the need for further research on this issue is clearly evident.

As one phase of the program for validating the DRT it was appropriate that relations between Diagnostic Rhyme Test scores and Fairbanks Rhyme Test scores be extensively investigated. From the results obtained

thus far it appears that the total DRT score provides a gross figure of system performance which is very nearly equivalent to that provided by the Fairbanks test. Even closer equivalence can be achieved by differential weighing of the various diagnostic scores yielded by the DRT. In this connection, it might be noted that a diversity of results have emerged to attest to certain serious deficiencies in the Fairbanks test In particular it has been found that the Fairbanks test is virtually insensitive to system deficiencies in the transmission of information with respect to the elementary consonant attribute, sibilation. Again however, the need for further research is evident.

Speaker Recognizability

Early in the course of the present contract an extensive program of research was undertaken to refine the multidimensional (semantic differential) method of classifying voices on the basis of perceived acoustic traits and for evaluating the fidelity with which speech processing systems (e.g., vocoders) transmit these traits An attempt was also made to identify the major acoustical correlates of these traits. (The bulk of this work has been described in a scientific report, AFCRL-66-24 December 1965) It was discovered among other things that the psychophysics of voice perception may undergo rather drastic qualitative changes under certain conditions involving impoverished speech. On one hand, the physical correlates of certain perceived acoustic traits may shift qualitatively when the "normal" physical correlates of these traits are obscured On the other hand the oblivation of certain minor, but perceptually significant acoustic features may act to increase the strength of certain major psychophysical relations - listeners evidently extract greater amounts of speaker identity from the remaining, unobscured physical voice traits While such effects as these do not vitiate the voice rating approach to the evaluation of speaker recognizability, they necessarily limit the scope of its usefulness. It has become evident in any case that further research with the voice rating method will be required to render it suitable for routine purposes of system evaluation. The possibilities which merit consideration in this connection include the provision of better standards for comparison in rating voices, improved instructions and illustrations of the listener's task and the development of training methods for listeners

Other research relating to the problem of speaker recognizability was concerned primarily with the cross-validation of the four-dimensional voice taxonomy developed in the course of extensive research with the voice rating method. This involved an experiment in which listeners judge the similarity of all possible pairs of voices for a sample of 24 speakers. An attempt was then made to scale the 24 voices using a method of "non-metric multidimensional scaling" developed by Shepherd (Shepherd, Roger N., "The Analysis of Proximities. Multidimensional Scaling with an Unknown Distance Function, I," Psychometrika 27, 125-140, 1962). Thus far the results have defied interpretation, though it is not possible at present, to determine whether the fault lies with the experimental procedure or with the method of analysis. Further research on this matter is needed. In particular it would be desirable to investigate other methods of scaling the present data.

Services

During the course of this program a number of evaluations (DRT) were performed on various vocoders in various operating modes. The results of these evaluations have been transmitted in monthly letter reports.

To meet the immediate needs of the present program and also to provide materials for future use in research with the AFCRL Patterns Vocoder, a diversity of speech materials was simultaneously recorded with three microphones. These include a high quality ceramic microphone placed at a distance of 20 cm from the speaker's; a dynamic microphone, slightly to the side, at a distance of 4 cm and a "larynx microphone" held in contact with the speaker's throat. Recorded materials included the following.

SUMMARY OF RECORDED SPEECH MATERIALS

MATERIALS	SPEAKERS					
	RD - trained	RC - neutral	SL - neutral	CH - neutral	BH - high	JC - low
DRT-III lists 1.33 sec time	1-8	1-4	3-8	3,4,7,8	3-6	1-4
FRT lists 1.33 sec time	2,3,4		6,7	4,10,11	7	6
Sentence list 1	---	---	4 times	4 times	2 times	2 times
Sentence list 2	---	---	4 times	4 times	2 times	2 times
Sentence list 3	---	---	4 times	4 times	2 times	2 times
"Gettysburg Address"	6 times	---	4 times	4 times	2 times	2 times
"Fairy Tales" (normal)	---	---	4 times	4 times	2 times	2 times
"Fairy Tales" (style)	---	---	2 times	2 times	1 time	1 time
"Battle of Berkeley" (normal)	---	---	4 times	4 times	2 times	2 times
"Battle of Berkeley" (style)	---	---	2 times	2 times	1 time	1 time
"On Writing"	---	---	4 times	4 times	2 times	2 times
"Water Divides"	---	---	4 times	4 times	2 times	2 times
"Psychology Test"	1 time	1 time	---	---	---	1 time

APPENDIX

DIAGNOSTIC EVALUATION OF SPEECH INTELLIGIBILITY

DIAGNOSTIC EVALUATION OF SPEECH INTELLIGIBILITY

INTRODUCTION

Voiers and his collaborators[1] have described the construction and use of an experimental, seven-dimensional test of initial consonant discriminability: the Diagnostic Rhyme Test (DRT). The DRT is a two-choice test, designed such that the listener's response to each stimulus word provides an indication of the discriminability of an elementary attribute of consonant phonemes. While conceived primarily for purposes of diagnosing the inadequacies of communications systems and components, the test appears to have some potential for evaluating the characteristics of speakers and listeners as well. This report summarizes the major results of experimental studies with various forms of the DRT and presents a refined version (Form III) based on the results of research with earlier forms of the test.

ISSUES IN THE DESIGN OF INTELLIGIBILITY TESTS

The Diagnostic Rhyme Test, in all forms, is based on a fundamentally different principle than other intelligibility tests in general use today. This difference in principle is perhaps most apparent when the DRT is compared to the venerable PB test of word recognizability,[2] but important differences also exist between the DRT and some of the more recently developed tests of consonant intelligibility (e.g., the Fairbanks Rhyme Test,[3] the Modified Rhyme Test of House *et al.*,[4] and the Phonemically Balanced Rhyme Test of Clarke[5]). A test for consonant recognizability developed by Stevens, Hecker and Kryter[6] is perhaps most similar to the DRT in terms of the motivation for its development and the general principles upon which it is based. Even here, however, there are significant differences in method and in the assumptions made concerning the nature of the speech recognition process.

Like the latter tests, the DRT is concerned exclusively with the issue of consonant intelligibility, since the primary motivation for its design was the need for methods of evaluating communication systems in which vowel intelligibility is normally a problem of relatively minor consequence. Unlike them, however, the DRT does not purport to test for consonant recognizability, *per se*. Rather its purpose is to test for the discriminability of minimal differences among consonant phonemes in terms of certain elementary phonemic attributes or distinguishing features. Thus the listener's task with the DRT is not to recognize speech sounds in the usual sense, but, in effect, to discriminate the states of various elementary phonemic attributes.

Recognition in the usual sense of the word, of course, presupposes the discrimination of one or more such attributes, but may also depend upon various extra-stimulus factors (e.g., the listener's expectations) which in the practical testing situation, may complicate attempts to isolate the effects of

specific channel variables, speaker variables, or nonextraneous listener variables. Therefore, even where ultimate concern is with phoneme recognizability in the usual sense, a test of the discriminability of various phonemic attributes may permit better control of extraneous variables and ultimately provide a more valid evaluation of the potential of a speaker, channel, or listener with regard to phoneme recognizability.

There are other theoretical and practical advantages to the discrimination approach, as exemplified by the DRT. These become particularly important where it is desirable to obtain not only a gross measure of intelligibility but also to pinpoint the specific deficiencies of a speaker, listener, or system under test. Consider in this connection the nature of the information yielded by more conventional tests of speech recognizability, such as the PB word recognition test. Here, the listener's task is to respond to a stimulus word by choosing among an unspecified, if not unknown, set of alternatives. A correct recognition response may indicate that some one or more phonemic attributes has been correctly evaluated, but the number and nature of these attributes remain unknown so long as the alternatives considered by the listener are likewise unknown.

Incorrect responses in a word recognition test are potentially sources of diagnostic information, but the diversity of incorrect responses potentially evokable by a given word is normally so great as to render either automatic or manual scoring on this basis altogether impractical. The fact that errors may occur in any or all of several phonemes comprising a given PB word serves only to compound this inherent shortcoming of the recognition approach.

The Fairbanks Rhyme Test and Modified Rhyme Test effectively remedy the more conspicuous shortcomings of the PB word test but retain several disadvantages of the recognition approach. In both tests the practical problems associated with an indeterminate response set are substantially reduced

through the restriction of stimulus ambiguity to a single phoneme. Moreover, other features of the two tests provide additional restriction of the response set. The Fairbanks test provides at least implicit restriction of the response set in that experienced listeners tend to confine their choices to elements of the basic corpus of stimulus words used in the test. Explicit specification of response options, as provided by the Modified Rhyme Test and Phonemically Balanced Rhyme Test, offers a more satisfactory solution, however, and serves effectively to control such factors as the listener's past experience, vocabulary and, perhaps, his intelligence.

There remains the question of how the response options to a given stimulus word are to be selected. It is essential that the significance of each erroneous choice should be unambiguous and determinable, for, in general, each of the erroneous choices will have different implications concerning the characteristics of the channel, the speaker, or the listener under evaluation. While this condition can be satisfied with multiple choice tests in general, diagnostic scoring of errors tends to be laborious and complicated with tests which provide more than one alternative to the correct response. The situation becomes especially complicated where it is desired that every response option also serve as a stimulus word, for the discriminative task facing the listener will change depending upon which member of an ensemble of three or more words is used as the stimulus. Thus, the diagnostic significance of a given erroneous response will vary depending upon which member of a set of response options serves as the stimulus word. It becomes prohibitively difficult, moreover, to generate different, but equivalent, random scramblings of a given corpus of stimulus materials where the listener is provided more than two response choices for each item. If only on practical grounds, therefore, a two-choice test in particular has much to recommend it over multiple choice tests in general. Here, the discriminative task facing the listener does not depend qualitatively upon which member of the pair serves as the stimulus word. It varies only in that for one case he is required to detect one

state of a distinguishing feature or attribute while for the other case he must detect the alternative state of the same feature. Random selection of the stimulus words from such pairs can thus be used to generate multiple versions or scramblings of the test materials without altering the validity of the test.

Consider, now, the issue of how the alternative of each stimulus word should differ from the stimulus word itself. Ideally, as suggested above, this difference should be restricted to a single phoneme. In addition, however, it is important that the critical phonemes of a stimulus word and its alternative differ minimally by some criterion or another, such that a correct choice between the members of a pair is as nearly tantamount to an elementary discriminative response as the use of speech stimuli will permit. The issue which remains to be resolved thus concerns the nature of the optimal set of phonemic attributes or test dimensions, i.e., the basic types of discrimination to be required of the listener. Here, the most important criteria of optimality are exhaustiveness and relevance.

It is essential, first, that the attributes which the listener is required to discriminate are sufficient to classify each phoneme of interest in a unique manner. Only then can measures of the discriminability serve as a basis for predicting recognizability. In addition, however, it is especially desirable that each of the attributes in question bear in a unique manner upon some aspect of the communication process; that each represents a dimension of interphonemic variation which is elementary in some sense or another. Ideally perhaps all such attributes should be definable in three ways — genetically, acoustically and perceptually. However, where knowledge of the psychophysics and psychophysiology of speech is insufficient to warrant the use of all three criteria, fewer must suffice. Where various criteria yield contradictory or ambiguous classifications, perceptual considerations

should perhaps be given priority, as in fact they were in formulating the consonant taxonomy upon which Form III of the Diagnostic Rhyme Test is based.

BACKGROUND AND RATIONALE OF THE DIAGNOSTIC RHYME TEST

The initial version of the Diagnostic Rhyme Test[7] was based on a consonant taxonomy similar to that formulated by Miller and Nicely.[8] However, the Miller-Nicely characterization of place of articulation as a ternary attribute was discarded in favor of a three-dimensional binary characterization as shown in Table 1. The remaining dimensions of the Miller-Nicely taxonomy were used without modification.

This version of the DRT provided tests of the discriminability of each attribute in only four vowel contexts and utilized only those consonant phonemes explicitly classified by Miller and Nicely. These did not include the affricates, the liquids, nor the glides. Thus the original DRT provided tests of the discriminability of some attributes only under a relatively limited sub-set of the circumstances in which they were potentially crucial to phonemic recognition. Arbitrary restrictions on the choice of test speech materials possibly served further to limit the validity of this form of the test. In spite of these limitations, the original DRT proved to be extremely sensitive to various forms of speech impoverishment when used for gross evaluation of speech intelligibility as well as for the detection of specific deficiencies of transmitted speech. But once this version had served effectively to validate the general principles upon which the test was based, various possibilities for refinement and extension became evident. Research with the DRT itself yielded valuable information in this connection. For example, a factor analytic study of the DRT[9] provided evidence that two of the "place" dimensions of the test, "middle vs back" and "back vs front" are quite redundant (though of reversed polarity), as the distinctive feature system of Jakobson, Fant and Halle (JFH)[10] might lead one to predict. Other, similar findings, coupled with the results of a review of the recent literature, attested to the inadequacy of the consonant taxonomy upon which the original DRT was based.

A re-examination of various systems (e.g., Halle[11]) which have been proposed to account for perceived relations among consonant phonemes led ultimately to the conclusion that the most parsimonious and perceptually valid characterization of English phonemes could be provided by a slightly altered version of the original JFH taxonomy.[12] The dimensions of this modified JFH taxonomy are shown in Table 1.

The first taxonomic dimension voicing, is recognized as an elementary consonant attribute in all of the major system of phonemic classification. All systems, moreover, are in agreement as to the manner in which this attribute is distributed across the population of English consonant phonemes.

Nasality, like voicing, is generally recognized as an elementary attribute of English phonemes. It serves to distinguish the phonemes /m/ and /n/ and /ŋ/ from all other English consonants.

Sustension corresponds to the continuant-interrupted opposition of JFH and of Halle. Among other things it distinguishes the plosive consonants from all others.

Sibilation corresponds to the strident-mellow opposition as originally employed by JFH. However, since Halle[13] has recently reclassified two phonemes /f/ and/v/ with respect to a similarly titled attribute, the term sibilation was substituted in an attempt to minimize ambiguity as to the essential nature of this attribute and as to the manner in which it is distributed over the population of English consonant phonemes.

With exceptions in the case of the affricates and palatal sibilants, the opposition, grave-acute, serves essentially the same function here as in the JFH taxonomy. Otherwise, the present taxonomy follows JFH rather than Halle in classifying /g/, /k/ and /j/ indifferently with respect to the grave-acute opposition. Graveness serves, among other things, to distinguish phonemes articulated at the front of the vocal cavity from all others.

TABLE 1. SYSTEMS FOR THE CLASSIFICATION OF CONSONANT PHONEMES

	Modified Miller-Nicely (DRT Forms I and II)*		Modified Jakobson *et al.* DRT Form III	
1.	Voicing	(present vs absent)	Voiced	(vs unvoiced)
2.	Nasality	(present vs absent)	Nasalized	(vs unnasalized)
3.	Friction	(vs plosion)	Sustained	(vs interrupted)
4.	Duration	(long vs short)	Sibilated	(vs unsibilated)
5.	Place	(front vs middle)	Grave	(vs acute)
6.	Place	(middle vs back)	Compact	(vs diffuse)
7.	Place	(back vs front)	Vowel-like	(vs nonvowel-like)

*Does not classify liquids, glides and affricates, although Miller and Nicely employ the term affrication synonymously with friction.

With minor exceptions, the attribute compactness performs the same taxonomic function here as in the original JFH taxonomy. Like graveness it serves to distinguish between phonemes articulated at different places in the vocal passage. In particular it distinguishes phonemes articulated at the back from those articulated at the middle and front.

Somewhat arbitrarily, the classification vowel-like is used to distinguish the liquids and glides from the remaining consonants. While perhaps over-simplified, this characterization seems somewhat more consistent with the facts of speech perception than Halle's characterization, which distinguishes between some members of this highly confusable group by means of as many as four binary features.

Table 2 presents the classification of twenty-three consonant phonemes in terms of the seven-dimensional taxonomy described above. Plus signs, minus signs and zeros serve to indicate the state of each attribute for each phoneme. This system of classification provided a basis for compiling a set of word pairs used in constructing Form III of the DRT. With minor exceptions noted on the following page, the members of each pair differ in terms of a single attribute.

TABLE 2. CONSONANT TAXONOMY USED IN THE CONSTRUCTION OF THE DRT (FORM III)

	/m/	/n/	/v/	/ð/	/z/	/ʒ/	/ʤ/	/b/	/d/	/g/	/w/	/r/	/l/	/j/	/f/	/θ/	/s/	/ʃ/	/ʧ/	/p/	/t/	/k/	/h/
Voicing	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
Nasality	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sustention	-	-	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	+
Sibilation	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-
Graveness	+	-	+	-	-	-	-	+	-	o	+	-	o	o	+	-	-	-	-	+	-	o	o
Compactness	-	-	-	-	-	+	+	-	-	+	-	-	o	+	-	-	-	+	+	-	-	+	o
Vowel-like	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-

*The DRT does not test for the discriminability of the opposition, vowel-like — nonvowel-like. However, test words are chosen so as not to confound this attribute with the six attributes for which discriminability is tested.

THE DIAGNOSTIC RHYME TEST (FORM III)

A corpus of 96 rhyming word pairs, shown in Table 3, constitutes the set of stimulus materials used in the Diagnostic Rhyme Test. The structure of the test is reflected in the Table. The items in each block of seven are arranged according as they are designed to test for the discriminability of a particular attribute of the initial consonant phoneme. The order is as follows:

(1) Voicing

(2) Nasality

(3) Sustention

(4) Sibilation

(5) Graveness

(6) Compactness

(7) Filler item (to be used for research purposes, etc.)

The positive state (e.g., grave of each attribute is represented in the left member of each pair; the negative state (e.g., acute) is represented in the right member of each pair. The discriminability of each attribute is tested in each of eight vowel contexts. As shown in Table 3, this involves two vowels from each "corner" of the vowel articulation diagram. Thus items in the four upper left blocks of Table 3 involve high, front vowels, whereas those in the four upper right blocks involve high, back vowels. The low, front vowels are represented in the four lower left blocks, while the low, back vowels are represented in the lower right blocks. No central vowels are used in the DRT.

There are two formally equivalent items (e.g., bean-peen and veal-feel) designed to test for the discriminability of each attribute in each vowel context. No words occur more than once in the corpus used for Form III, though repetitions were

TABLE 3. SPEECH MATERIALS USED IN FORM III
OF THE DIAGNOSTIC RHYME TEST

99.*	BEAN-PEEN	43.	VEAL-FEEL	50.	ZOO-SUE	106.	DUNE-TUNE
51.	NEED-DEED	107.	MEAT-BEAT	2.	MOOT-BOOT	58.	NUDE-DUDE
3.	FEET-PEAT	59.	THEE-DEE	66.	FOOL-POOL	10.	THEW-TOO
67.	SEEM-THEME	11.	CHEESE-KEYS	18.	JUICE-GOOSE	74.	CHEW-COO
19.	PEACH-TEACH	75.	WEED-REED	82.	WOMB-ROOM	26.	MOON-NOON
83.	SHE-SEE	27.	KEEP-PEEP	34.	YOU-WOO	90.	COOT-TOOT
35.**	TEAL-KEEL	91.**	WIELD-YIELD	98.**	RUSE-USE	42.**	BOON-GOON
71.	GIN-CHIN	15.	BIG-PIG	22.	DOTE-TOTE	78.	GOAD-CODE
23.	MITT-BIT	79.	NIP-DIP	86.	NOSE-DOZE	30.	MOAN-BONE
87.	FIN-PIN	31.	THIN-TIN	38.	FORT-PORT	94.	VOTE-BOAT
39.	CHINK-KINK	95.	SINK-THINK	102.	JOE-GO	46.	CHOKE-COKE
103.	BID-DID	47.	PIP-TIP	54.	POST-TOAST	110.	MODE-NODE
55.	SHIFT-SIFT	111.	KIT-PIT	6.	GOAL-DOLE	62.	COAL-POLE
7.**	TILT-KILT	63.**	BILL-GILL	70.**	BOLD-GOLD	14.**	SO-SHOW
8.	DENT-TENT	64.	BENT-PENT	1.	GAUZE-CAUSE	57.	JAW-CHAW
72.	MEND-BEND	16.	NET-DEBT	65.	GNAW-DAW	9.	MOSS-BOSS
24.	FEND-PEND	80.	VEST-BEST	17.	THONG-TONG	73.	FAWN-PAWN
88.	JEST-GUEST	32.	JET-GET	81.	SOUGHT-THOUGHT	25.	CHALK-CAULK
40.	PEST-TEST	96.	BED-DEAD	33.	PALL-TALL	89.	BONG-DONG
104.	YEN-WEN	48.	SHED-SAID	97.	CAUGHT-TAUGHT	41.	YAWL-WALL
56.**	PEN-KEN	112.**	TED-KED	49.**	BALL-GALL	105.**	RAW-YAW
36.	VAST-FAST	92.	GAFF-CALF	85.	BOX-POX	29.	DOT-TOT
100.	NAB-DAB	44.	MAT-BAT	37.	MOM-BOMB	93.	KNOCK-DOCK
52.	FAN-PAN	108.	THAN-DAN	101.	VON-BON	45.	FOND-POND
4.	CHAMP-CAMP	60.	SAD-THAD	53.	CHOCK-COCK	109.	JOT-GOT
68.	BAD-DAD	12.	MAP-NAP	5.	POP-TOP	61.	MOB-NOB
20.	CAST-PAST	76.	CAN-TAN	69.	SHOCK-SOCK	13.	COD-POD
84.**	RAM-YAM	28.**	BASS-GAS	21.**	SOT-SHOT	77.**	BOB-GOB

*Numbers to the left of each pair indicate the position of the item in each block of 112 items on the listeners answer sheet.

**Filler items. The manner in which these spaces are filled is at the option of the experimenter. Among other things they may be used for testing experimental items.

necessary in the earlier forms of the test due to the constraints inherent in the method of compiling test items. Even with the greater freedom provided by the design of Form III of the DRT, however, the population of acceptable word pairs was somewhat limited. From Table 3 it is perhaps apparent that there was insufficient latitude to permit any degree of selectivity on the basis of frequency of word occurrence in speech or printed matter. There is, however, a fairly substantial basis for the assumption that frequency-of-occurrence is a factor in word recognizability primarily as it affects the listener's expectations with regard to a given stimulus word. Where the listener is provided with other, more explicit bases for estimating the probability that a given stimulus word will occur, the effects of his past experience (and the expectations it generates) should be minimal. In fact Pollack, Rubenstein and Decker[14] have found that frequency of occurrence has virtually no effect on word intelligibility when the choices permitted the listener are specified for each stimulus word.

It may also be noted in Table 3 that there are several minor exceptions to the rule of "uni-dimensional" difference between members of each word pair. These occur in the case of items designed to test for the discriminability of compactness. While the phoneme pairs /k-p/, /g-b/, /k-t/ and /g-d/ differ primarily with respect to compactness, they might be considered to differ secondarily in terms of graveness since the first member of each pair has a neutral status with respect to graveness while the second member of each pair has a positive status with respect to this attribute. However, the exclusion of all of these pairs would result in a seriously biased sample of the cirucmstances in which compactness is criterial of phonemic identity. Subject to the results of further research on the issue, it was decided, therefore, to include these "contaminated" items in the DRT. It might be noted, that the filler items shown in Table 3 were selected to serve the purposes of such research.

It is clearly impractical in each vowel context to provide complete representation of the cirucmstances in which a given attribute is crucial to phoneme recognizability. However, the attempt was made to achieve a fairly high degree of representativeness for each group of related vowels. Thus while only two "critical phoneme pairs," at most, may be employed to test the discriminability of a given attribute in a given vowel context, greater representation is generally provided within classes (e.g., "high front") of vowels than for individual vowels.

METHODS AND MATERIALS

Selection of Stimulus Materials

The first steps in the preparation of test speech materials involve determining a sequential arrangement of items and selecting a stimulus word from each item. It proves convenient for purposes of computer scoring to order the various items so that the discriminability of each attribute is retested once every seventh item. The vowel context is also changed with each item such that the eight vowels are completely cycled every eight items.

For general testing purposes, the list of test items is cycled four times (normal administration), one stimulus word being selected from each pair on each cycle to yield a total of 448 stimulus words (including 64 filler words). Selection between the words of each pair is random but for the restriction that each attribute is represented an equal number of times in both states in each vowel context. This restriction may result in the equal occurrence of both words from a given item. Otherwise, where the positive state of an attribute occurs a given number of times for one member of a pair of equivalent items, the negative state occurs the same number of times for the other member of the pair (if, for example, "bean" is selected three times from the item bean-peen, "feel" will be selected three times from the item veal-feel). Often, however, it may be desirable to require "perfect balance," i.e., to require that all test words occur as the stimulus with the equal frequency (twice in the case of a normal administration). While this latter arrangement may provide the listener with potentially useful, extra-stimulus information as to the identity of a particular stimulus word, listeners appear unable to make use of this information, (i.e., to use patterns of previous responses as a basis for responding to a given stimulus word) even when instructed to attempt this.

Recording of Stimulus Materials

The stimulus words are normally recorded without carrier phrase at a rate of one word per 1.3 seconds. This rate has proved to be optimal on several grounds.[15] Listeners report that it provides the most comfortable working pace. It yields higher scores and smaller standard errors than faster or slower rates. Finally, it minimizes various types of "testmanship" as factors in listener response, for, as indicated above, listeners have insufficient time between items to permit them to utilize patterns of previous response as a basis for responding. At this rate, a complete test (448 items) can be accomplished in approximately ten minutes.

The recordings of materials used in the research described below were made in a Silence, Inc. sound insulated room of approximately 6' x 6' x 6' interior dimensions. A General Radio microphone (1560-P5) was placed at a distance of 20 cm from the lips of the speaker. A padded head restrainer was used to prevent changes in the speaker's position. This arrangement was found to yield particularly satisfactory recordings but various other equipments and procedures are undoubtedly adequte, depending upon the uses to be made of the test materials and the environment in which recordings are to be made.

Selection of Speakers

For most purposes, one of two types of speakers may be used. It is perhaps preferable ideally, to use a "neutral" voice, i.e., one judged by listeners to be most typical with respect to the "perceived acoustic traits" pitch-magnitude, loudness-roughness, clarity-beauty, and animation-rate, as described by Voiers.[16] However, a trained voice of normal pitch and general American dialect will probably suffice for most purposes. We have obtained quite similar results with these two types of speakers.

For special purposes, voices judged by listeners to be distinguished by various perceived characteristics are used. However, test results obtained with such speakers may vary substantially from those obtained with a neutral voice, particularly where the test speech has been vocoded or otherwise impoverished.

Selection and Training of Listeners

For purposes of evaluating communications equipment and devices a crew of at least eight minimally trained listeners is desirable, although a smaller crew may suffice for special purposes of system diagnosis, e.g., where the system under test is suspected to be particularly deficient in the transmission of a single attribute. Because the test exhibits a degree of listener sensitivity, however, some care should be exercised in selecting listeners. All should have clinically normal hearing over the range from 250 to 8000 Hz. Stricter standards based on previous performance with the DRT should be employed with crews of fewer than eight members. While, in general, effects attributable to practice or familiarity with the DRT have failed to appear, it is perhaps advisable to use one session (448 items) to familiarize the listener with his task before using him for experimental purposes.

Administration of the Test

Because of the importance of timing in the presentation of DRT materials, the use of "live" presentation procedures may be somewhat impractical. The use of prerecorded materials as described above is generally to be preferred. We have standardized on diotic presentation over good quality headphones (e.g., Permoflux PDR-8) though other methods will undoubtedly yield comparable results.

The DRT appears to be relatively insensitive to level. Variation in presentation level over a range of 8 dB (76-84 dB re .0002 dynes cm^2) appears to have negligible influence upon

listener scores. For routine purposes of system evaluation, however, an average vowel peak level of approximately 80 dB SPL appears to be most generally satisfactory.

Scoring the Diagnostic Rhyme Test

Listener response data on the DRT can be evaluated to yield a number of different scores, depending upon the interests of the investigator. Generally, however, greatest interest will attach to seven of these. They include a total score, based on all test items, and six diagnostic scores, each representing the discriminability of one of the six elementary consonant attributes.

Separate scores representing the discriminability of each of the two states of each attribute may be obtained. These are likely to be of particular interest where the effect of a channel variable (e.g., voicing threshold of a channel vocoder) may act to alter the bias rather than the basic sensitivity of a system with respect to a given attribute. Gross measures of consonant discriminability in various vowel contexts may be of value on some occasions, but are unlikely to be of interest for routine purposes of system evaluation. Finally, separate scores, representing the discriminability of each attribute in each vowel context, may be obtained, but these should, perhaps, be interpreted rather cautiously since, in routine testing situations, the amount of data on which they are based will generally be too small for adequate reliability. Moreover, there is reason, as noted above, to question whether a sufficiently representative sample of the instances in which a particular attribute is crucial to phonemic recognition is provided in each vowel context.

It is perhaps apparent that DRT data are eminently suited to evaluation in a framework of signal detection theory. However, a somewhat simpler approach to the scoring problem provides a solution which is probably adequate for most practical purposes.

It involves a crude correction for guessing, accomplished by means of the following formula:

$$S = \frac{100(R - W)}{(T)} ,$$

where S is the "true" percent-correct responses, R is the observed number of correct responses, W is the observed number of incorrect responses and T is the total number of items involved. This correction is applied to all DRT scores including the gross or total score. It finds some practical justification in that it yields DRT total scores which are very nearly equivalent, numerically, to Fairbanks Rhyme Test scores obtained under comparable circumstances. Discrepancies between the two scores can usually be attributed to system deficiencies for which the Fairbanks test is relatively insensitive (for example, deficiencies in the transmission of the physical correlates of sibilation).

EXPERIMENTAL EVALUATION OF THE DIAGNOSTIC RHYME TEST

Normative Data

To provide reference points for evaluating the various forms of communication deficiency, the results of a series of tests involving undegraded speech are presented in Table 4. Shown in the Table are averaged diagnostic scores (two administrations) obtained with a crew of eight experienced listeners for each of five speakers. It appears that, even under ideal conditions, the various attributes of consonant phonemes are neither perfectly nor equally discriminable, though the differences are generally small. It also appears that individual consonant attributes are not equally discriminable for all speakers. Again the differences are generally small, but it will be shown at another point that the DRT is sensitive to individual differences in speech.

Reliability and Sensitivity

House *et al.*[17] have noted that the sensitivity of present day intelligibility tests does not remain constant for all degrees of speech degradation. Rather, it tends to decrease (i.e., results in larger standard errors) up to a point with increasing degradation and then to increase as the speech approaches complete unintelligibility. Results consistent with this trend can also be observed in the case of the DRT. Figure 1 shows the standard error of the mean total DRT score plotted against the mean itself for crews of eight listeners who took the DRT under a diversity of conditions involving impoverished speech. (Somewhat larger standard errors are of course to be expected for the various diagnostic scores.) These results suggest, among other things, the importance of considering inherent test sensitivity when attempting to evaluate intelligibility tests on the basis of gain function (test score change per dB S/N) or other related figures of

TABLE 4. DIAGNOSTIC RHYME TEST NORMS FOR SELECTED SPEAKERS (TWO ADMINISTRATIONS, EIGHT LISTENERS)

Test Dimension	SPEAKER				
	Trained	Neutral(1)	Neutral(2)	Low-Pitched	High-Pitched
Voicing	98	98	96	93	97
Nasality	99	99	99	98	96
Sustention	99	96	94	96	95
Sibilation	99	98	95	98	97
Graveness	100	98	96	98	95
Compactness	98	99	97	98	97
Average	99	98	96	97	96

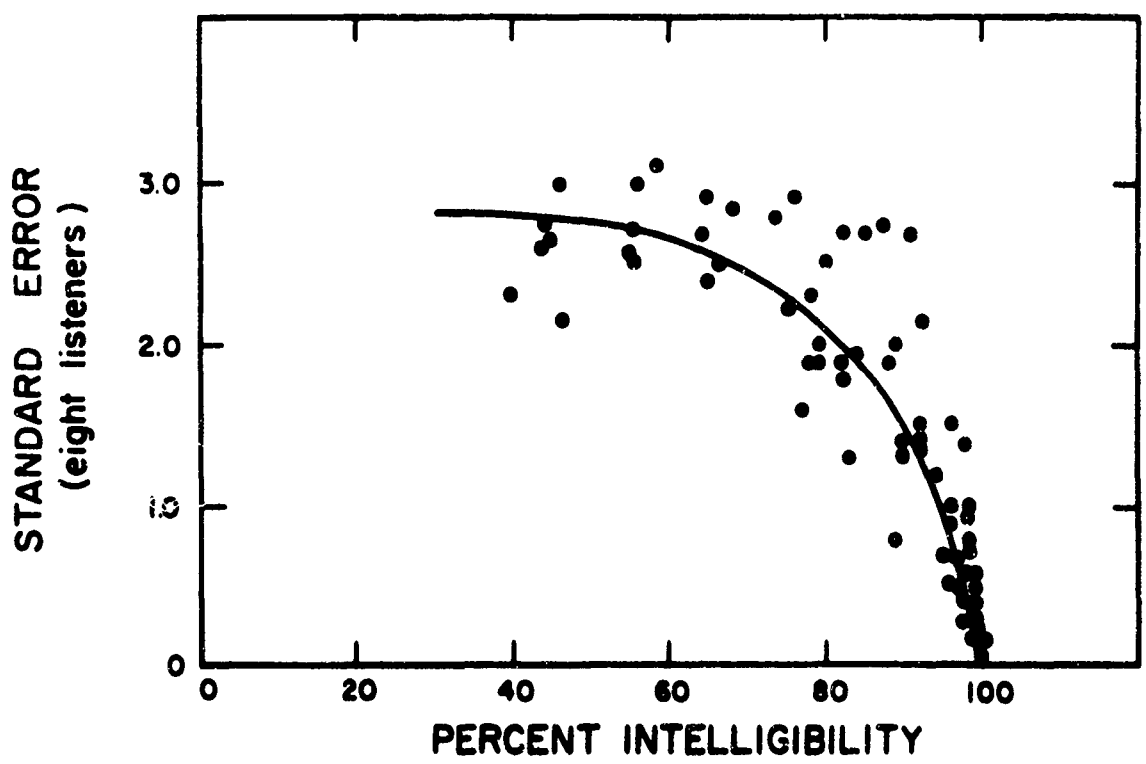


FIG. 1 Standard error of the mean as a function of mean DRT score.

merit. The value of a gain function is meaningful only in relation to the standard error characteristic of the range of test scores involved.

Sensitivity of the DRT to Characteristics of the Channel

In this section the potential of the DRT as an instrument for evaluating the communication channel or link between the speaker and listener is examined. Here, the speaker and listener are effectively held constant (at least in a statistical sense), such that observed variation in DRT scores is attributable to the channel or various of its sub-components (e.g , a vocoder, filter or other processing device). Except where noted otherwise, stimulus materials are provided by recordings of a "neutral" speaker, judged by listeners to be the most representative of a sample of 24 male general American speakers. All listening crews are composed of eight normal-hearing males between the ages of 16 and 19 With minor exceptions, the same crew of listeners performed in all of the experiments described below.

Effects of Noise on DRT Scores

Noise from one source or another is inevitable in every communications situation. It is appropriate, therefore, that first consideration be given to the effects of this channel variable upon consonant intelligibility and, more particularly, upon the discriminability of various consonant attributes. In the experiment performed to evaluate these effects, the speech (vowel) level was approximately 80 dB at the listener's ear. It was held constant over all noise conditions. Both speech and noise were low-pass filtered at 4000 Hz and high-passed at 200 Hz.

Figure 2 shows the effects of band limited noise upon total DRT scores and of Fairbanks Rhyme Test scores obtained under identical conditions. It appears, here that the two

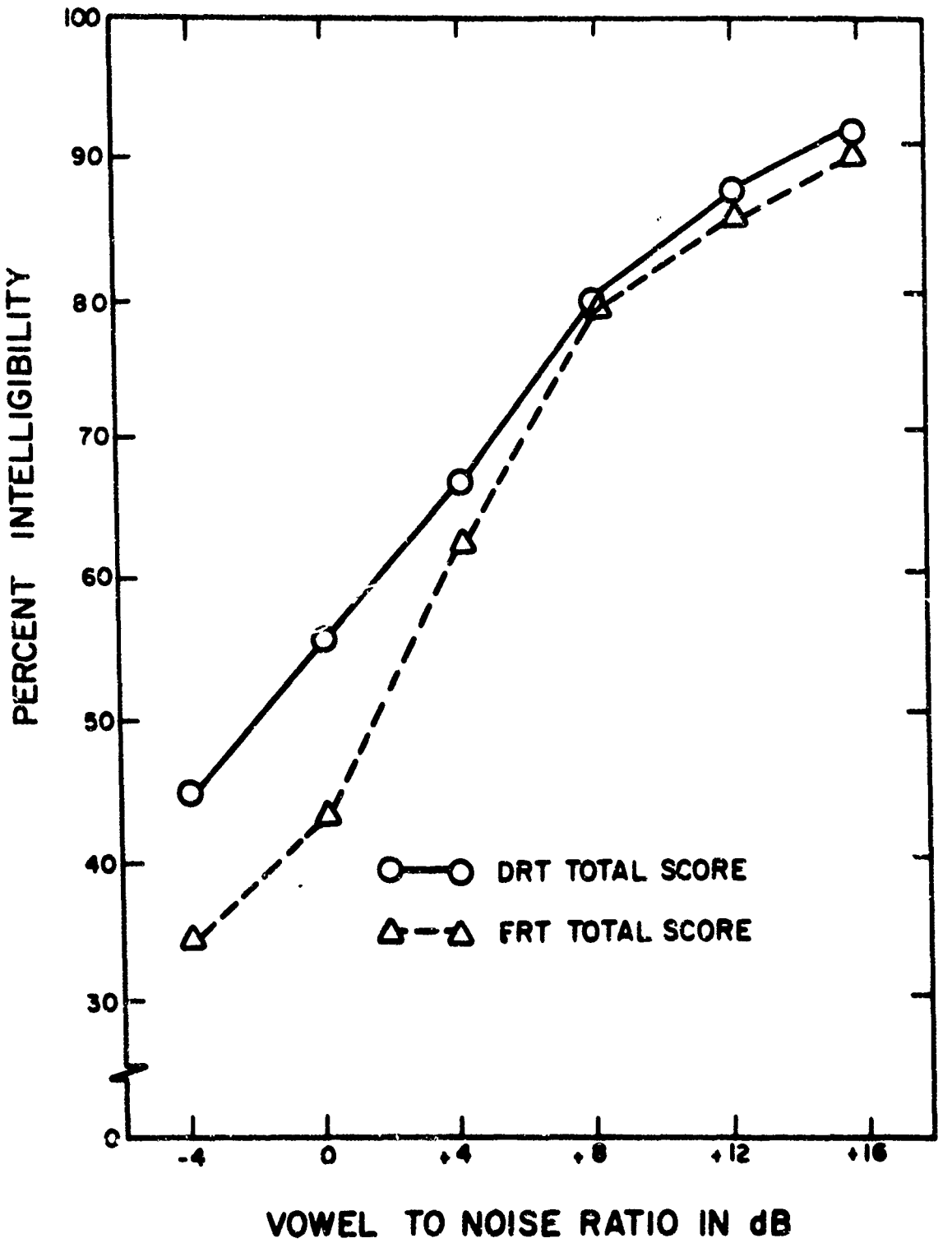


FIG. 2 Mean DRT score as a function of vowel/noise ratio.

tests have very nearly equal sensitivity to noise over the range of speech degradation most likely to be of interest in evaluating modern day speech processing devices. However, the Fairbanks test exhibits somewhat greater sensitivity under conditions of extreme speech impoverishment.

Figure 3 shows, however, that not all diagnostic scores are equally affected by noise. Rather, some are relatively refractory, while others are extremely sensitive. Thus while the DRT is somewhat less sensitive to noise than the Fairbanks Rhyme Test when used for the gross evaluation of intelligibility, it is substantially more sensitive when used diagnostically. One reason for this is that noise appears to have relatively little effect upon the discriminability of nasality and voicing. Even under conditions which reduce over-all intelligibility to less than 40%, listeners are able to discriminate these attributes with approximately 80% accuracy. However, noise drastically reduces the discriminability of the remaining attributes.

Extrapolation of the trends shown in Fig. 3 leads to the conclusion that sustention is the first attribute to be affected by noise, but that these effects are relatively small until the noise level approaches the level of the speech signal. Thereafter, the discriminability of this attribute decreases rapidly with increasing noise level. Low level noise appears to have relatively little effect upon the discriminability of sibilation, but over the range of speech-to-noise (i.e., vowel-to-noise) ratios below 12 dB, the discriminability of this attribute decreases sharply with decreasing speech-to-noise ratio. Both of the "place" attributes, graveness and compactness, appear to be quite sensitive to noise over the entire range of conditions investigated here.

In light of what is known of the physical bases of the various attributes, it is appropriate to ask whether, in general, both states of each attribute are equally subject to the effects of noise. The results presented in Figs. 4 through 9

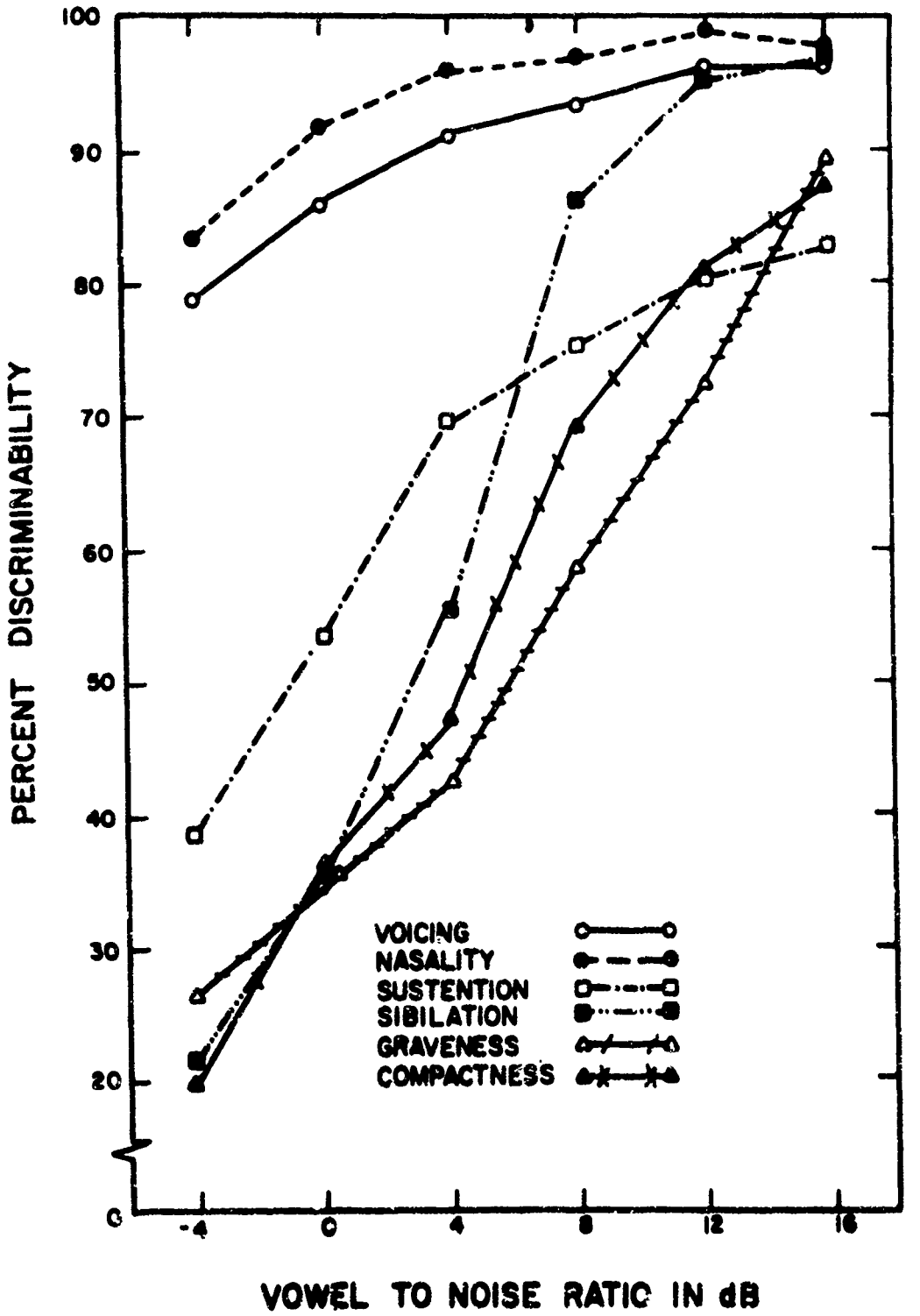


FIG. 3 Individual diagnostic scores as a function of vowel/noise ratio.

bear on this issue. Here, discriminability scores are presented separately for each state of each attribute under the various experimental conditions.

In Fig. 4 it appears that relatively low levels of noise affect the discriminability of both states of the attribute, voicing, to about the same degree. Higher levels of noise, however, have different effects upon the discriminability of the two states. As the speech-to-noise ratio falls below +4 dB, the presence of voicing becomes only slightly less discriminable. Under the same conditions, however, the discriminability of the unvoiced state is substantially reduced.

Figure 5 reveals that the effects of noise upon the discriminability of nasality are quite similar for the two states of this attribute.

The results presented in Fig. 6 are somewhat ambiguous but suggest that low levels of noise tend to bias listener responses in favor of the positive state of the attribute, sustention, while higher levels of noise induce a bias in the opposite direction.

The results presented in Fig. 7 are perhaps predictable from knowledge of the primary physical correlate of sibilation: high frequency noise. Not only does noise tend generally to reduce the discriminability of this attribute, it tends in particular to obscure the physical correlate of the positive state of this attribute. Under conditions of high noise level, listeners fail to achieve a chance level of performance in detecting the presence of sibilation. (Recall that all discriminability scores reported here have been adjusted in an attempt to remove the effects of chance or guessing. Thus an adjusted score of zero represents the chance level of performance.)

Figure 8 shows the effects of noise on the discriminability of graveness. Again, noise affects the discriminability

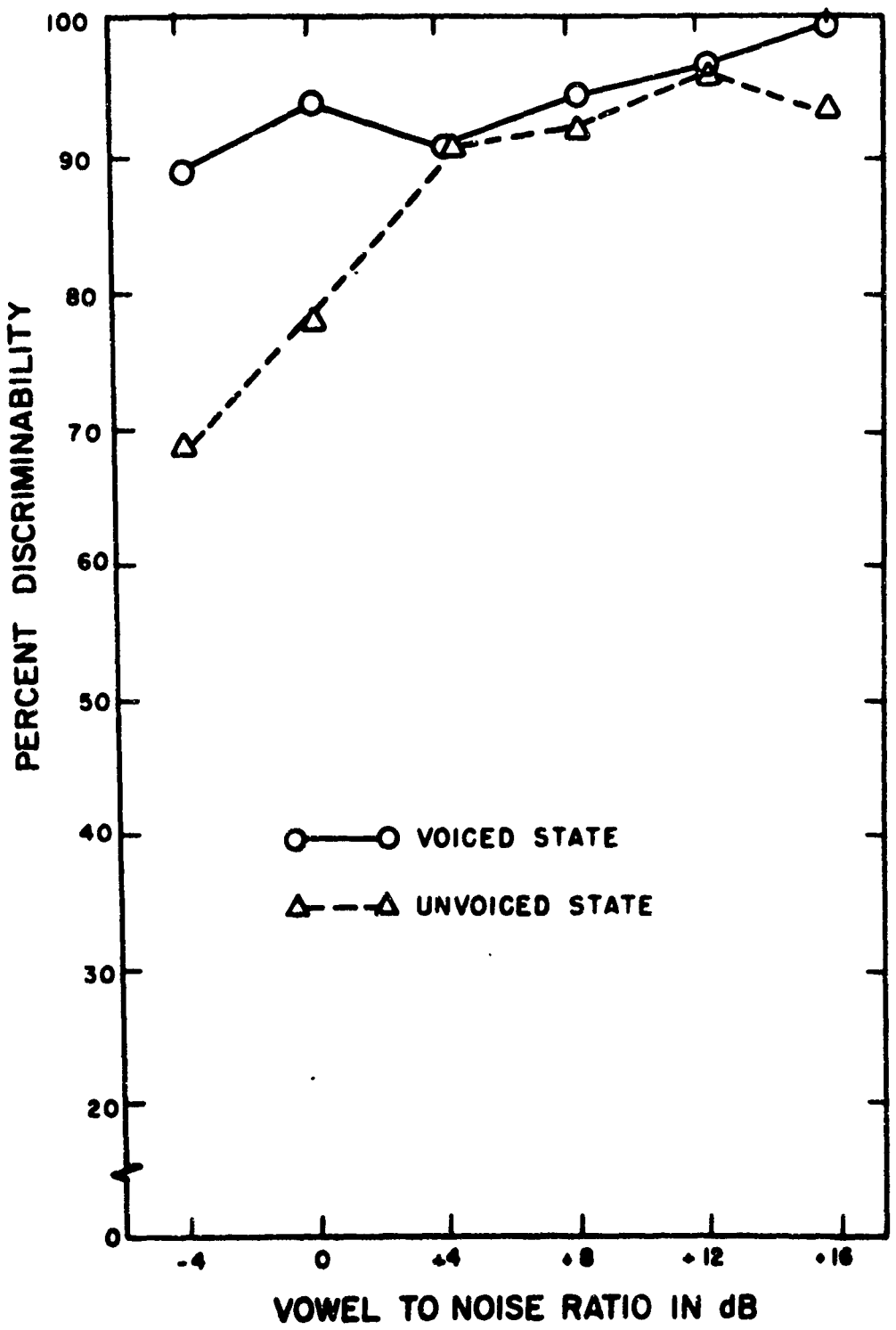


FIG. 4 The discriminability of each state of the attribute voicing as a function of vowel/noise ratio.

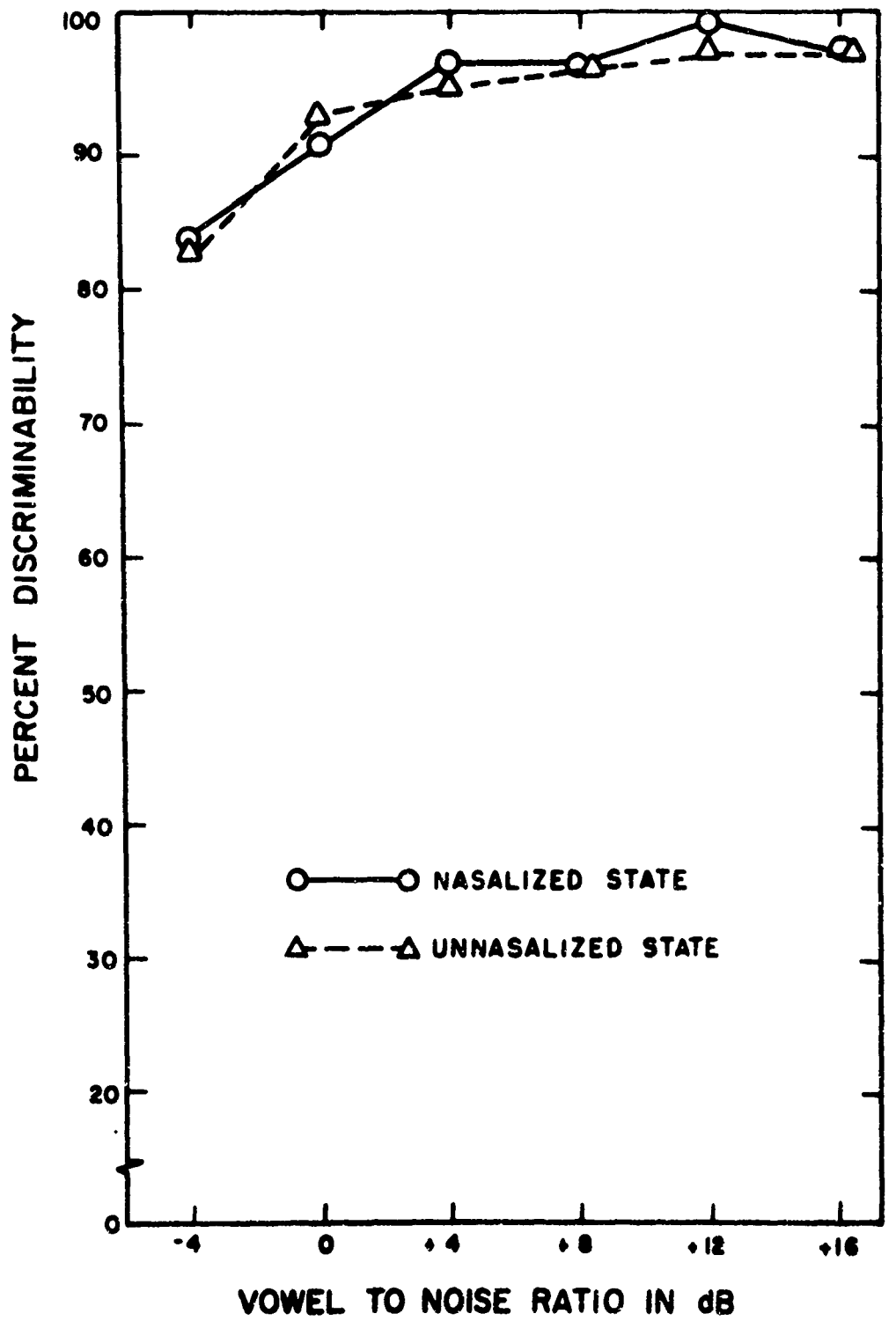


FIG. 5 The discriminability of each state of the attribute nasality as a function of vowel/noise ratio.

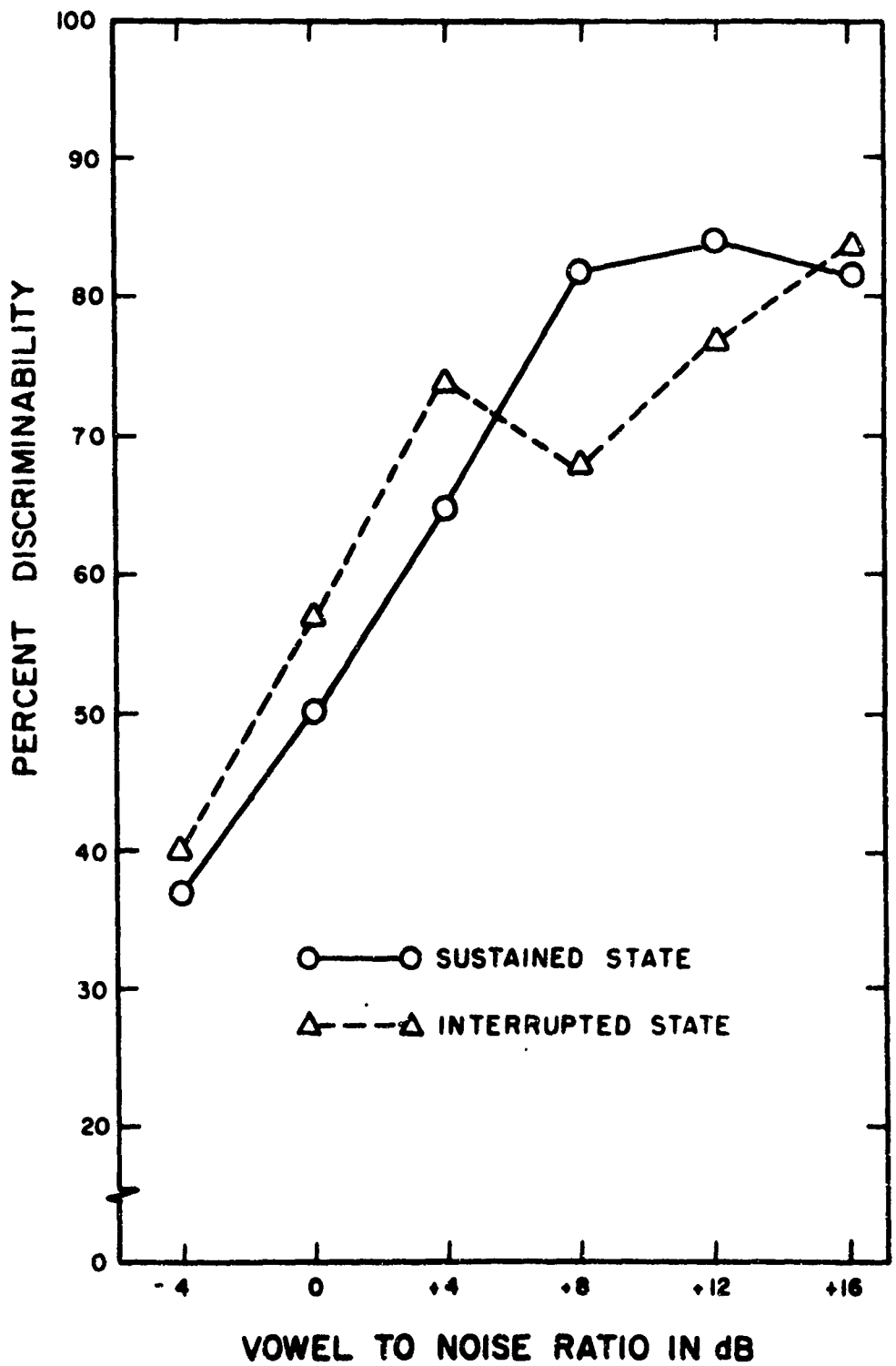


FIG. 6 The discriminability of each state of the attribute sustention as a function of vowel/noise ratio.

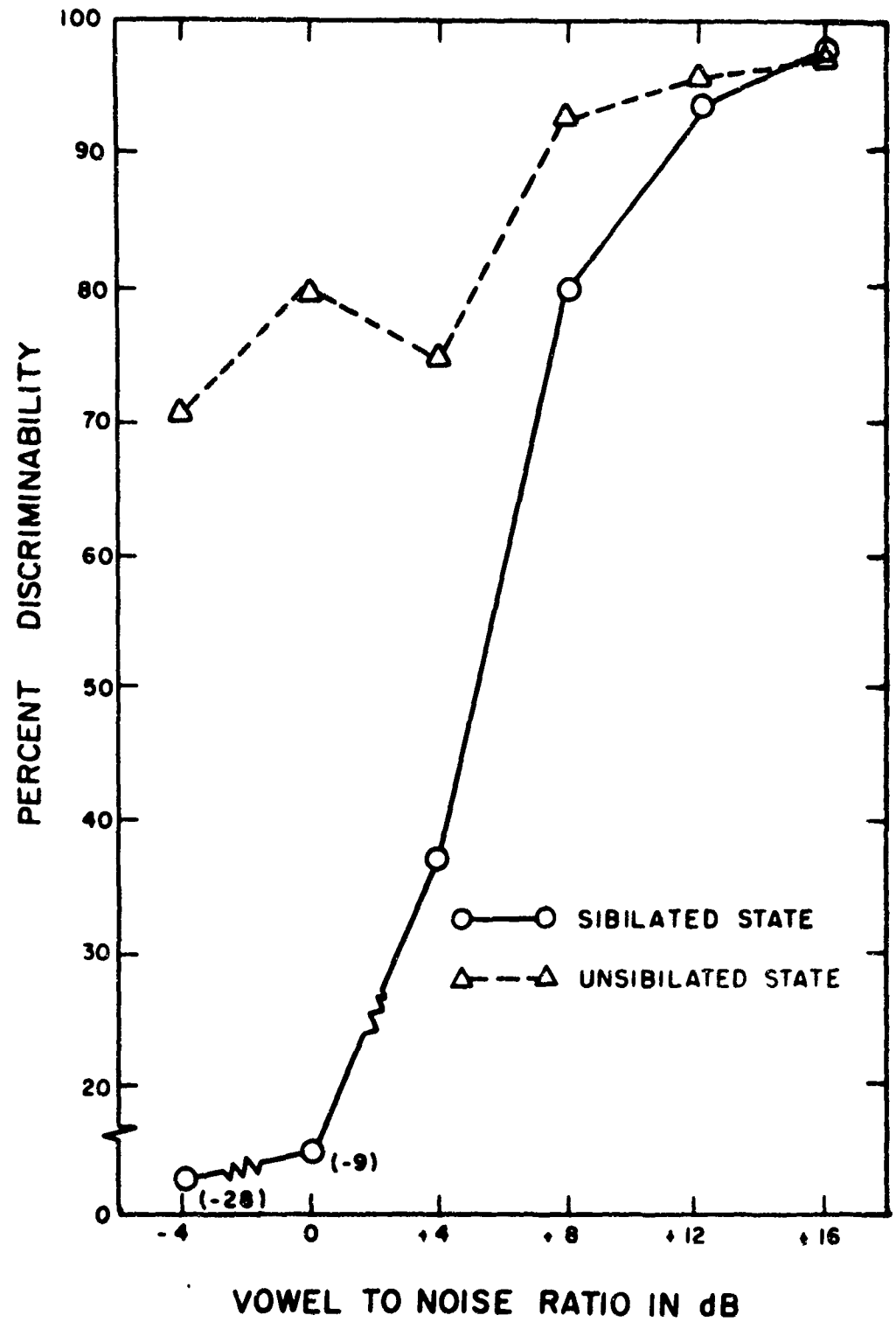


FIG. 7 The discriminability of each state of the attribute sibilation as a function of vowel/noise ratio.

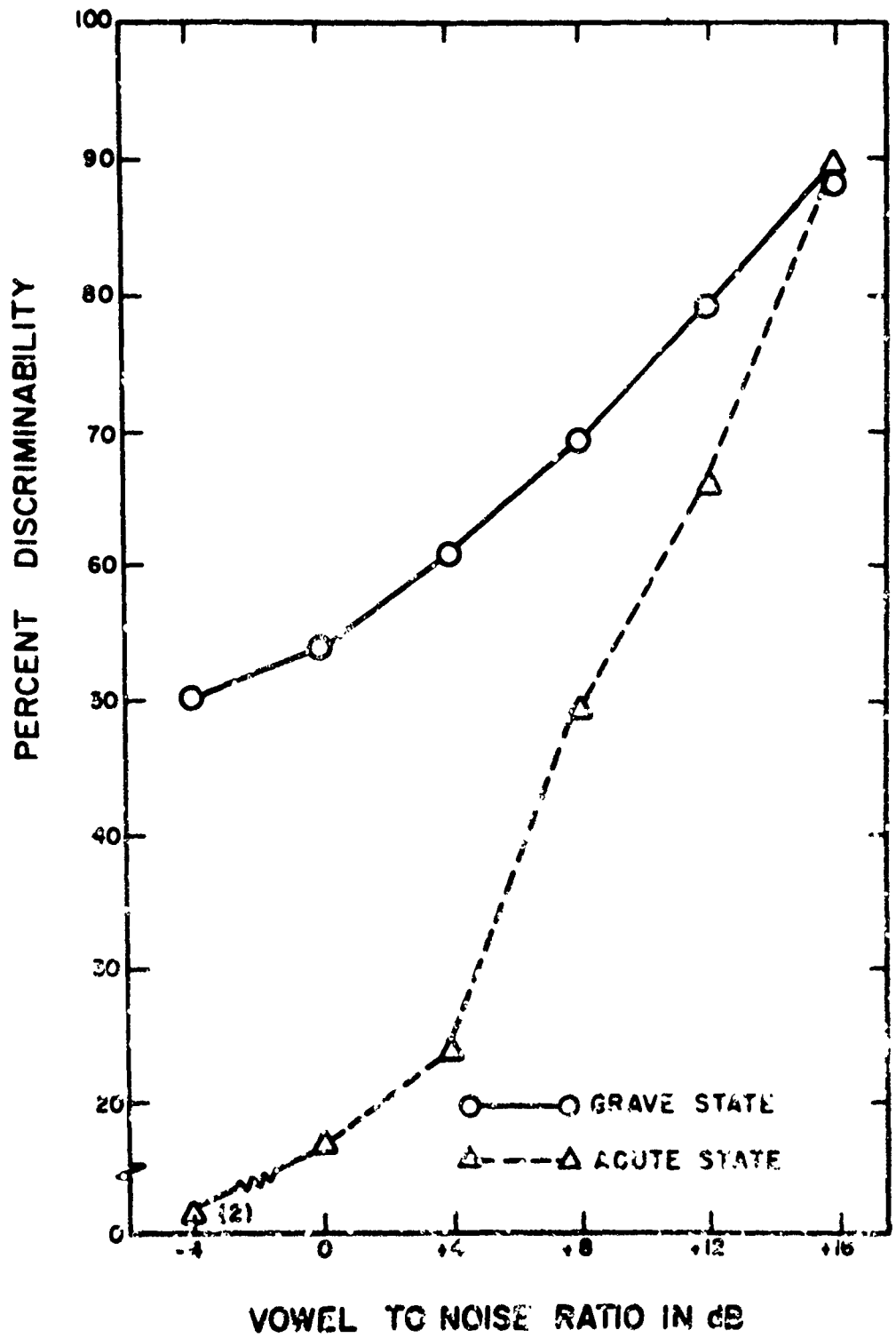


FIG. 8 Discriminability of graveness as a function of vowel/noise ratio.

of the two states of an attribute in different degrees. While it tends to obscure both states of this attribute to a significant degree, it tends in addition to bias listener responses in favor of the grave state of the response dichotomy.

Figure 9 shows that low to moderate levels of noise tend to affect the discriminability of compactness in a relatively unbiased manner. However, high levels of noise possibly bias listener responses somewhat in favor of the negative or diffuse state of this attribute.

It is of some interest that, while gross measures of listener performance reveal little difference in the effects of noise on the discriminability of graveness and compactness, separate treatment of the two states of each attribute reveals the experimental independence of these two attributes. Noise tends to bias listener perception of graveness while affecting the perception of compactness in a relatively unbiased manner.

The various results described above bear generally upon an important aspect of the validity of the DRT: the experimental independence of the six attributes for which it provides discriminability measures. The results of the following investigation also bear on this issue.

Effects of Frequency Distortion on Diagnostic Scores

Like noise, frequency distortion in some form or another is commonly encountered in modern voice communications systems. The sensitivity of the Diagnostic Rhyme Test to this type of speech degradation is thus an issue of both practical and theoretical importance. An experiment conducted in an attempt to resolve this issue involved a crew of eight experienced listeners who took the DRT under ten different frequency pass conditions. Different scramblings of the DRT materials were used for these various conditions. Nominally, five of these conditions involved high-passed speech and five low-passed

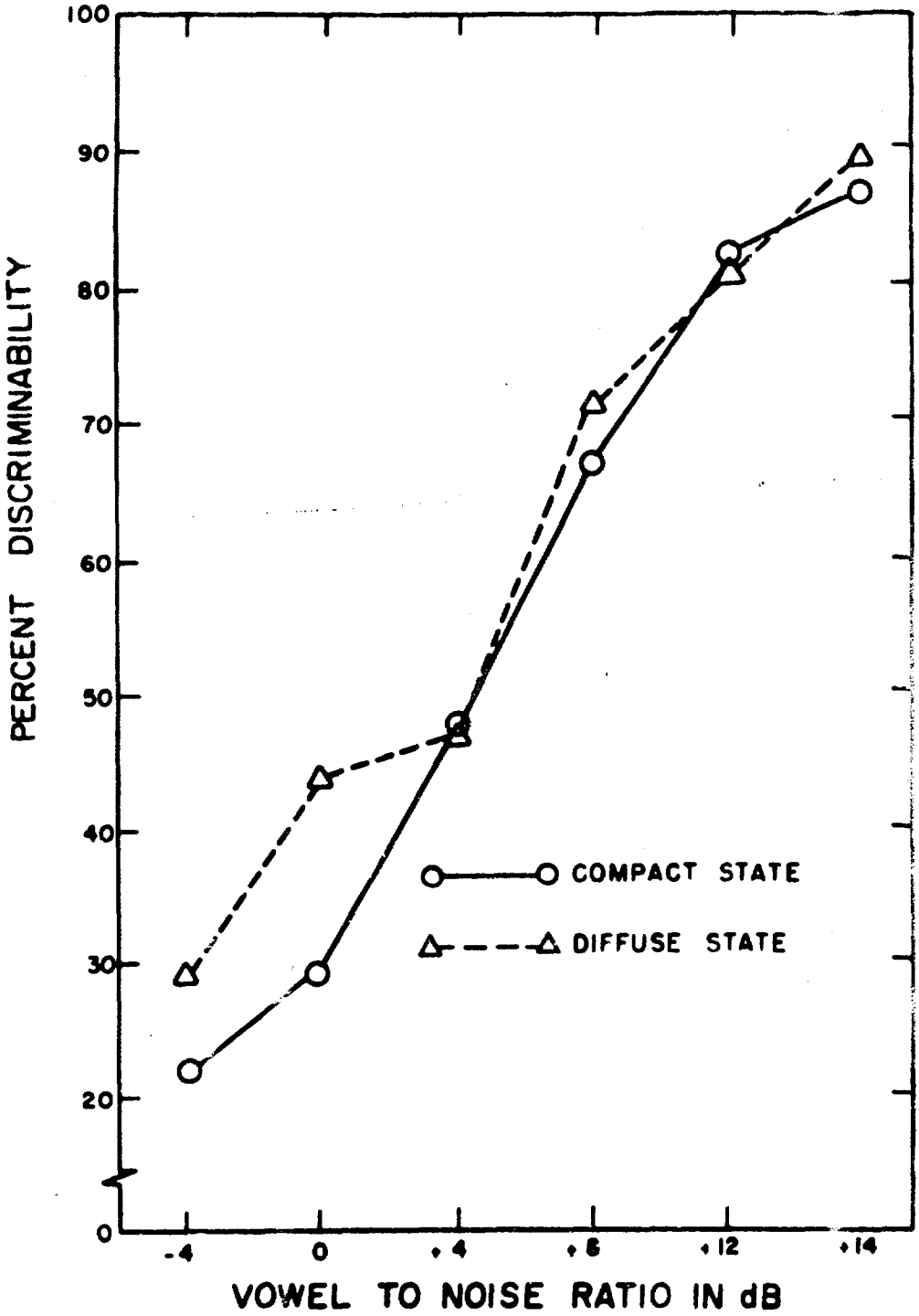


FIG. 9 The discriminability of each state of the attribute compactness as a function of vowel/noise ratio.

speech. In all of the low-pass conditions, however, the speech was also high-passed at 200 Hz. All high-passed speech was also low-passed at 4000 Hz. (The purpose of this procedure was to provide comparative data for subsequent use in evaluating the "spectral efficiency" of channel vocoders.) In all cases, frequencies beyond the nominal filter cutoff were attenuated at a rate of 42 dB/octave. The unfiltered level of speech (vowel) presentation was 80 dB re .0002 dynes/cm^2 as measured with a true rms meter. Deviations from this level were attributable only to the effects of filtering. A different scrambling of the DRT materials was used for each condition.

Figure 10 shows the effects of various degrees of two types of frequency distortion upon the DRT total score. Scores obtained with the full scale Fairbanks Rhyme Test under identical conditions are also presented. From the Figure it appears that high-pass filtering has relatively little effect upon DRT total scores over the range of conditions examined here. However, there is a consistent decrease in DRT scores with increase in the lower cutoff frequency. Predictably, perhaps, attenuation of high frequencies has much greater effects on DRT scores than low frequency attenuation. This effect becomes particularly pronounced as the upper cutoff falls below 1450 Hz.

In general, the Diagnostic Rhyme Test (total score) and Fairbanks Rhyme Test exhibit about the same degree of sensitivity to frequency distortion. However, the DRT appears to be somewhat more sensitive to extreme high frequency distortion (i.e., low-pass filtering). Some possible reasons for this are discussed in connection with the effects of frequency distortion on individual diagnostic scores.

Figure 11 shows the effects of high-pass filtering upon the six gross diagnostic scores yielded by the DRT. From the graph it appears that, over the range of conditions treated here, low-frequency distortion has little effect upon the discriminability of the various consonant attributes. The

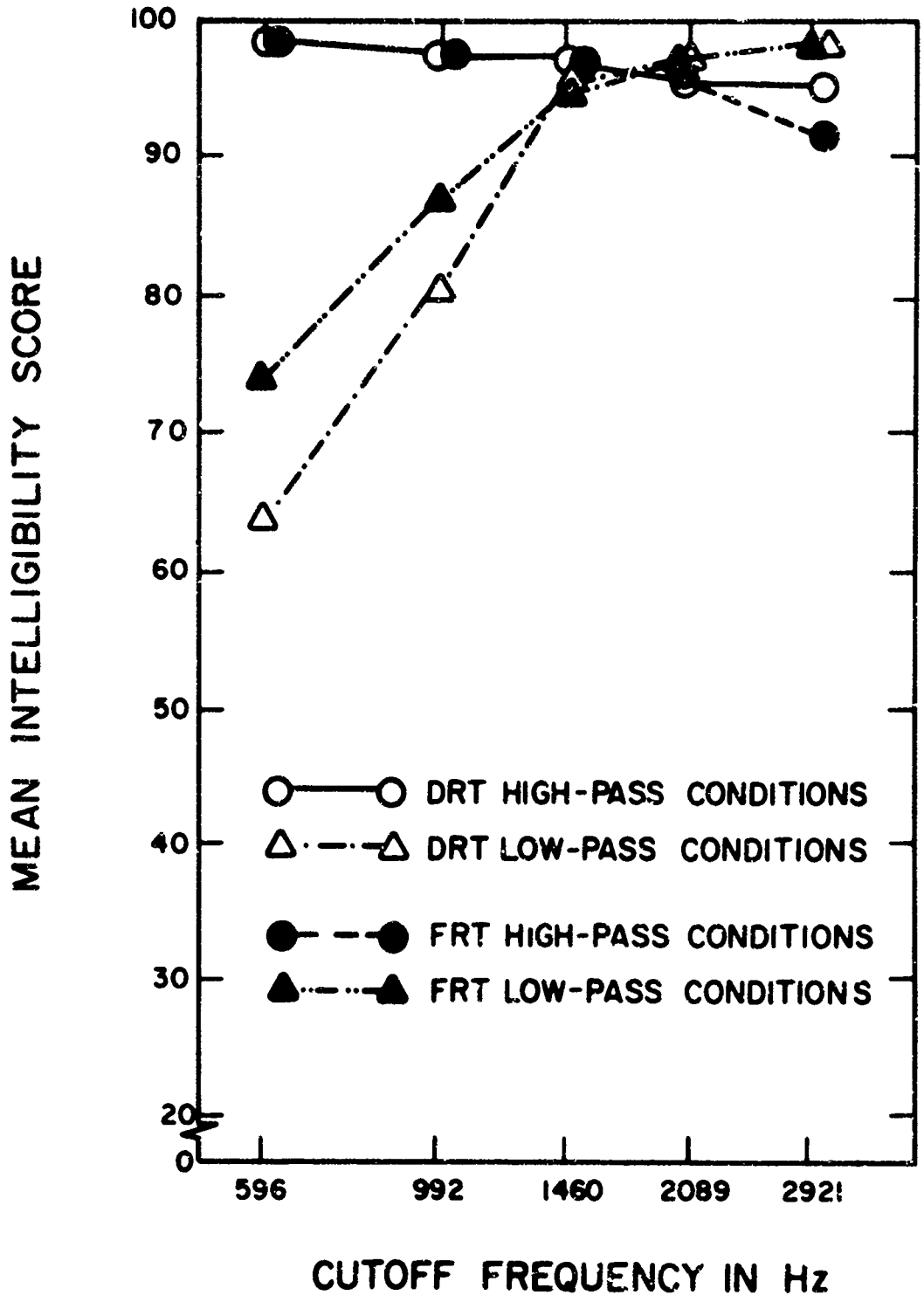


FIG. 10 Effects of frequency distortion on mean Diagnostic Rhyme Test scores and Fairbanks Rhyme Test scores.

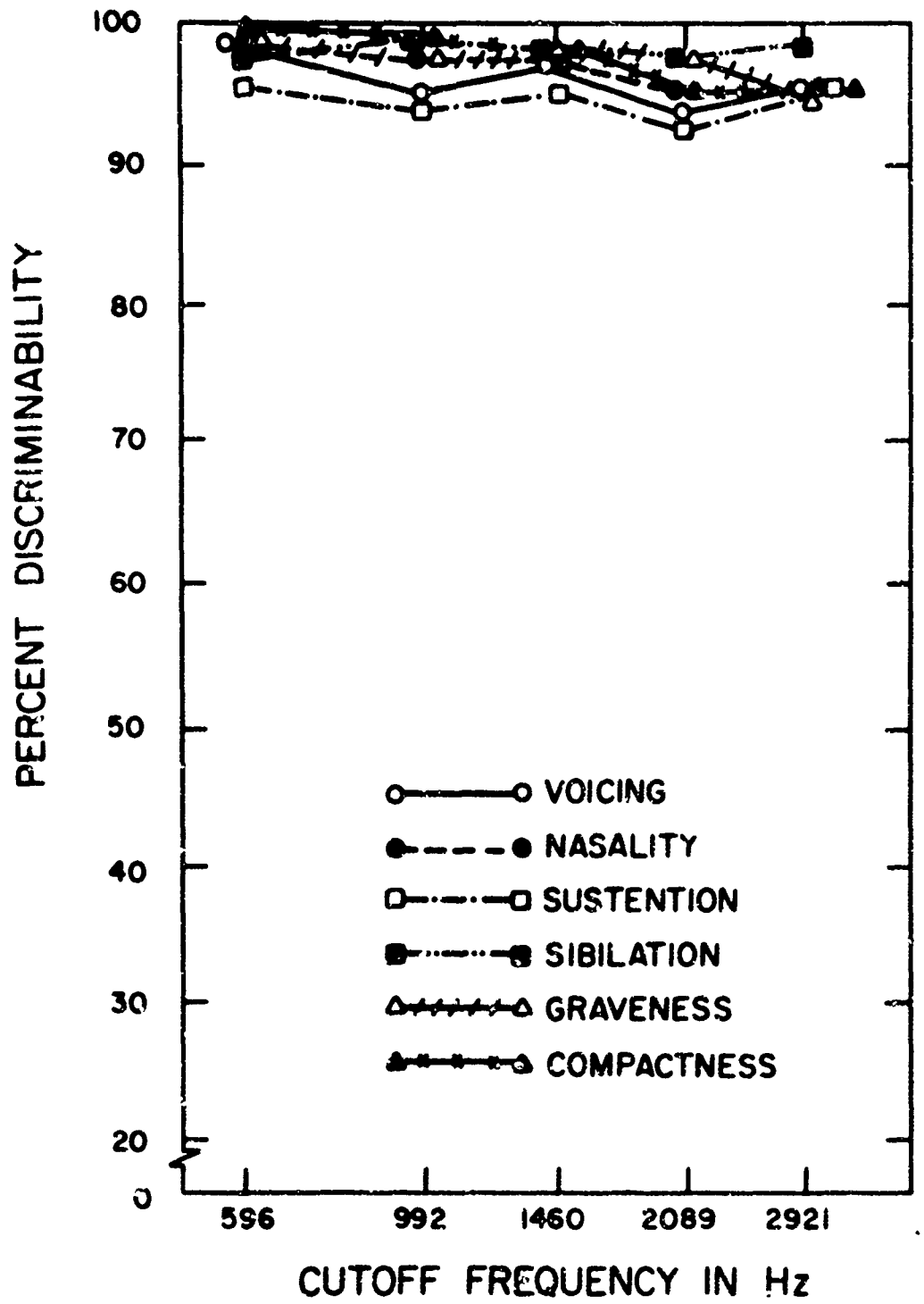


FIG. 31 Effect of high-pass filtering upon gross diagnostic scores.

discriminability of sibilation is virtually unaffected under any condition, though the remaining scores exhibit a slight decreasing trend with increases in the low-frequency cutoff. Quantitatively, these results are somewhat at variance with what might have been predicted from the results of Miller and Nicely.[18] Qualitatively, however, they are more in line with the latter in that, for example, sibilation (the near equivalent of the duration feature of Miller and Nicely) is the least affected by this form of speech impoverishment.

Miller and Nicely suggest that, while high-frequency distortion tends to induce systematic errors of listener response, low frequency distortion is more condusive to random errors. Restriction of the listener response options, as in the case of the DRT, might thus be expected to affect listener performance differently under these two conditions.

Figure 12 shows the effect of high-frequency distortion on gross diagnostic scores. These results are more nearly in line with those of Miller and Nicely. Voicing and nasality, remain quite discriminable over the range of conditions investigated here, though consistent decreases in scores for these attributes are evident as the upper cutoff frequency is decreased.

Sustention (which corresponds most nearly to "affrication," as used by Miller and Nicely) is moderately affected by high frequency distortion whereas sibilation and the two "place" attributes exhibit greatly reduced discriminability with attentuation of frequencies above 1460 Hz.

Further insights are provided when the effects of frequency distortion are examined separately for each state of the six consonant attributes. This is permitted by Figures 13 through 18.

From Fig. 13 it is evident that frequency distortion has relatively little effect upon the discriminability of either

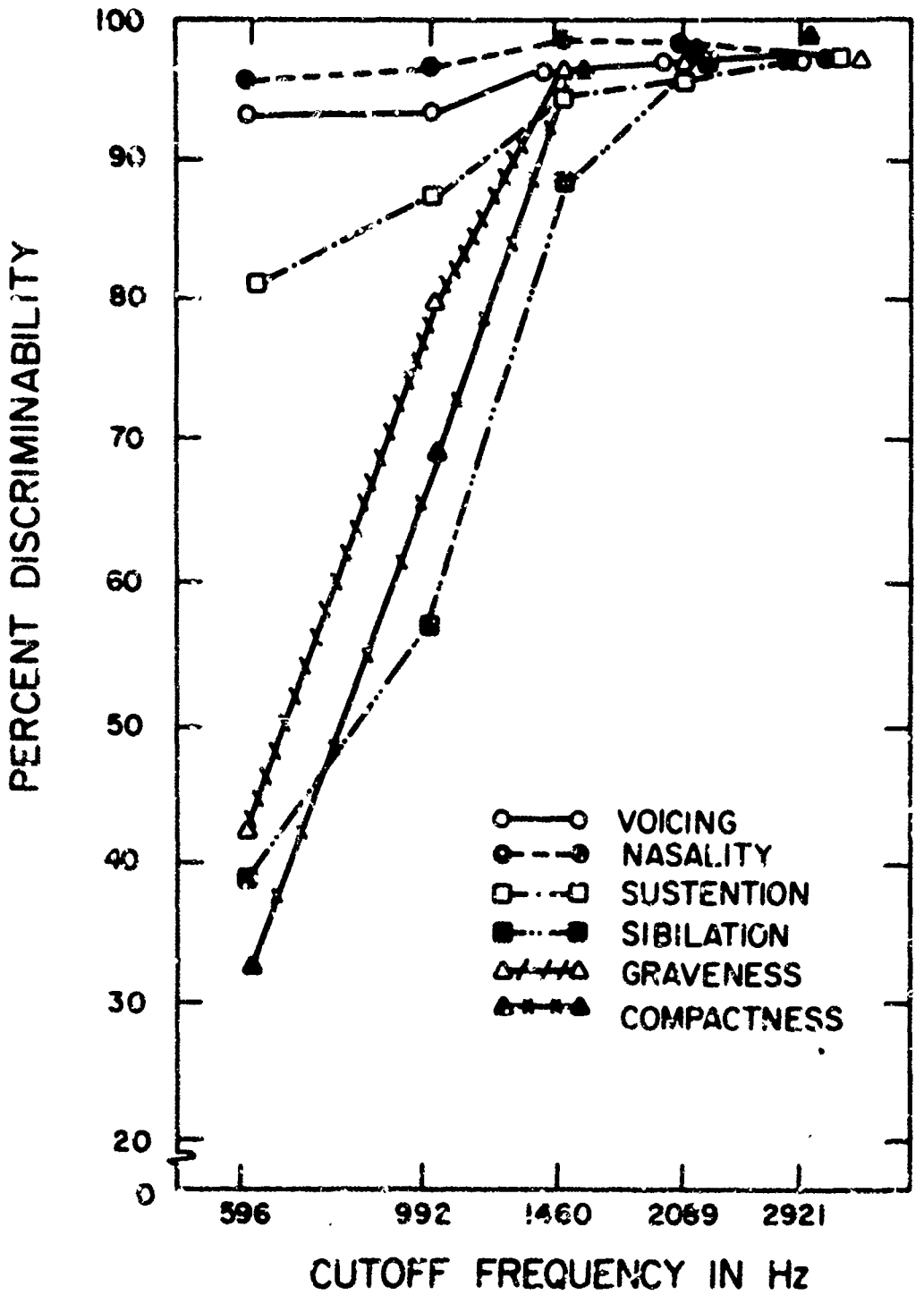


FIG. 12 Effect of low-pass filtering upon gross diagnostic scores.

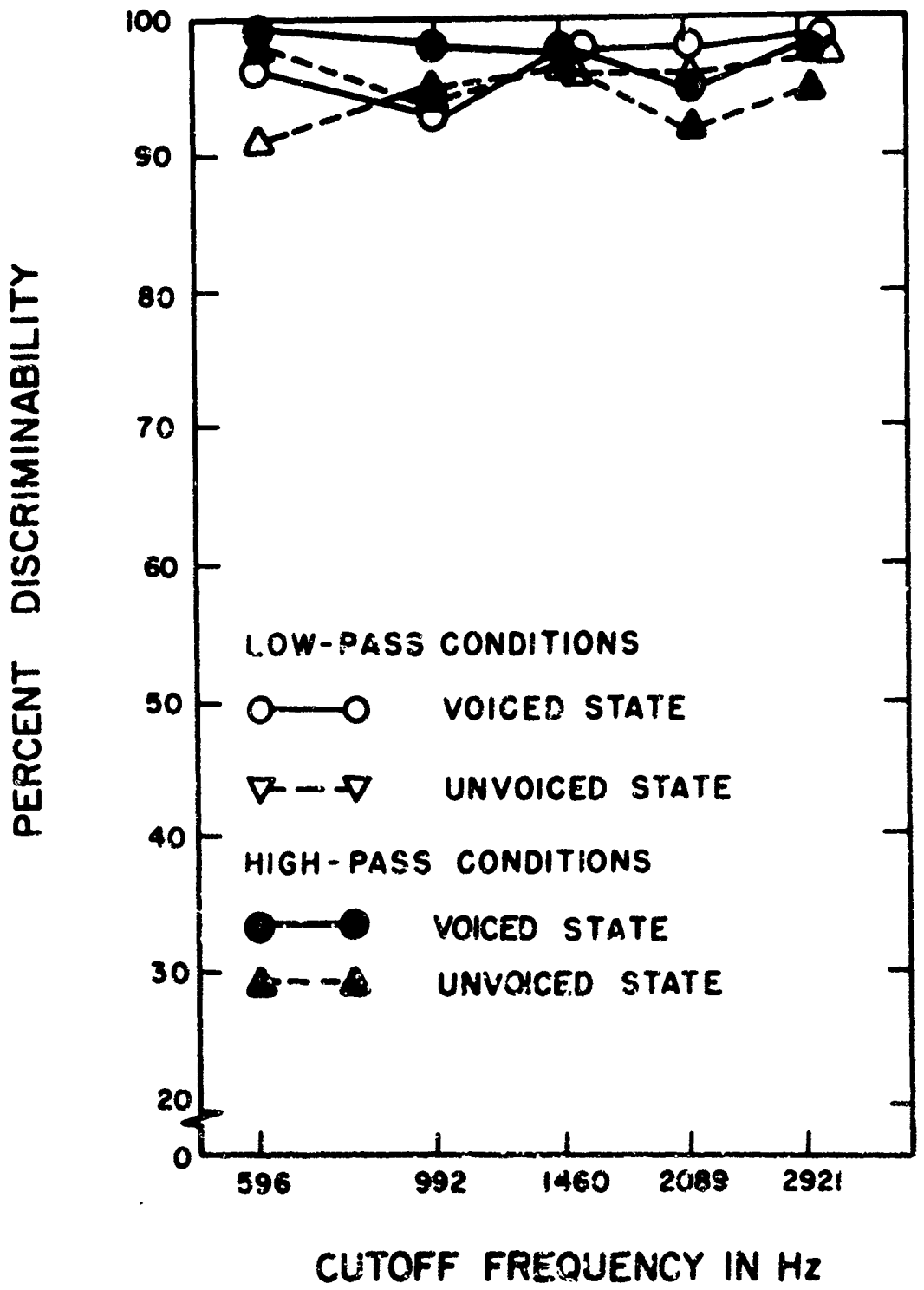


FIG. 13 Effects of frequency distortion on the discriminability of voicing.

state of the attribute, voicing, but there is some indication that high frequency attenuation tends to bias listener responses to this attribute. As the upper cutoff is decreased, listeners tend somewhat to attribute voicing to unvoiced sounds.

There is a slight indication in Fig. 14 that the two types of frequency distortion tend to bias listener response to this attribute in opposite directions. As the upper cutoff frequency is lowered, listeners exhibit increasing failures in detecting the presence of nasality; as the lower cutoff frequency is raised, listeners tend increasingly to attribute nasality to non-nasal speech sounds.

Figure 15 shows the effects of frequency distortion for the two states of the attribute sustention. Although, generally, high-pass filtering has a rather small effect in the case of this attribute it appears to bias listener responses in favor of the positive state of this attribute. However, the discriminability of this attribute decreases consistently with decreases in the upper frequency cutoff, which effect becomes especially pronounced as the cutoff falls below 1460 Hz. This latter result may account in part for the observed differences between the DRT and the Fairbanks Rhyme Test in terms of sensitivity to high frequency attenuation. Of the 250 items comprising the Fairbanks Test, approximately 45 are potentially available to reflect transmission deficiencies with respect to the attribute, sustension, (i.e., for 45 items there are acceptable response words which differ from the stimulus wo only with respect to sustention). The results obtained for the case of sibilation, also bear on this issue.

Figure 16 shows in detail, the effects of frequency distortion on the discriminability of sibilation. It appears that low frequency attenuation has a negligible effect over the range of conditions studies here. Predictably, however, attenuation of the higher frequencies drastically affects the discriminability of this attribute generally and the detectability of its positive state in particular. These results

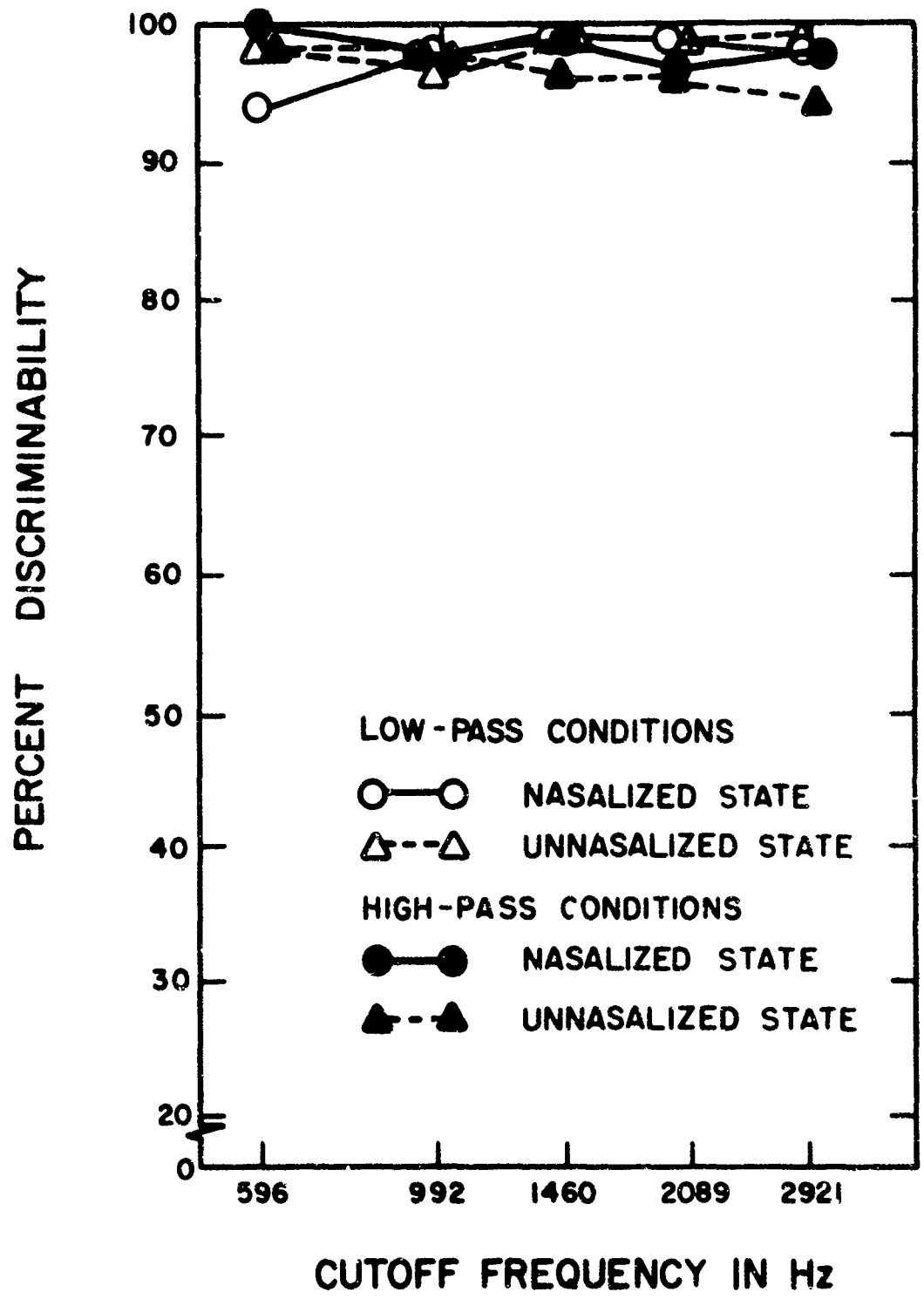


FIG. 14 Effects of frequency distortion on the discriminability of nasality.

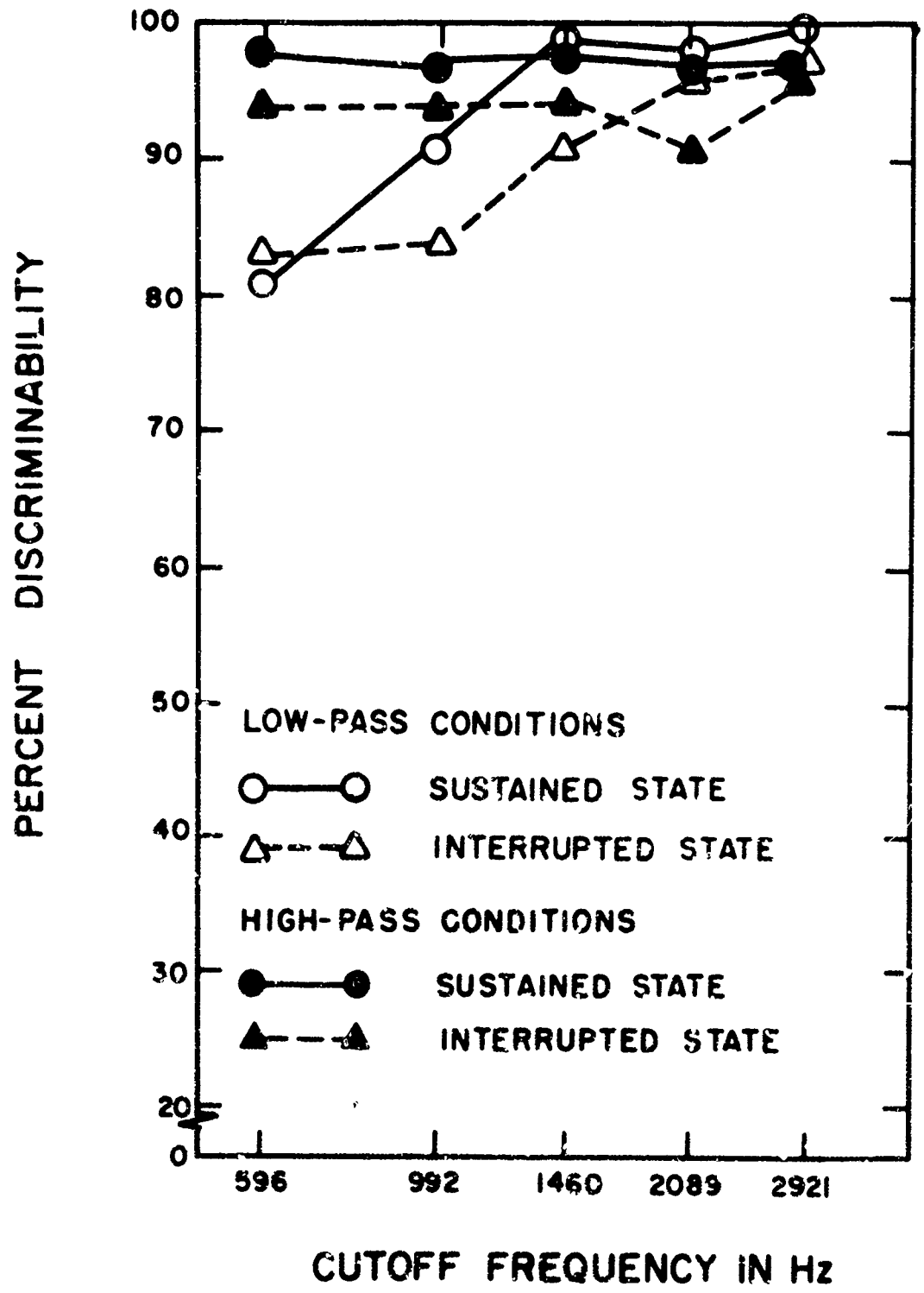
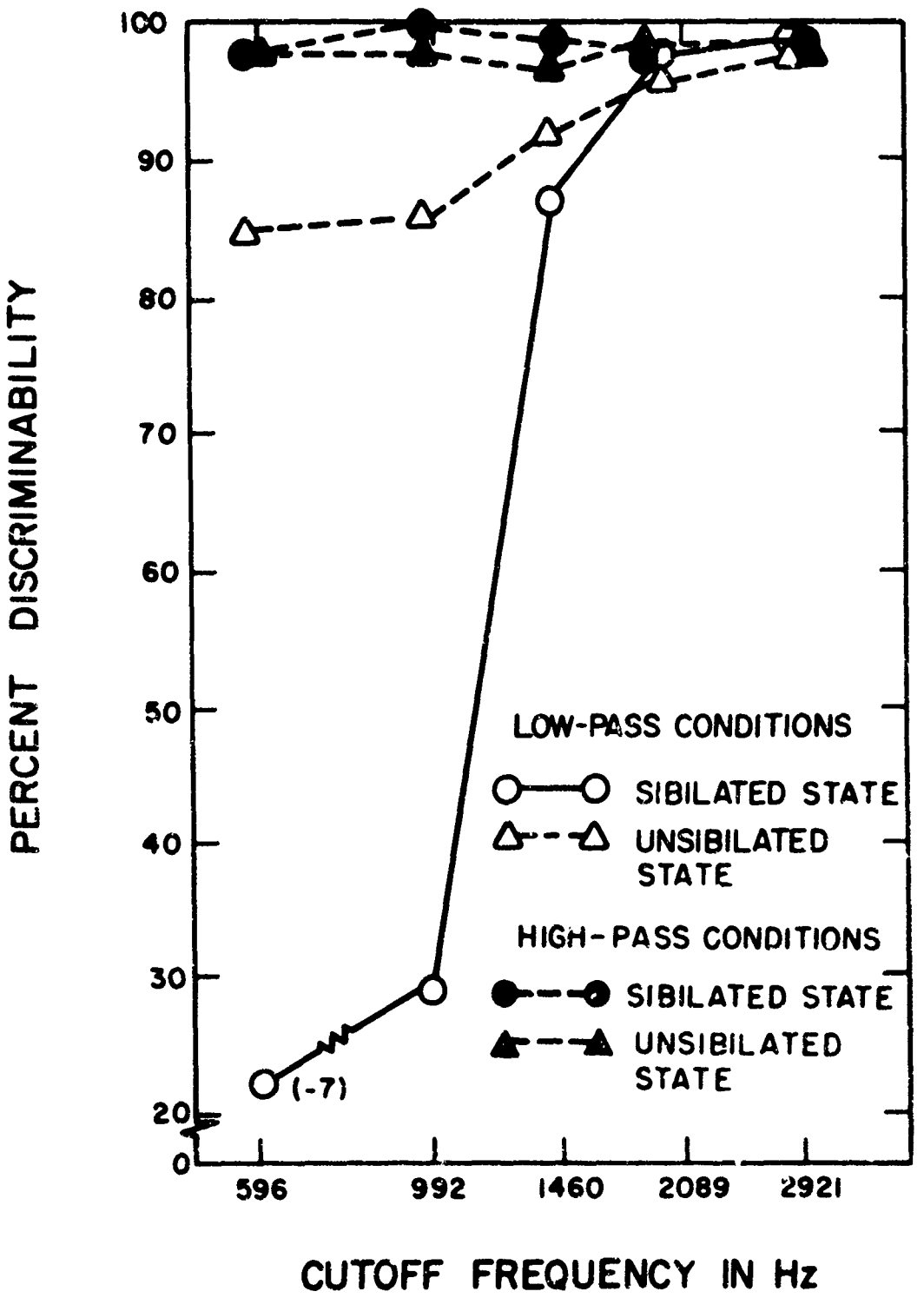


FIG. 15 Effects of frequency distortion on the discriminability of sustention.



on,

FIG. 16 Effects of frequency distortion on the discriminability of sibilation.

also suggest an additional basis for the difference between the DRT and the Fairbanks Rhyme Test in terms of sensitivity to high frequency attenuation. The Fairbanks Rhyme Test contains no more than three items on which a listener's failure to discriminate sibilation can alone lead to an erroneous response. Thus, systems that are particularly deficient in the capacity to transmit this attribute will usually yield lower DRT total scores than FRT scores.

As in the case of several other attributes, high-pass filtering leads to a slight but consistent decrease in the discriminability of graveness as the lower cutoff frequency is increased. In Fig. 17 there appears at the same time to be a gradually increasing bias in favor of the acute state of this attribute. The effects of high frequency attenuation are more pronounced. While the attenuation of frequencies above 1460 Hz has little effect in the case of this attribute, discriminability is sharply reduced as the upper cutoff frequency is lowered beyond this point. At the same time there is an increasing bias in favor of the positive or grave state of this attribute. Listeners tend increasingly to mistake acute phonemes for their grave cognates.

As in the case of the other "place" attribute, low frequency attenuation has a small, but consistent effect upon the discriminability of compactness, but Fig. 18 also reveals an increasing bias with increasing lower cutoff point. Listeners tend increasingly to favor compactness as low frequency attenuation becomes more extensive.

The foregoing results demonstrate the sensitivity of the DRT to a commonly encountered form speech degradation and also, perhaps, provide some minor insights concerning the spectral distribution of phonemic identity information. In any case, they provide additional evidence of the experimental independence of the various diagnostic scores.

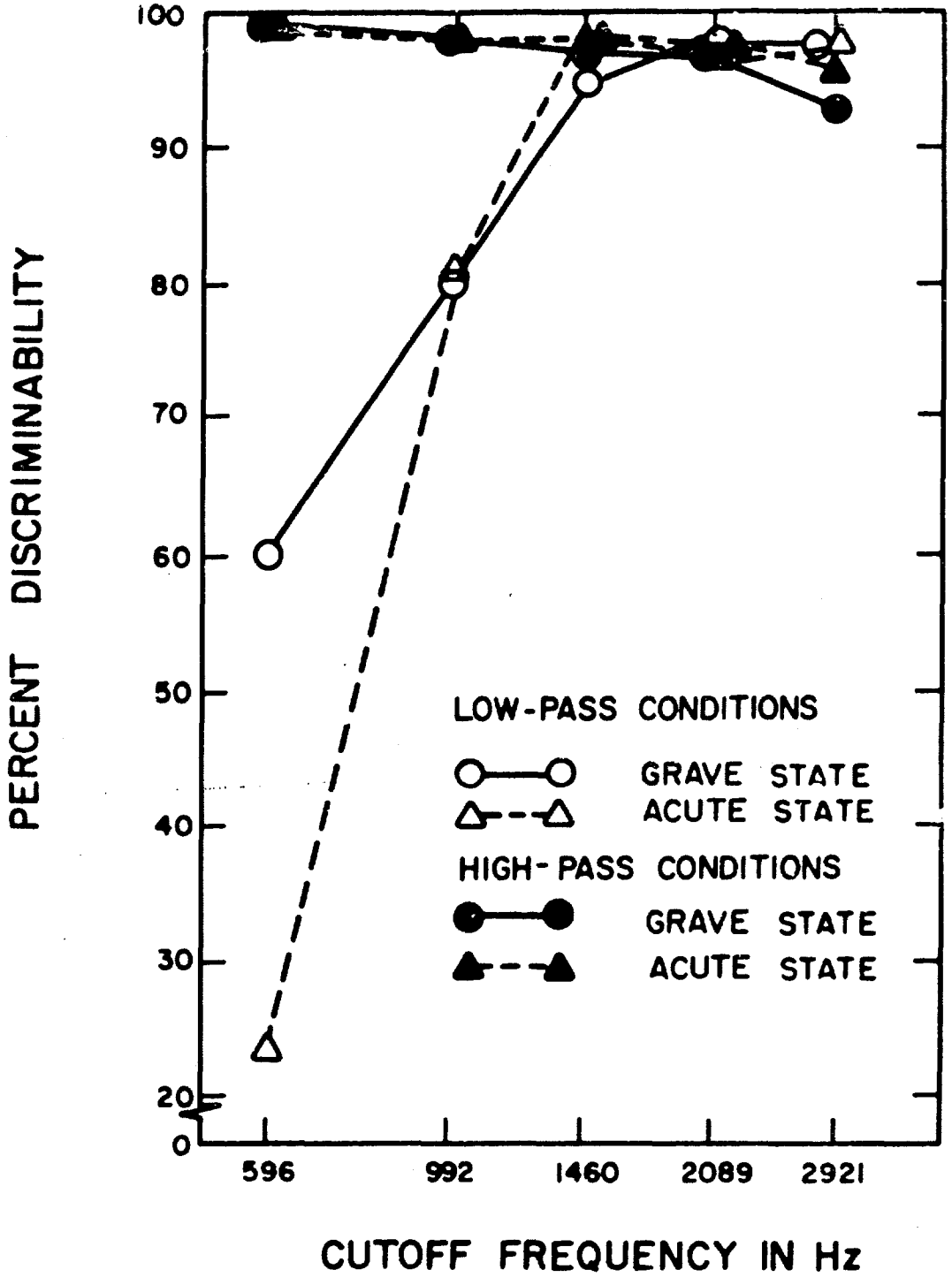


FIG. 17 Effects of frequency distortion on the discriminability of graveness.

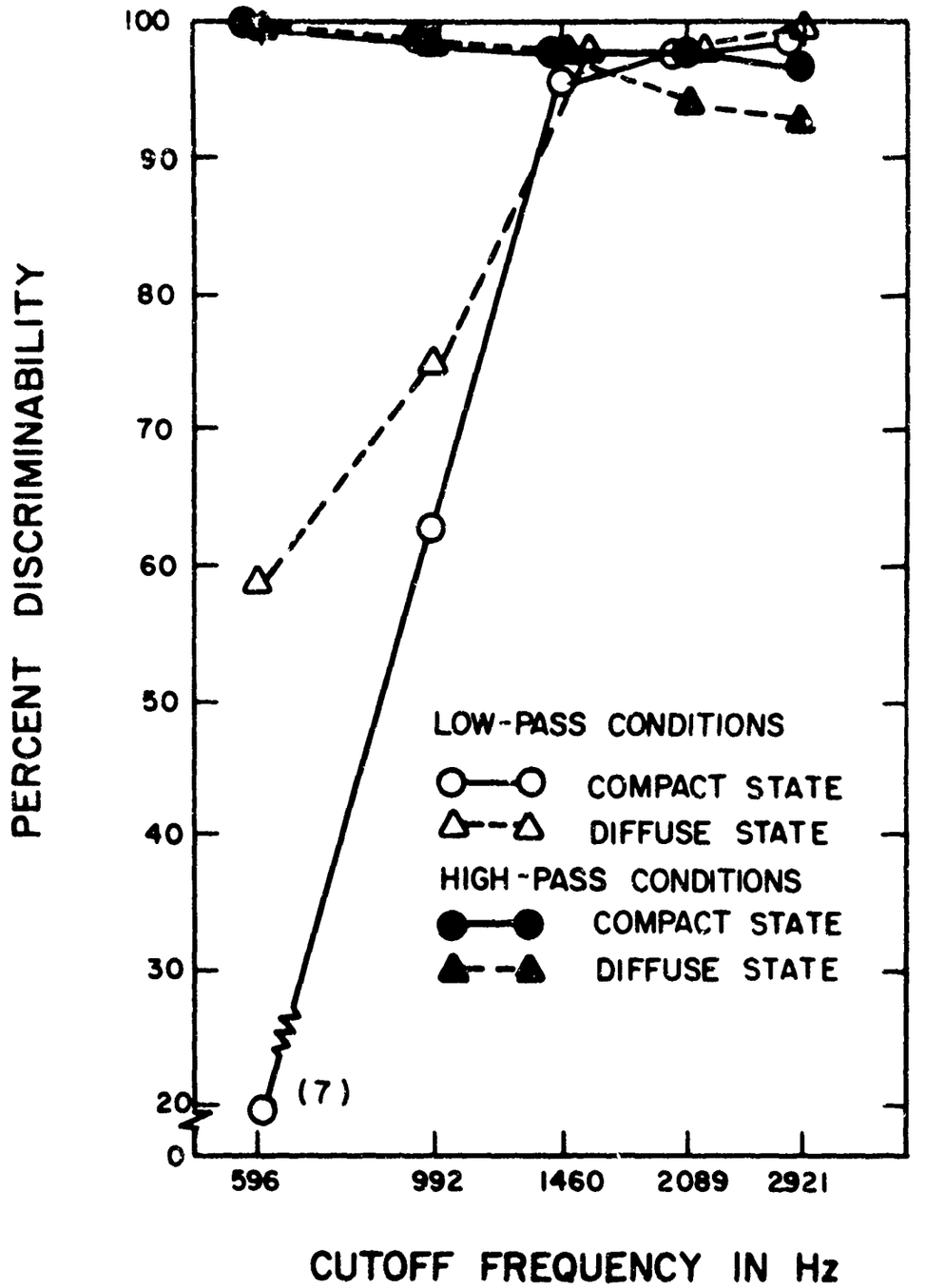


FIG. 18 Effects of frequency distortion on the discriminability of compactness.

Effects of Speech Processing Devices

Form III of the DRT has yet to be used extensively for routine evaluation of voice communications systems, though Form I has been used in more than 150 evaluations of vocoder systems alone. On recalling that Forms I and II of the DRT are (with certain scoring modifications), formally equivalent to Form III, one may reasonably use results obtained with earlier versions of the test to make inferences concerning Form III. Accordingly the following are examples of the kinds of results that may be expected where Form III is used in evaluating vocoder systems.

One study has shown that the voicing scale of the DRT is uniquely sensitive to deficiencies in the voicing detectors of channel vocoders. In this study, the voicing switch of a channel vocoder was "locked out." Listener scores on the voicing scale were reduced approximately 30%. No other scores were affected.

It has been frequently observed that the digitalization of vocoded speech acts primarily to depress DRT scores on the scale now identified by the term, sustention. Differences of the order of sixty points on the Form I equivalent of this scale are not exceptional in comparisons between digital vocoders and their analog counterparts. Still other examples of the uses of the DRT for purposes of system evaluation have been described elsewhere,[19] and need not be recounted here.

One important issue, concerning validity of the DRT as an instrument for system evaluation remains to be considered. This is the issue of the validity of the DRT total score as a gross indicant of system performance. While, a priori, various theoretical considerations might lead one to question the usefulness of such a score (in light of the fact, for example, that it gives equal weight to each of six aspects of systems performance which are undoubtedly of unequal importance for overall intelligibility) the available experimental evidence argues for an affirmative answer to this question.

Figure 19 shows a plot of Fairbanks Rhyme Test score against Diagnostic Rhyme Test (Form I) total score for a sample of vocoded speech conditions. A substantial correlation is evident. Moreover, an examination of the diagnostic scores obtained in deviant cases has revealed that these cases usually involve systems which are particularly deficient with respect to sibilation; a deficiency to which the Fairbanks test appears to have quite limited sensitivity.

The limited data available thus far indicate that Form III of the DRT is also highly correlated with the Fairbanks Rhyme Test and tends generally to yield scores more nearly numerically equal to those of the FRT than is the case with Form I of the DRT. The fact that Form I gives effectively double weight to compactness (an attribute that is particularly sensitive to vocoding) possibly accounts for its tendency to yield somewhat lower scores than both Form III and the Fairbanks Rhyme Test when used in evaluating speech compression systems.

Sensitivity of the DRT to Characteristics of the Speaker

Just as speech processing or transmission systems may vary in the fidelity with which they transmit the various criterial attributes of consonant phonemes, speakers may vary in the precision with which they produce the acoustical correlates of these attributes. While differences among normal speakers in this respect are not highly pronounced in situations involving undegraded speech, they are usually amplified by various types of speech degradation.

Vocoding, in particular, tends to emphasize differences among speakers, and these differences are readily detected by means of the Diagnostic Rhyme Test. The results of a study involving vocoded speech of eight different male speakers are relevant in this connection. Of the eight speakers, one was a trained speaker (former radio announcer); the other seven were selected on the basis of listener ratings which they had

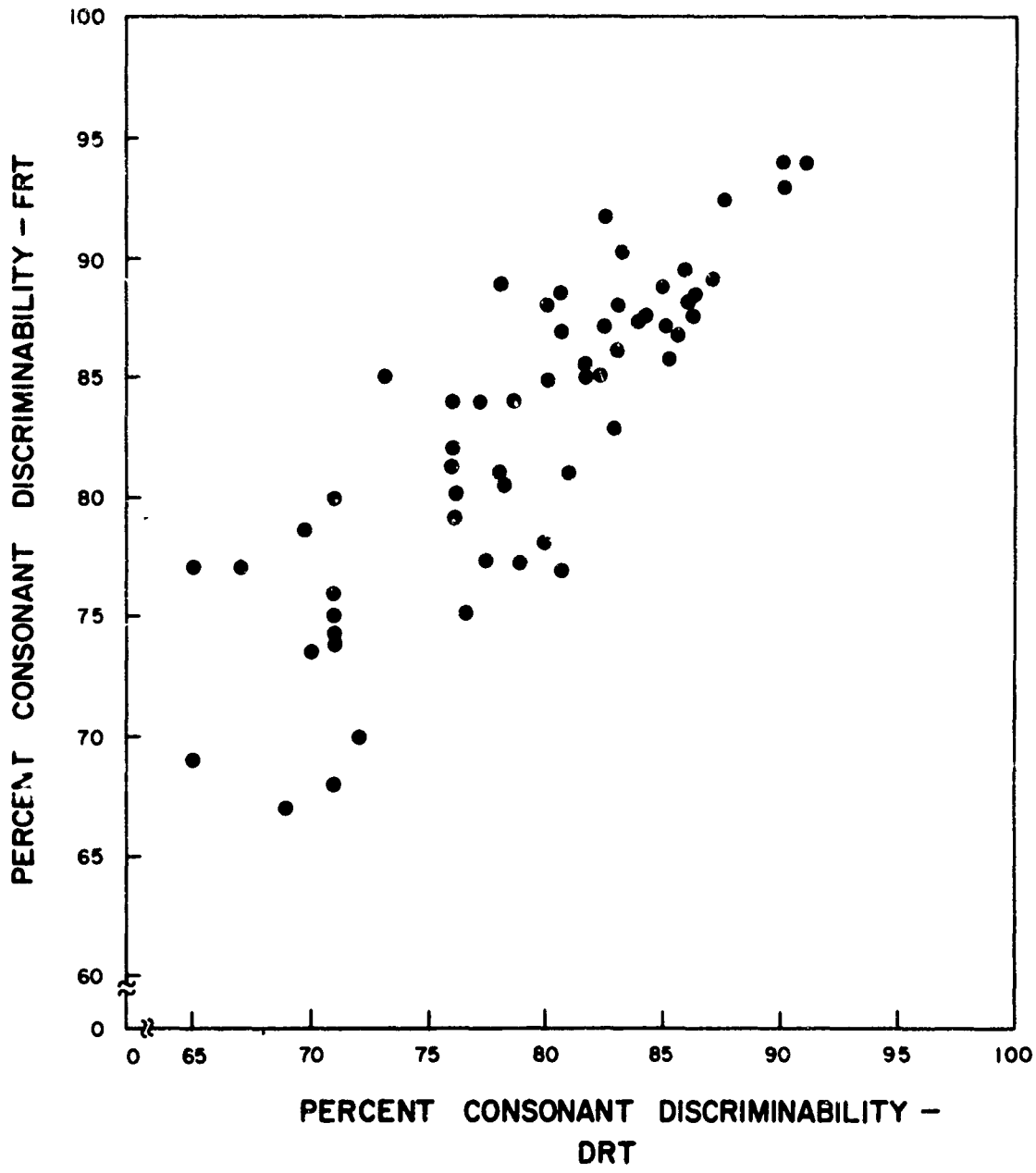


FIG. 19 Scatter plot of Fairbanks Rhyme Test scores vs Diagnostic Rhyme Test scores for a sample of vocoder systems.

received on each of four perceived acoustic traits.[20] DRT recordings by each speaker were processed by each of four different but similar experimental configurations of a modern analog channel vocoder. Various scramblings of the basic test materials were employed. Recordings of the processed materials were then administered to a crew of eight male listeners. The results are summarized in Table 5 where the range and diversity of speaker differences are evident. These results might be taken to suggest some of the articulatory bases of various perceived voice characteristics. However, it should be recalled that the scores shown here may depend not only upon the inherent characteristics of the speaker but also upon the manner in which these characteristics interact with the process of vocoding. Collectively, however, they serve to illustrate the sensitivity of the various DRT scores to individual speaker differences in speech.

A somewhat more dramatic illustration of this sensitivity is provided by the case of a speaker with a hearing loss and an associated speech difficulty which, however, had been substantially reduced as a result of intensive therapy. An audiogram of this speaker's best ear reveals a loss of approximately 15 dB for frequencies below 1000 Hz; no sensitivity to frequencies of 2000 Hz or higher could be demonstrated.

A recording of the DRT Form III test materials by this speaker was presented to an experienced crew of listeners with normal hearing. The results are presented in Table 6. While any attempt at a detailed interpretation of this pattern of diagnostic scores would be premature at this point, the pattern is generally consistent with what might be predicted from the audiometric data for this speaker. Gross scores for the various attributes are not seriously depressed, but some significance perhaps attaches to the pronounced bias which is evident in all cases but those of voicing and nasality. Relatively poor discriminability of sibilants produced by the

TABLE 5. DIAGNOSTIC SCORES FOR VOCODED SPEECH BY EIGHT SELECTED SPEAKERS

	SPEAKER								
	Trained	Neutral	Low	High	Smooth	Rough	Clear	Unclear	Mean
Voicing*	97	95	81	94	96	85	95	93	92
Nasality	96	98	93	94	96	94	95	98	95
Sustention	90	94	92	91	88	94	92	93	92
Sibilation	99	90	95	94	96	91	98	92	94
Graveness	88	92	83	86	86	78	89	88	88
Compactness	98	97	95	91	92	89	96	95	94
Mean	95	94	90	92	92	89	94	93	93

*These results were obtained with Form I of the DRT, however, the various diagnostic scores are labeled with their Form III equivalents. Standard errors of approximately 2.0 are typical for average diagnostic scores. Standard errors of 0.5 are typical for the averaged total scores.

TABLE 6. DIAGNOSTIC SCORES FOR A SPEAKER WITH HIGH-FREQUENCY HEARING LOSS

Attribute	Positive State	Negative State	Average
Voicing	93	87	90
Nasality	99	100	99
Sustention	88	96	92
Sibilation	79	97	88
Graveness	98	86	94
Compactness	99	83	91
Mean			92

speaker is particularly consistent with the available audiometric data as well as with data on this individual's DRT performance as a listener.

Sensitivity of the DRT to Characteristics of the Listener

Early in the history of the DRT it became apparent that the test was sensitive in various ways to individual differences in discriminatory capacity. In fact, such differences served on occasion as the basis for selection of listeners, particularly in instances where pure-tone audiometric criteria did not provide a clear-cut basis for decision. On other occasions, DRT scores have isolated individuals with histories of suspected auditory perceptual difficulty where routine audiometric tests revealed no deficiency. Such instances as these seemed clearly to warrant further investigation of the potential of the DRT for purposes of diagnostic evaluation of the auditory perceptual characteristics of listeners. The use of the DRT itself as a selection device, of course, operated to minimize individual differences within the listening crews used for routine experimental purposes, particularly, in situations involving undegraded speech. However, where impoverished speech is involved, stable, significant individual differences become quite evident.

In one study, a crew of six previously screened listeners was repeatedly tested with a set of vocoded DRT materials, a different scrambling of the materials being used for each test. The obtained data were subjected to analysis of variance. The results of this analysis are summarized in Table 7. A highly significant F-ratio for the listener effect is to be seen in all cases except sustention, indicating that an average of the results of four or more tests is sufficient to provide a reliable indication of an individual's discriminative performance with respect to five of the six elementary consonant attributes. There remains, however, the question of the nature of the listener factors underlying such performance differences.

TABLE 7. LISTENER DIFFERENCES IN DRT PERFORMANCE:
RESULTS OF ANALYSIS OF VARIANCE FOR INDIVIDUAL ATTRIBUTES

Attribute	Source	SS	df	MS	F*	P
Voicing	Listeners	46.37	5	9.27	11.42	<0.01
	Scramblings	7.12	3	2.37	2.94	ns
	LXS	12.12	15	0.80	-----	
Nasality	Listeners	42.71	5	8.54	18.02	<0.01
	Scramblings	1.45	3	0.48	1.07	ns
	LXS	6.79	15	0.45	-----	
Sustention	Listeners	58.83	5	11.76	3.39	ns
	Scramblings	4.33	3	1.44	------	ns
	LXS	52.16	15	3.47	-----	
Sibilation	Listeners	75.70	5	15.14	14.08	<0.01
	Scramblings	4.12	3	1.37	1.27	ns
	LXS	16.02	15	1.07	-----	
Graveness	Listeners	398.71	5	79.74	26.51	≤0.01
	Scramblings	8.12	3	2.71	-----	
	LXS	45.12	15	3.01	-----	
Compactness	Listeners	131.83	5	26.36	7.21	≤0.01
	Scramblings	7.16	3	2.39	-----	ns
	LXS	54.83	15	3.65	-----	

*For 5 and 15 degrees of freedom $P \leq 0.01$ for $F \geq 4.56$;
For 3 and 15 degrees of freedom $P \leq 0.01$ for $F \geq 5.42$.

To the extent that listener differences with respect to the various attributes are highly intercorrelated one must consider the hypothesis that such differences are collectively attributable to factors of motivation or attentiveness rather than to more elementary factors of discriminative capacity. To the extent, however, that the uncorrelated component of variance in listener response to a given attribute is significant, one would be justified in favoring an explanation in terms of listener differences in discriminative capacity. A variance design with listeners and attributes as main effects provides results which bear on this issue. Here, a significant interaction between listeners and attributes would imply that the uncorrelated component of listener variance is significant in the case of one or more attributes and in turn support the discriminational rather than the motivational explanation of individual differences in DRT performance. Table 8 presents the results of the analysis called for by this line of reasoning.

The large F-ratio associated with the listener effect indicates that individual differences in discriminative performance with respect to the various attributes tend to be intercorrelated in some degree. Conceivably, therefore, motivational factors contribute to individual differences in DRT performance. A significant F in the case of attributes, may be taken simply to indicate that the physical correlates of the various attributes are transmitted with differing degrees of fidelity by the vocoder used in this experiment. Finally, however, the significant F-ratio for the interaction of listeners and attributes signifies an independent component of the listener effect in the case of one or more attributes. This result is consistent with the hypothesis that listeners vary significantly in terms of basic capacity to discriminate one or more of six attributes in question. Such results raise the possibility that the DRT may have some value in the field of clinical audiology. Although this possibility has yet to be systematically explored, the results for a single

TABLE 8. LISTENER DIFFERENCES IN DRT PERFORMANCE:
RESULTS OF ANALYSIS OF VARIANCES ALL ATTRIBUTES

Source	SS	df	MS	F*	P
Listeners	536.05	5	107.41	52.8	≤0.01
Attributes	421.55	5	84.31	9.7	≤0.01
Interaction	217.11	25	8.68	4.3	≤0.01
Residual	219.50	108	2.03	---	

*For 5 and 25 degrees of freedom $P \leq 0.01$ for $F \leq 3.86$;
for 25 and 108 degrees of freedom $P \leq 0.01$ for $F \leq 1.95$.

case may be at least suggestive of the potential of the DRT in this respect.

Table 9 shows the results obtained where an individual with hearing loss, as described above, served as the listener for recordings of the DRT by a trained speaker. While not predictable in detail from the audiometric data for this listener the present results are generally consistent with them. Much additional research will of course be required to evaluate the potential of the DRT for purposes of clinical audiology. The present results serve at least to indicate the existence of this potential.

TABLE 9. DIAGNOSTIC SCORES FOR A LISTENER WITH HIGH-FREQUENCY HEARING LOSS

Attribute	Positive State	Negative State	Mean
Voicing	94	94	94
Nasality	94	100	97
Sustention	69	69	69
Sibilation	60	75	67
Graveness	62	31	46
Compactness	50	78	64
Average			73

SUMMARY

The results described above serve to demonstrate the general sensitivity of the Diagnostic Rhyme Test to the characteristics of the speaker, the speech transmission channel and the listener. They also serve to indicate its possible diagnostic uses in each of these applications. However, further research will be required to assess fully its potential for evaluation of speakers and listeners.

ACKNOWLEDGEMENTS

Many people have contributed directly or indirectly to the development of the Diagnostic Rhyme Test in its present form. The contributions of five of them must be acknowledged.

Mr. Juozas Mickunas was responsible for data collection in most of the studies conducted thus far with the DRT. He was also responsible for the recording and preparation of the bulk of the speech materials used in this study.

Miss Marion Cohen conducted a number of important methodological studies during the early stages of the development of the DRT and has subsequently provided valuable advice at many points.

Mr. Jack Miller compiled the item pool from which the original form of the DRT was constructed. He has also provided valuable assistance in data collection and analysis during the course of recent research with the DRT.

Miss Virginia Miethe developed the computer scoring techniques employed with the DRT and has also assisted in data collection and analysis.

Mr. Caldwell Smith encouraged the original development of the DRT. His patience and advice on many occasions are gratefully acknowledged.

Two issues relating to the DRT are perhaps most subject to controversy: one concerns the fundamental principle upon which it is based; the other, the form of the consonant taxonomy used in reducing this principle to practice. The writer assumes full responsibility for the manner in which these issues are resolved.

FOOTNOTES

1. 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).

2. Egan, J. P., Articulation Testing Methods, Laryngoscope 58, 955-991, (1948).

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

4. House, A. S., Williams, C. E., Hecker, H. L., and Kryter, K. D., Articulation Testing Methods: Consonantal Differentiation with a Closed-Response Set, J. Acoust. Soc. Am. 37, 158-166, (1965).

5. Clarke, F. R., Technique for Evaluation of Speech Systems, Final Report, Contract No. DA28-043-AMC-00227(E), USAEL, (1965).

6. Stevens, K. N., Hecker, M. H. L., and Kryter, K. D, An Evaluation of Speech Compression Systems, Technical Documentary Report RADC-TDR-62-171, Rome Air Development Center, Griffiss Air Force Base, (March 1962).

7. Voiers, Cohen and Mickunas, op.cit.

8. Miller, G. A. and Nicely, P. A., An Analysis of Perceptual Confusions Among Some English Consonants, J. Acoust. Soc. Am. 27, 338-352, (1955).

9. Voiers, W. D., The Elementary Dimensions of Transmission Failure in Digital Vocoders, unpublished research at the Sperry Rand Research Center (1966). This study involved the factor analysis of Fairbanks and Diagnostic Rhyme Test scores for a sample of 20 experimental digital vocoder systems.

10. Jakobson, R., Fant, C. G. M., and Halle, M., Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates, Technical Report No. 13, Acoustics Laboratory, MIT, (1952).

11. Halle, M., On the Bases of Phonology, in Fodor, J. A., and Katz, J. J. (eds.), The Structure of Language, (Prentiss-Hall, Inc., Englewood Cliffs, New Jersey, 1964), pp. 324-333.

Footnotes (cont'd)

12. During the time that Form III of the DRT was being developed, I was unfortunately unaware of the highly significant work of Wickelgren, Distinctive Features and Errors in Short-Term Memory for English Consonants, J. Acoust. Soc. Am. 39, 388-398 (1966). I do not concur with Wickelgren's conclusions concerning the taxonomic implications of his results, although my analysis of these results has subsequently led me to consider certain additional modifications of the JFH taxonomy (e.g., the assignment of a neutral or indifferent status to the liquids and glides on the attribute, sustention), which may lead to further modifications of the DRT. However, no modification thus far considered would require changes in the present design of the DRT.

13. Halle, op.cit.

14. Pollack, I., Rubenstein, H., and Decker, L, Intelligibility of Known and Unknown Message Sets, J. Acoust. Soc. Am. 31, 273-279, (1959).

15. Voiers, Cohen and Mickunas, op.cit.

16. Voiers, W. D., the Perceptual Bases of Speaker Identity, J. Acoust. Soc. Am. 36, 1065-1073, (1964) and Voiers, W. D., Performance Evaluation of Speech Processing Devices II. The Role of Individual Differences, Scientific Report AFCRL-66-24, Contract No. AF19(628)-4987, AFCRL, (1965).

17. House, Williams, Hecker and Kryter, op.cit.

18. Miller and Nicely, op.cit.

19. Voiers, Cohen and Mickunas, op.cit.

20. Voiers, op.cit., (1964).

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

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

1 ORIGINATING ACTIVITY *(Corporate author)* Sperry Rand Research Center Sudbury, Massachusetts 01776	2a REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP

3 REPORT TITLE

PERFORMANCE EVALUATION OF SPEECH PROCESSING DEVICES III. DIAGNOSTIC EVALUATION OF SPEECH INTELLIGIBILITY

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Final Scientific Report: 1 March 1965 - 30 November 1966 Approved 23 Feb. 1967

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

Voiers, William D.

6. REPORT DATE February 1967	7a TOTAL NO OF PAGES 73	7b NO. OF REFS --

8a CONTRACT OR GRANT NO. AF19(628)-4987 b PROJECT NO and Task No. 4610-02 c DOD Element No. 62405304 d DOD Subelement No. 674610	9a ORIGINATOR'S REPORT NUMBER(S) SRRC-CR-67-6 Final Report 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* AFCRL 67-0101

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Hq. AFCRL, OAR (CRB) UNITED STATES AIR FORCE L.G. HANSCOM FIELD, BEDFORD, MASS.

13. ABSTRACT

The report summarizes the results of a program of research on communication system evaluation from the standpoint of speech intelligibility and speaker recognizability.

The history and present status of the Diagnostic Rhyme Test (DRT) Form III are described along with the results of research relating to the validity of the DRT in various applications.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED

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						
Vocoders						

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.