ESD-TR-66-677

(FINAL REPORT)

INTELLIGIBILITY TEST METHODS AND PROCEDURES FOR THE EVALUATION OF SPEECH COMMUNICATION SYSTEMS

Carl E. Williams
Michael H. L. Hecker
Kenneth N. Stevens
Barbara Woods

December 1966

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

(Prepared under Contract No. AF 19(628)-5659 by Bolt Beranek and Newman Incorporated, Cambridge, Massachusetts.)

LEGAL NOTICE

When U.S. Government drawings, specifications or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

OTHER NOTICES

Do not return this copy. Retain or destroy.

Distribution of this document is unlimited.

ESD-TR-66-677

(FINAL REPORT)

INTELLIGIBILITY TEST METHODS AND PROCEDURES FOR THE EVALUATION OF SPEECH COMMUNICATION SYSTEMS

Carl E. Williams
Michael H. L. Hecker
Kenneth N. Stevens
Barbara Woods

December 1966

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

(Prepared under Contract No. AF 19(628)-5659 by Bolt Beranek and Newman Incorporated, Cambridge, Massachusetts.)

FOREWORD

This document was prepared by Bolt, Beranek and Newman, Inc., of Cambridge, Massachusetts, and is the final report under Air Force Contract AF 19(628)-5659. This contract was initiated under Project 2808, "Psychoacoustic Standards for Voice Communications," monitored under the direction of Mr. Allen C. Busch, ESVHA, of the Decision Sciences Laboratory.

This technical report has been reviewed and is approved.

Allen C Busch
ALLEN C. BUSCH
Project Officer
Decision Sciences Laboratory

James S. Duva
JAMES S. DUVA
Technical Director
Decision Sciences Laboratory

INTELLIGIBILITY TEST METHODS AND PROCEDURES FOR THE EVALUATION OF SPEECH COMMUNICATION SYSTEMS

ABSTRACT

In further exploring the Modified Rhyme Test (MRT), a recently developed intelligibility test designed for the evaluation of speech communication systems under operational military conditions, research has been conducted in the following areas: (a) the relation between MRT scores and other intelligibility test scores for various types and levels of speech distortion; (b) the influence of the closed-response format and listening experience on MRT scores; and (c) speaker intelligibility and the selection of speakers for recording the test lists. The present report describes the work undertaken in each of these areas. The ultimate objective of the work is the development of valid procedures for the efficient evaluation of speech communication systems.

The major experimental results demonstrate that (1) the relation between scores obtained with different intelligibility test materials is not unique but depends considerably on the type of speech distortion employed, (2) neither the closed-response format nor prior listening experience appreciably affects MRT scores, and (3) less intelligible speakers tend to be those whose voiceless consonants are generated with lower intensity, particularly in word-final position.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SECTION I

INTRODUCTION

Under Contract AF19(628)-382, Study of Methods for Psychoacoustic Evaluation of Speech Communication Systems, a "Modified Rhyme Test", suitable for measuring the intelligibility of speech transmitted over communication systems was developed and evaluated.[5] Laboratory tests and tests conducted in the field with Air Force communication systems indicated that the new test should prove to be a practical and valuable tool for the measurement of speech intelligibility.

While meeting an existing need for a brief intelligibility test that can be used under operational military conditions, the Modified Rhyme Test (MRT) was found to need further exploration if it is to be utilized properly and to its greatest advantage. During the evaluation of the test several questions arose suggesting areas for further research. These questions pertained to (1) the relation between MRT scores and other intelligibility test scores for various types and levels of speech distortion, (2) the influence of the closed-response format and listening experience on MRT scores, and (3) speaker intelligibility and the selection of speakers for recording the test lists.

The purpose of the present contract was to undertake work in each of these areas. The ultimate objective of the research reported herein is the development of valid procedures for the efficient evaluation of speech communication systems.

SECTION II

RELATION BETWEEN INTELLIGIBILITY SCORES FOR VARIOUS TEST METHODS AND DIFFERENT TYPES OF SPEECH DISTORTION

Among the several intelligibility tests currently being used to evaluate speech-communication systems are the Harvard PB-Word Intelligibility Test,[1] the Fairbanks Rhyme Test,[2] the Modified Rhyme Test, and the Harvard Test Sentences.[1] Inherent in each test are certain advantages and limitations which must be considered when a test is selected for a particular application. Individuals engaged in the development of speech-communication systems would often like to compare the performance of a system evaluated with one test with the performance of another system that has been evaluated with a different test. In order to make comparisons that are meaningful, knowledge is required about the relation between various intelligibility test scores.

Although some data are available on the relation between test scores,[7,8] such data have usually been obtained for only one type of distortion, namely speech masked by noise. Furthermore, the test materials have usually been recorded by only one speaker.* Some indirect comparisons have been made among test scores reported in the literature,[7] but such comparisons must be interpreted with caution. Differences observed among test scores may be due not only to the particular tests employed, but also to the noise spectra, the method of measuring signal-to-noise ratio,

* An exception is a study by Hirsh, Reynolds, and Joseph[4] who obtained data showing the relation between the intelligibility of monosyllabic, disyllabic, and polysyllabic words, and nonsense syllables for various cut-off frequencies of high and low pass filtering and for different signal-to-noise ratios. These investigators found the relations to be different for the two types of distortion.

and the particular speakers and listeners involved. For these reasons, it was deemed desirable to establish the relation between scores of different intelligibility tests in a single experiment wherein such factors that are likely to influence the results could be more carefully controlled.

The present study explores the relation between scores obtained with the Harvard PB-Word Test, the Fairbanks Rhyme Test, the Modified Rhyme Test, and the Harvard Test Sentences for various levels of three types of speech distortion. The three types of speech distortion employed were: additive speech-shaped noise (i.e., random noise whose spectrum level was uniform up to 500 Hz and decreased at a rate of 9 dB per octave above that frequency), peak clipping preceded by low-pass filtering (1 KHz, attenuation rate: 24 dB/oct), and processing by a digital channel vocoder (Hughes CV1546). Block diagrams of the instrumentation used to achieve these three types of distortion may be seen in Fig. 1.

A preliminary study was conducted to determine how many conditions would be required for adequate comparisons among the four intelligibility tests. On the basis of informal listening experiments employing the Harvard PB-Word Test, it was considered appropriate to cover the range 45 to 95 percent correct with the 15 experimental conditions shown in Table 1.

For the additive noise, the conditions were seven signal-to-noise ratios ranging from +10 to -10 dB. The signal-to-noise ratio was measured with a VU meter, and the speech level was determined by averaging the speech peaks of the individual test words in a given test list. For the peak clipping preceded by low-pass filtering, the conditions were four levels of clipping ranging from no

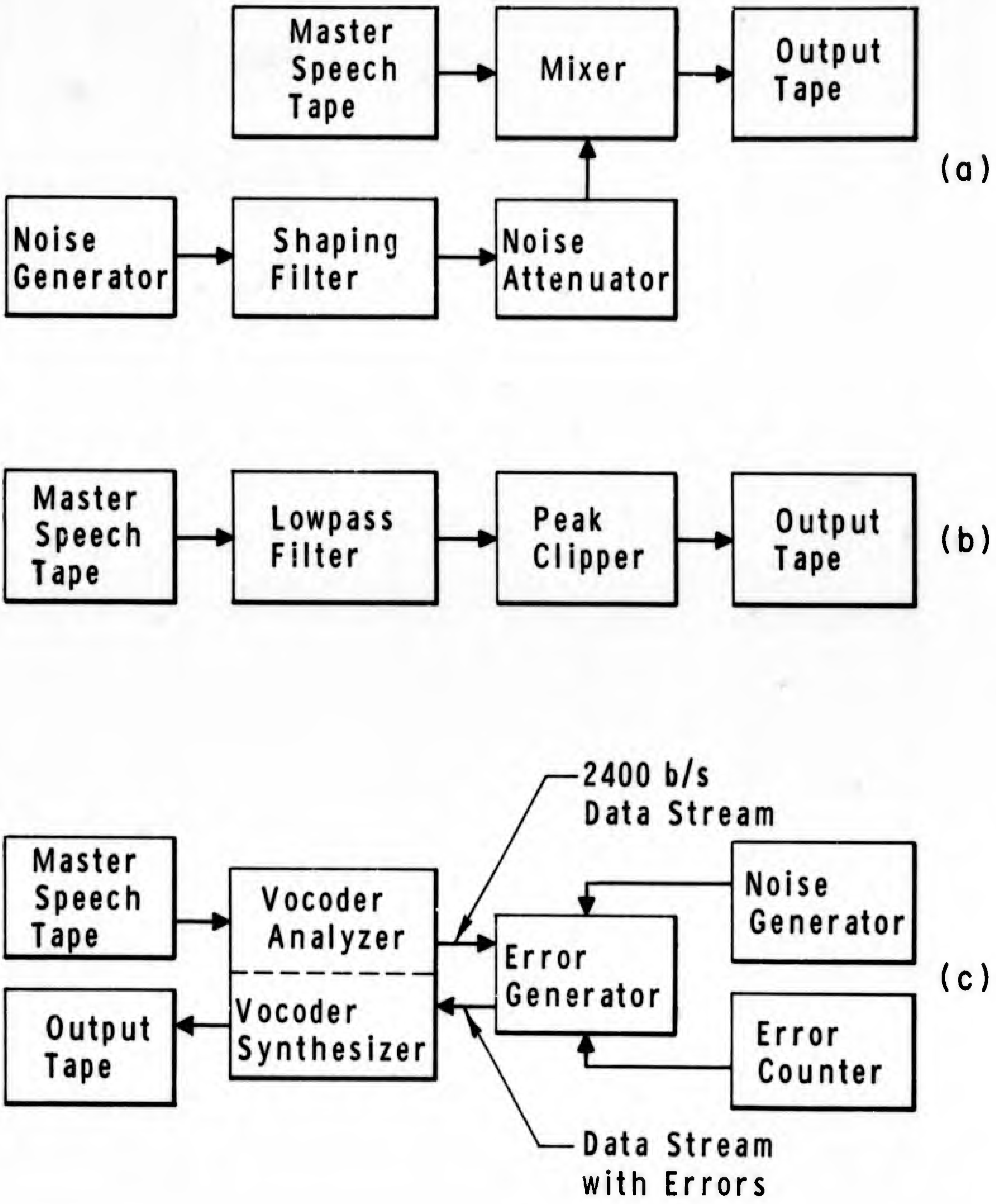


FIG. 1 BLOCK DIAGRAMS OF INSTRUMENTATION USED TO ACHIEVE THREE TYPES OF SPEECH DISTORTION: (a) ADDITIVE NOISE, (b) PEAK CLIPPING, AND (c) VOCODERIZATION.

BLANK PAGE

Table 1. Experimental conditions employed to obtain scores for various intelligibility test materials.

Condition No.	Type of Distortion	Level of Distortion	Test Material PB	RT	MRT	HTS
1	Additive Noise (speech-shaped)	S/N Ratio: +10 dB	x			
2		+ 5	x	x	x	
3		0	x	x	x	x
4		- 3	x	x	x	x
5		- 5	x	x	x	x
6		- 8		x	x	x
7		-10				x
8	Peak Clipping (preceded by low-pass filtering)	Clipping: 0 dB	x			
9		4	x	x	x	
10		16	x	x	x	x
11		22	x	x	x	x
12	Vocoderization (digital channel vocoder)	Error Rate: 0%	x	x	x	
13		2	x	x	x	
14		5	x	x	x	x
15		8	x	x	x	x

clipping to 22 dB of clipping. For the vocoderized speech, the conditions were four error rates ranging from 0 to 8 percent. The errors were introduced into the digital bit stream at random. Appendix I shows the mean error counts obtained when the test materials were processed with the different vocoder error rate conditions. As can be seen in Table 1, no one test material was represented by all 15 experimental conditions. Certain conditions for a given test material were excluded to keep the study within reasonable limits. Also, it was felt that for some conditions results could be predicted from earlier studies.

A given test material and experimental condition was represented by three test lists for each of two adult male speakers. To provide at least three test lists for a given material and condition, each speaker recorded: two randomizations of the twenty Harvard PB-Word lists, seven randomizations of the five Fairbanks Rhyme Test lists, six randomizations of the six Modified Rhyme Test lists, and 28 lists of selected Harvard Test Sentences. The recordings were made on a high-quality system in a sound-treated recording studio. The speakers attempted to maintain a constant vocal effort throughout each test list.

Master tape recordings of each of the four test materials were processed by the three types of distortion. Subsequent processed tapes were then edited and assembled in accordance with a matrix designed to provide appropriate randomizations of the conditions involved. Appendix II shows a sample test session design for each of the four test materials.

The tests were administered in the following order: the Harvard PB-Word Test, the Fairbanks Rhyme Test, the Modified Rhyme Test, and the Harvard Test Sentences. Prior to the administration of the Harvard PB-Word Test, the listeners were thoroughly familiarized with the 1000-word vocabulary. For each test material, a number of test lists representing some of the more severe test conditions were presented to the listeners for training purposes.

The listeners were ten high-school seniors, each of whom exhibited normal hearing. The test materials were administered to the listeners monaurally with a dummy phone covering the opposite ear. The speech was presented at an average level of 80 dB SPL. Listening sessions were held over a period of 20 two-hour sessions, spanning a period of approximately five weeks.

Figure 2 shows the mean-percent-correct listener responses obtained with the additive noise. (Mean listener scores obtained with each of the three distortions are also presented in Appendix III, together with standard deviations calculated from the individual listener scores.) Each point on the curves represents a mean intelligibility score based on responses by ten listeners to three test lists. The top portion of the figure shows results obtained for tests recorded by Speaker 1 and the bottom portion shows results obtained for tests recorded by Speaker 2. Not shown in Figure 2 are the mean-percent-correct listener responses obtained with the Harvard PB-Word Test at the +10 dB S/N condition. Scores of 98% and 97% were obtained for Speaker 1 and Speaker 2, respectively.

You will note that no sentence scores are shown for Speaker 2. The sentences recorded by Speaker 2 were characterized by an appreciable decrease in level after the first two or three words. Considerable difficulty was therefore encountered in setting the speech level during processing. While these sentences were administered to the listeners, the signal-to-noise ratios at which listener scores were obtained could not be stated with certainty and it was felt that to include the scores for Speaker 2 might be misleading. They are, however, shown in Appendix III.

In looking at the rank ordering of the tests according to intelligibility, it can be seen that the sentences provide the highest intelligibility curve and the PB words the lowest. With a few minor exceptions, the two rhyme-test curves are almost identical. The difference between rhyme test scores obtained at the -3 dB signal-to-noise condition for Speaker 2 was found to be statistically significant at the .05 level of confidence. The obtained value of $\underline{t}$, as determined by a $\underline{t}$ test for related measures, can be seen in Table 2. The table shows $\underline{t}$ values calculated for certain specific conditions, selected on the basis of visual inspection of the plotted data (Figures 2 through 5).

Upon comparing scores for the two speakers it is seen that Speaker 1 exhibits the higher scores. Whereas, for Speaker 1, the scores for the two rhyme tests and the PB words are quite similar at signal-to-noise ratios better than 0 dB, the PB word scores for Speaker 2 are considerably below those of the rhyme tests, even at a signal-to-noise ratio of +4 dB. (See statistical results in Table 2.) The appreciable differences between scores for the two speakers support the findings of earlier research and indicate the need for tests by more than one speaker in evaluating communication systems.

The mean-percent-correct responses for the peak-clipped speech are shown in Fig. 3. Here the rank ordering of the tests according to intelligibility is: the two rhyme tests, the sentences, and the PB words. As in the previous figure, the scores for the two rhyme tests are quite similar. Significant differences were found at the 22 dB clipping condition for Speaker 1 and the 16 dB clipping condition for Speaker 2. Except for the 16 dB clipping condition, the scores for the two speakers are almost identical. The difference between scores for the two speakers at this condition was significant for all test materials.

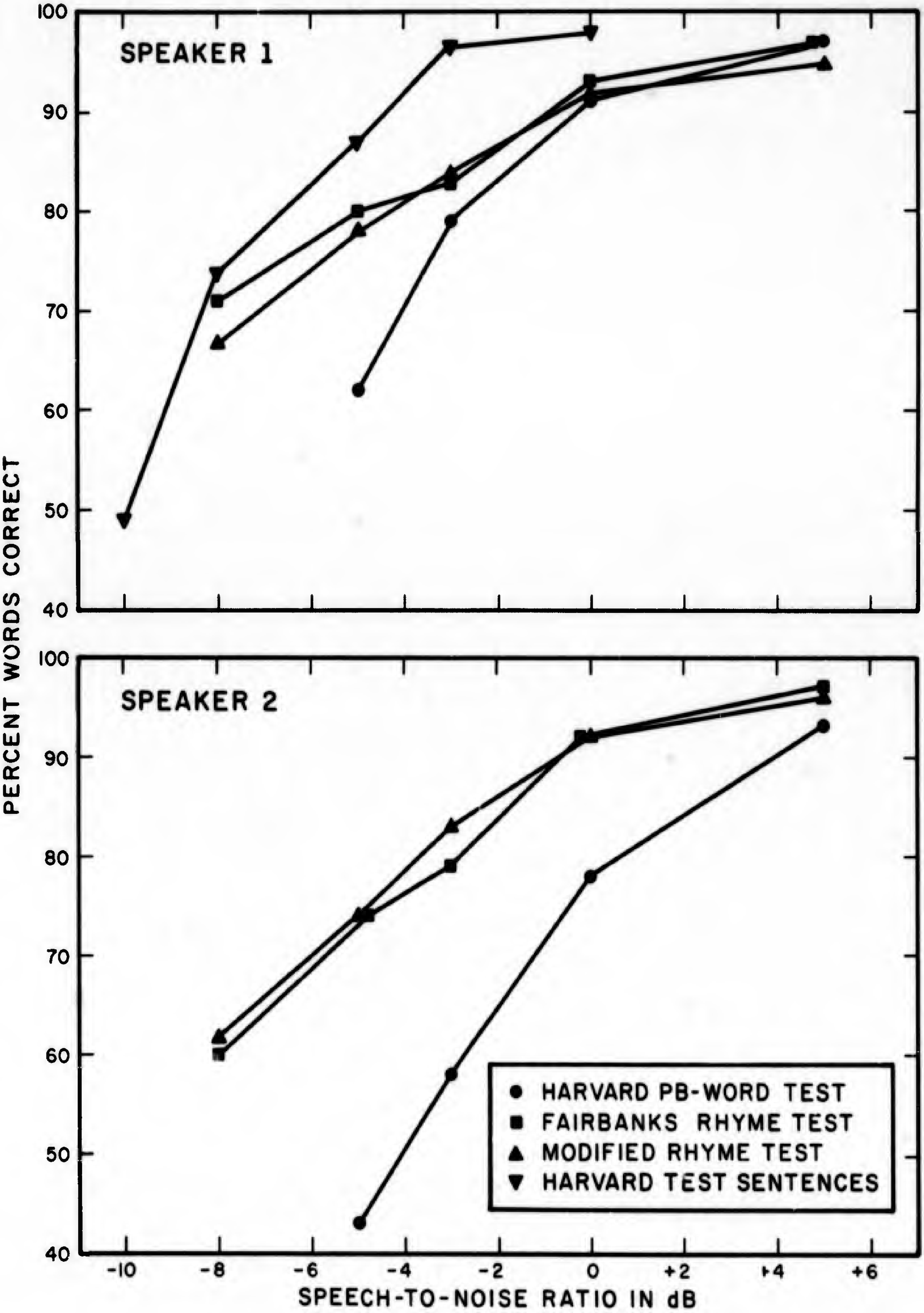


FIG. 2 PERCENT WORDS CORRECT AVERAGED OVER 10 LISTENERS FOR SPEECH WITH ADDITIVE NOISE.

BLANK PAGE

Table 2. Values of *t* for the evaluation of differences between test scores for certain conditions selected on the basis of visual inspection of the plotted data in Figs. 2 through 5.

Between Test Materials

Material	Type of Distortion	Level of Distortion	Speaker	t
RT and MRT	Additive Noise	-3 dB S/N	2	2.84
" " "	" "	-8 dB S/N	1	1.96*
" " "	Peak Clipping	16 dB	2	3.58
" " "	" "	22 dB	1	3.38
" " "	Vocoderization	0% error rate	2	3.03
" " "	"	2% error rate	1	2.43
" " "	"	2% error rate	2	3.38
" " "	"	5% error rate	2	2.93
PB words and RT	Additive Noise	0 dB S/N	1	1.82*
"	" "	0 dB S/N	2	8.51
PB words and MRT	" "	0 dB S/N	1	.83*
"	" "	0 dB S/N	2	9.63

Between Speakers

Material	Type of Distortion	Level of Distortion		t
PB words	Additive Noise	+5 dB S/N		4.76
" "	Peak Clipping	16 dB		5.89
" "	Vocoderization	5% error rate		1.71*
RT	Additive Noise	-3 dB S/N		3.65
"	Peak Clipping	16 dB		2.70
MRT	Additive Noise	-5 dB S/N		2.94
"	Peak Clipping	16 dB		2.77
"	Vocoderization	0% error rate		3.01

*Nonsignificant at .05 level.

It is evident that the scores for both rhyme tests remain largely independent of the level of clipping, whereas the scores for the PB words and for the sentences drop off as expected. A possible explanation for this may be related to the fact that no vowels are tested in the rhyme tests. Because peak clipping affects primarily the vowels, which are known to have a greater average intensity than the consonants, the intelligibility of the consonants may remain largely unaffected by the different amounts of peak clipping. In the case of the PB-word and sentence materials, on the other hand, both vowels and consonants are tested.

Figure 4 shows the mean-percent-correct responses for the vocoderized speech. The rank ordering of the tests according to intelligibility is the same as that obtained with additive noise: the Harvard Test Sentences, the two rhyme tests, and the PB words. Here again, the two rhyme tests produce similar scores, and the differences between speakers are minimal. Significant differences were found between the two rhyme test scores at the 2 percent error rate condition for Speaker 1 and at the 0, 2, and 5 percent error rate conditions for Speaker 2. There was a significant difference between MRT scores obtained for the two speakers at the 0 percent error rate condition. The points on the PB-word curves are in close agreement with results obtained by Steele and Cassel[10] in a related earlier study employing the Philco HY-2 vocoder.

Figure 5 shows scores for the Fairbanks Rhyme Test, the Modified Rhyme Test, and the Harvard Test Sentences, plotted as functions of PB-word scores. This arrangement of the data shows a little more clearly the relations between the different test materials for the three types of distortion. Except for the Harvard Test Sentences, the points on the curves represent means

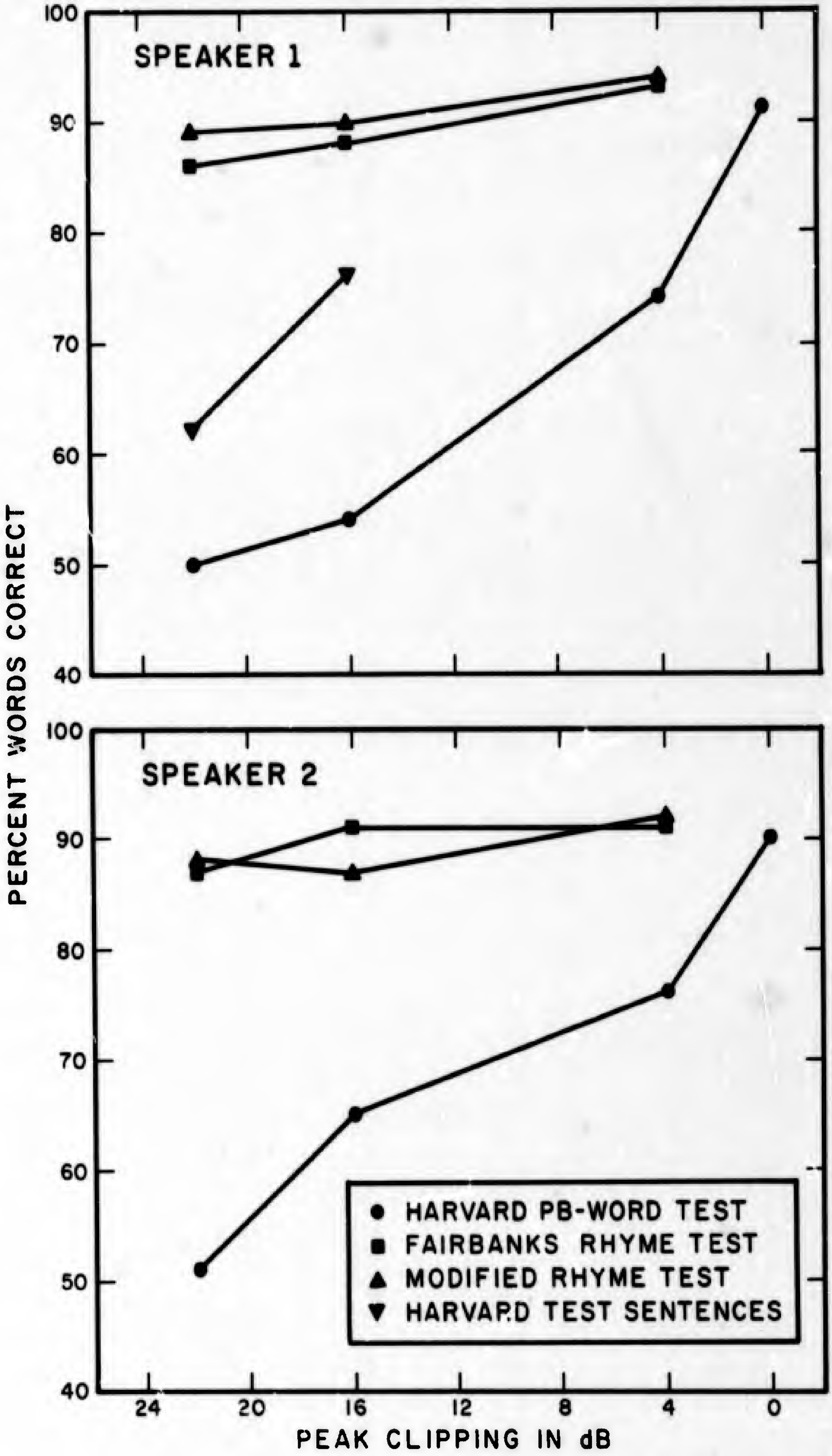


FIG. 3 PERCENT WORDS CORRECT AVERAGED OVER 10 LISTENERS FOR PEAK-CLIPPED SPEECH.

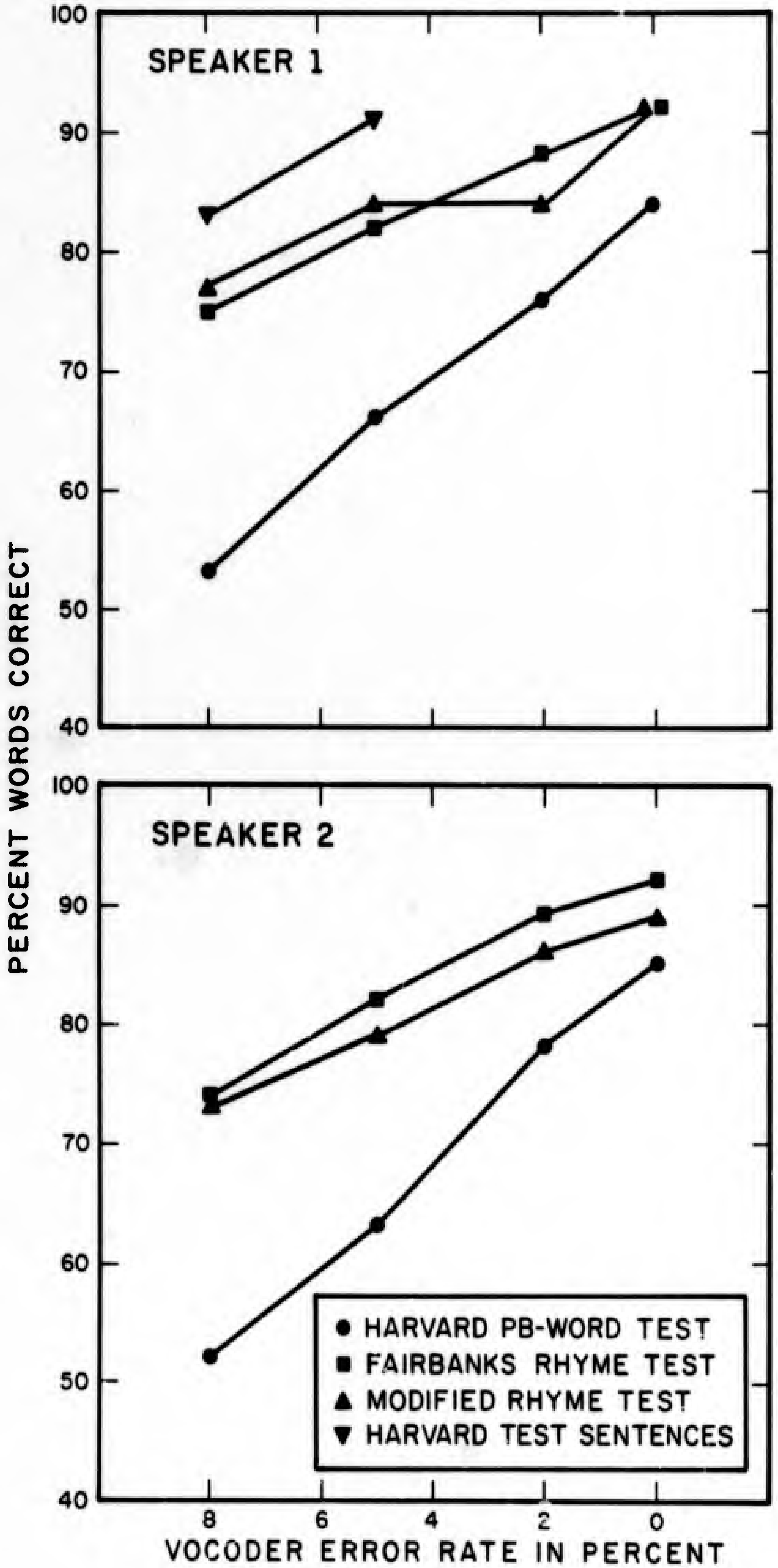


FIG. 4 PERCENT WORDS CORRECT AVERAGED OVER 10 LISTENERS FOR VOCODERIZED SPEECH.

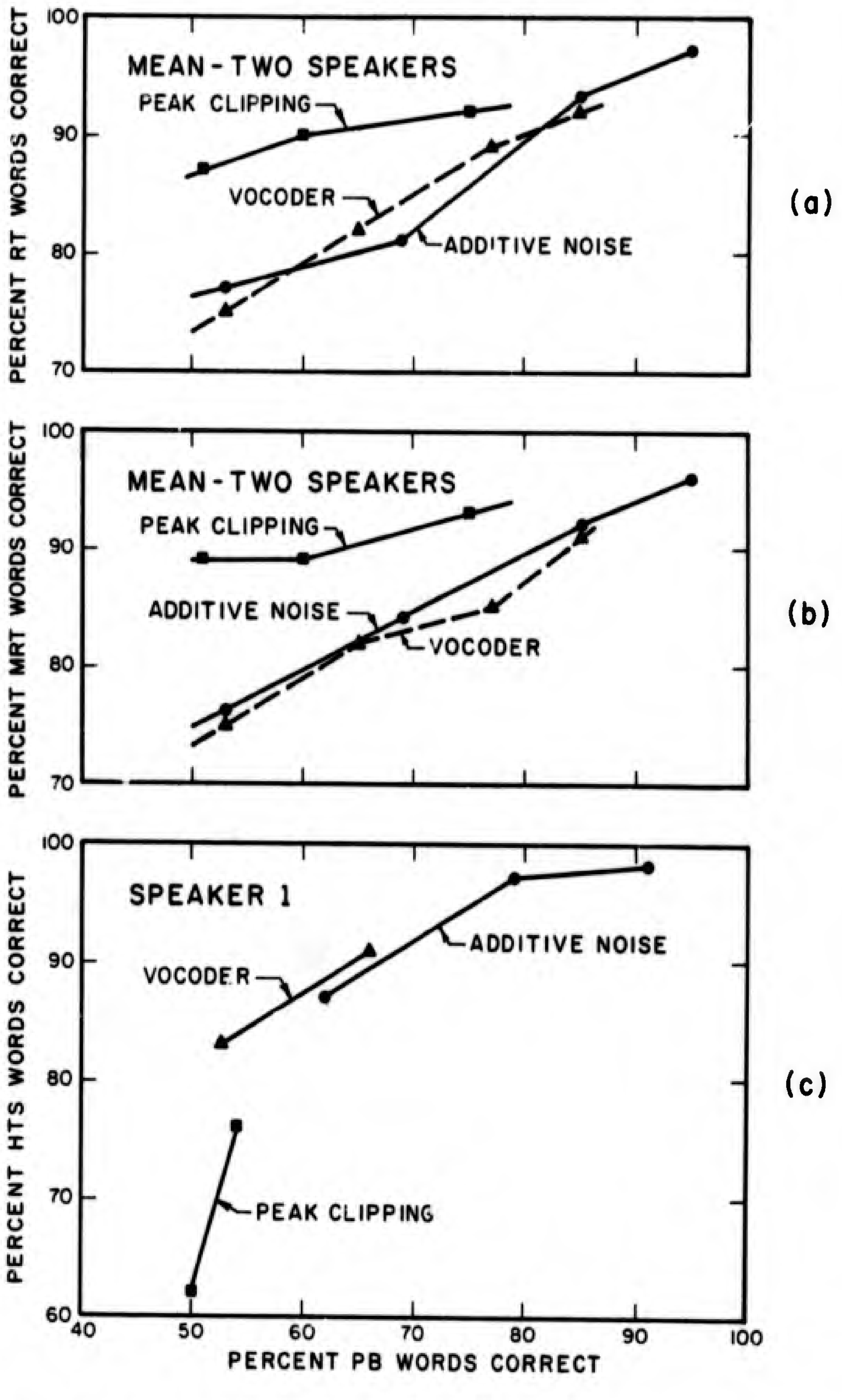


FIG. 5 RELATION BETWEEN SCORES OBTAINED WITH THE HARVARD PB-WORD TEST AND SCORES OBTAINED WITH (a) THE FAIRBANKS RHYME TEST, (b) THE MODIFIED RHYME TEST, AND (c) THE HARVARD TEST SENTENCES, FOR THREE TYPES OF SPEECH DISTORTION.

for both speakers. The curves for the additive noise and vocoderized speech are quite similar in all three plots. The plots for the two rhyme tests are very similar. Peak clipping preceded by low-pass filtering stands out as a unique type of distortion as far as these relations between test scores are concerned.

Two conclusions may be drawn from the results of this study:

(1) The relation between various test scores is not unique but depends considerably on the type of speech distortion involved. This finding implies that in converting scores obtained for a given speech-communication system with one test to scores that might be obtained with a different test, care should be taken to employ only data that is representative of the type of distortion involved. The results support the findings of Hirsh et al[4] reported earlier.

(2) Some types of distortion exaggerate speaker differences with respect to intelligibility, whereas others minimize such differences. The number of speakers that should be employed in an intelligibility-test program for evaluating speech-communication systems should, therefore, depend to some extent on the particular distortion involved.

SECTION III

EFFECT OF THE CLOSED-RESPONSE FORMAT ON MODIFIED RHYME TEST SCORES

The Modified Rhyme Test (MRT) with its closed-response format is more convenient to administer and score than the Harvard PB-Word Test, which has an open-response format. Untrained listeners may be employed, and the scoring of responses is readily adaptable to automation. The principal limitation of this test is that it appears to be less capable of discriminating among highly intelligible communication systems than the Harvard PB-Word Test. The present study has been undertaken to determine whether this shortcoming can be attributed to the closed-response format.

The materials of the MRT consist of 50 ensembles of six related words that differ only with respect to their initial or final consonants. These materials are recorded as six 50-word lists, and the listener is provided with a special answer sheet that shows the six response alternatives for each test item. A sample response form is shown in Appendix IV. The listener selects his answer for a given test item by marking one of the six words. If he should perceive a word which is not one of the response alternatives, he knows that this word is incorrect and he may select as his answer the next most-probable word from the closed-response set.

The Modified Rhyme Test could also be administered to listeners who are instructed to write down each perceived word on blank answer sheets. The response set could still be considered closed in the sense that the language limits the number of alternatives for each test item. However, since the number of possible words is frequently much greater than six, it is convenient to refer to an open-response set in this case.

An experiment was conducted in which the MRT vocabulary was administered to listeners using both the closed-response (i.e., multiple-choice) and open-response (i.e., write-down) formats.

One adult male speaker recorded four randomizations of each of the six test lists. Sufficient time was left between test items to enable listeners to write down the words. Two test conditions were produced by adding random noise to the speech signal and attenuating the speech to achieve signal-to-noise ratios of 0 dB and -8 dB. The noise spectrum level was uniform up to 500 Hz and decreased at a rate of about 9 dB per octave above that frequency. The recorded lists were organized according to a test matrix designed to take into account possible order effects for test lists and test conditions.

Thirteen high-school seniors and college students, none of whom had been previously exposed to the test vocabulary, participated in the experiment. The students were assigned at random to one of two listener groups. The experimental design employed for the two listener groups is presented in Table 3.

The six listeners constituting Group I came for six test sessions during each of which they heard the six test lists at each test condition. The write-down format was employed for the first five sessions, and the multiple-choice format for the last session. To familiarize the listeners with the test vocabulary as rapidly as possible, training sessions preceded Test Sessions 2 through 5. During training, the same speaker who recorded the test materials read the six lists live to the listeners, using a

Table 3. Experimental design employed to test the effect of the closed-response format on Modified Rhyme Test scores.

Test Session	Listener Group I	Listener Group II
1	12 tests (write-down format)	12 tests (multiple-choice format)
2	Training 12 tests (write-down format)	
3	Training 12 tests (write-down format)	
4	Training 12 tests (write-down format)	
5	Training 12 tests (write-down format)	
6	12 tests (multiple-choice format)	

microphone system that provided a signal-to-noise ratio of 20 dB. The randomizations of the lists used for training were different from those used in the actual tests. Immediately following the reading of a list, the correct words were read directly to the listeners (i.e. face-to-face) and they scored their responses.

The seven listeners constituting Group II came for a single test session during which they also heard the six test lists at each test condition. Only the multiple-choice format was employed. Appendix V shows the test design for one test session for each of the two listener groups.

The results of the experiment are shown in Fig. 6. Each point in Fig. 6 represents the mean score obtained for the six word lists. It can be seen that for both test conditions the scores of listeners in Group I, taking tests with the write-down format, leveled off after the fourth test session. This level, which was attained after a total of 14 exposures to the test vocabulary, was achieved by the listeners in Group II in a single exposure using the multiple-choice format. The standard deviations exhibited by the two groups of listeners are also comparable. A _t_ test for unrelated measures revealed that Group I's scores obtained during the fifth session with the write-down format did not differ significantly from Group II's scores obtained during the single session with the multiple-choice format. (See Table 4 for values of _t_.) Thus, it appears that the closed-response format does not, in and of itself, influence the discriminative capabilities of the MRT.

In this study, no attempt was made to examine other factors that might limit the discriminative capabilities of the MRT. If the test is to be considered for laboratory evaluation as well as

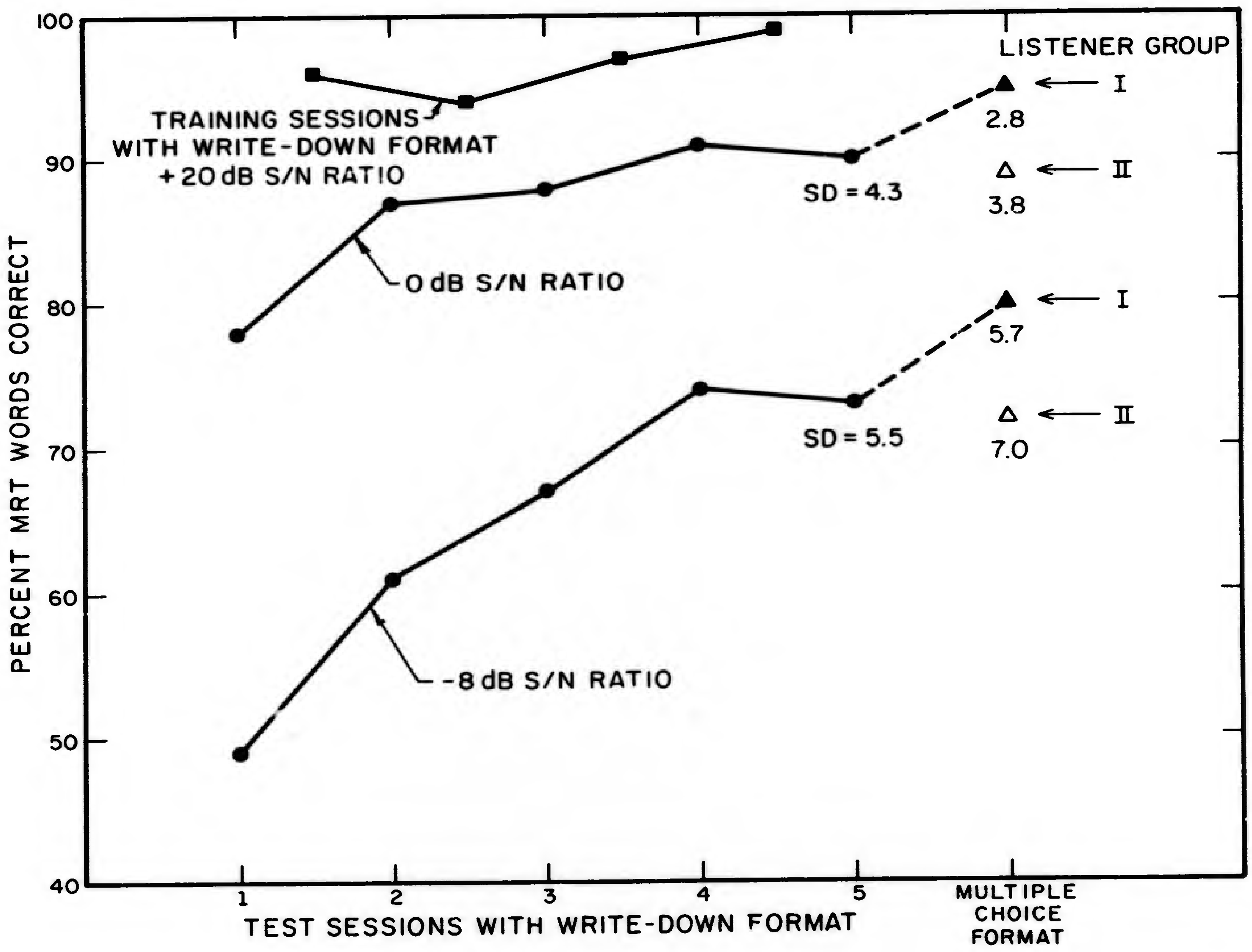


FIG. 6 PERCENT MRT WORDS CORRECT, AS A FUNCTION OF TEST SESSION, OBTAINED WITH TWO TEST FORMATS.

BLANK PAGE

Table 4. Results of t tests calculated for MRT scores obtained from four groups of listeners. t values with asterisk denote significance at the 0.05 level.

A. 0 dB S/N Ratio

	Listener Group and (Test Session)			
	I(5)	II	III(1)	IV(1)
I(5)	-	1.41	1.04	1.82
II		-	2.75*	0.49
III(1)			-	3.17*
IV(1)				-

B. -8 dB S/N Ratio

	Listener Group and (Test Session)			
	I(5)	II	III(1)	IV(1)
I(5)	-	0.42	0.83	2.50*
II		-	1.13	1.99
III(1)			-	2.63*
IV(1)				-

field evaluation of speech communication systems, additional studies must be undertaken to identify the limiting factors and improve the test in this regard. The present finding is encouraging in that the chief advantages of the test depend upon its closed-response format.

It can be observed that the scores achieved by Group I listeners using the multiple-choice format were appreciably higher than the scores achieved by Group II listeners. While the former scores were obtained with a procedure never employed in normal usage of the MRT, this observation raised three questions:

(1) Would another group of listeners, when tested under the same conditions as Group II, produce scores similar to those obtained from Group II, or would their scores more closely approximate those obtained from Group I with the multiple-choice format?

(2) To what extent would scores obtained from another group of listeners be temporally stable? Although it can be argued that the multiple-choice format excludes the possibility of vocabulary learning, repeated exposure to the test vocabulary, together with increasing familiarity with a speaker's voice in the presence of a particular type of speech distortion, may still influence listener performance.

(3) Would prior listening experience gained during exposure to the same speaker and test conditions, but with an entirely different vocabulary, result in higher MRT scores than are shown for Group II?

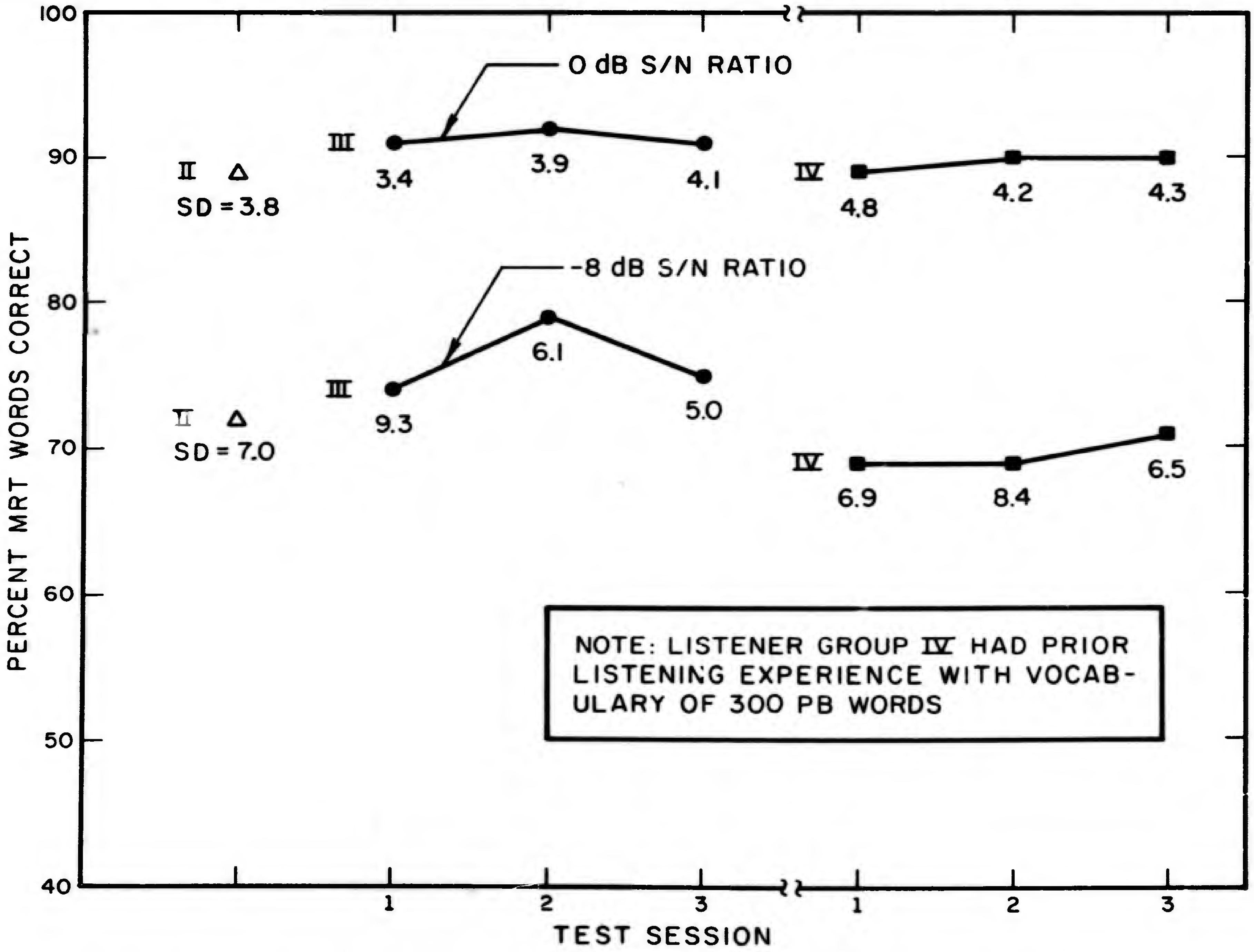
In an attempt to answer these questions, two additional groups of listeners, one of six (Group III) and one of nine (Group IV), were formed. Again high-school seniors and college students were employed as listeners, none of whom had previous experience in listening to speech tests.

Both groups of listeners came in for several test sessions. Table 5 shows the experimental design employed for these two groups of listeners. During some sessions the listeners were tested with the six lists of the MRT, using the regular multiple-choice format, and during other sessions they were tested with six lists of the Harvard PB-Word Test. Group III listeners came for six sessions. During the first three sessions, they heard the MRT. Following this, there were three test sessions in which they heard the restricted PB-word vocabulary of 300 words. Group IV listeners came for seven sessions. During the first four sessions, they heard different randomizations of the six PB-word lists. Following this, there were three sessions in which they heard the MRT. Test designs for a single test session for the two listener groups are shown together with those for Groups I and II in Appendix V.

Figure 7 shows, for both groups of listeners, the scores achieved on the MRT as a function of test session. Each point represents a mean of the six lists constituting the total 300-word vocabulary. The mean scores obtained for Group II listeners are shown again in this figure to allow comparison with the results obtained from Group III listeners. It is readily seen that Group III listeners produced scores similar to those of Group II listeners and, except for the unexpected rise in scores for the -8 dB condition during Test Session 2, their scores did not improve with repeated exposure to the test. While scores for the two groups are similar, it should be noted (see Table 4) that at the 0 dB S/N condition they were significantly different.

Table 5. Experimental design employed to test the effect of listening experience on Modified Rhyme Test scores.

Test Session	Listener Group III	Test Session	Listener Group IV
1	12 Modified Rhyme Tests	1	12 Harvard PB-Word Tests (300 word vocabulary)
2	"	2	"
3	"	3	"
1	12 Harvard PB-Word Tests (300 word vocabulary)	4	"
2	"	1	12 Modified Rhyme Tests
3	"	2	"
		3	"



NOTE: LISTENER GROUP IV HAD PRIOR LISTENING EXPERIENCE WITH VOCABULARY OF 300 PB WORDS

FIG.7 PERCENT MRT WORDS CORRECT AND STANDARD DEVIATIONS, AS FUNCTIONS OF TEST SESSION, FOR TWO GROUPS OF LISTENERS.

BLANK PAGE

Group IV listeners, who had prior listening experience with the restricted PB-word vocabulary, produced scores which were no higher than those obtained from Groups II and III. In fact, this group produced slightly lower scores at the -8 dB condition. Their scores were significantly different from the scores produced by Group III at both the 0 dB and -8 dB S/N condition.

These results demonstrate that scores, as obtained with the regular multiple-choice format, are similar for different groups of inexperienced listeners, and that repeated exposure to the test does not result in higher listener scores. The results also show that prior listening experience, obtained using a different test vocabulary, does not result in higher MRT scores.

Figure 8 shows, for both groups of listeners, the scores achieved on the PB-word lists as a function of test session. Group IV listeners, who had no prior listening experience before taking the tests, do not show significant improvement until after the third test session. Group III listeners, who had prior listening experience with the MRT, show considerable improvement after the first test session. In fact, Group III scores for the second session equal or surpass Group IV scores obtained during the fourth session. Whereas listening experience prior to the administration of a test with a closed-response format does not result in higher scores, listening experience prior to the administration of a test having an open-response format appears to accelerate learning of the test vocabulary.

The results shown in this and the previous figure do not explain the sudden increase in MRT scores for Group I listeners when they proceeded from the open-response format to the closed-response format. The fact that such high scores were never achieved using the regular multiple-choice format, even after repeated exposure to the test, leads us to believe that the sudden increase in scores may have been due to a learning of the different randomizations involved, or to experimental error.

In conclusion, the results of this study indicate that neither the closed-response format nor prior listening experience appreciably affects Modified Rhyme Test scores. The results also provide further evidence of the temporal stability of MRT scores for a given type and level of distortion.

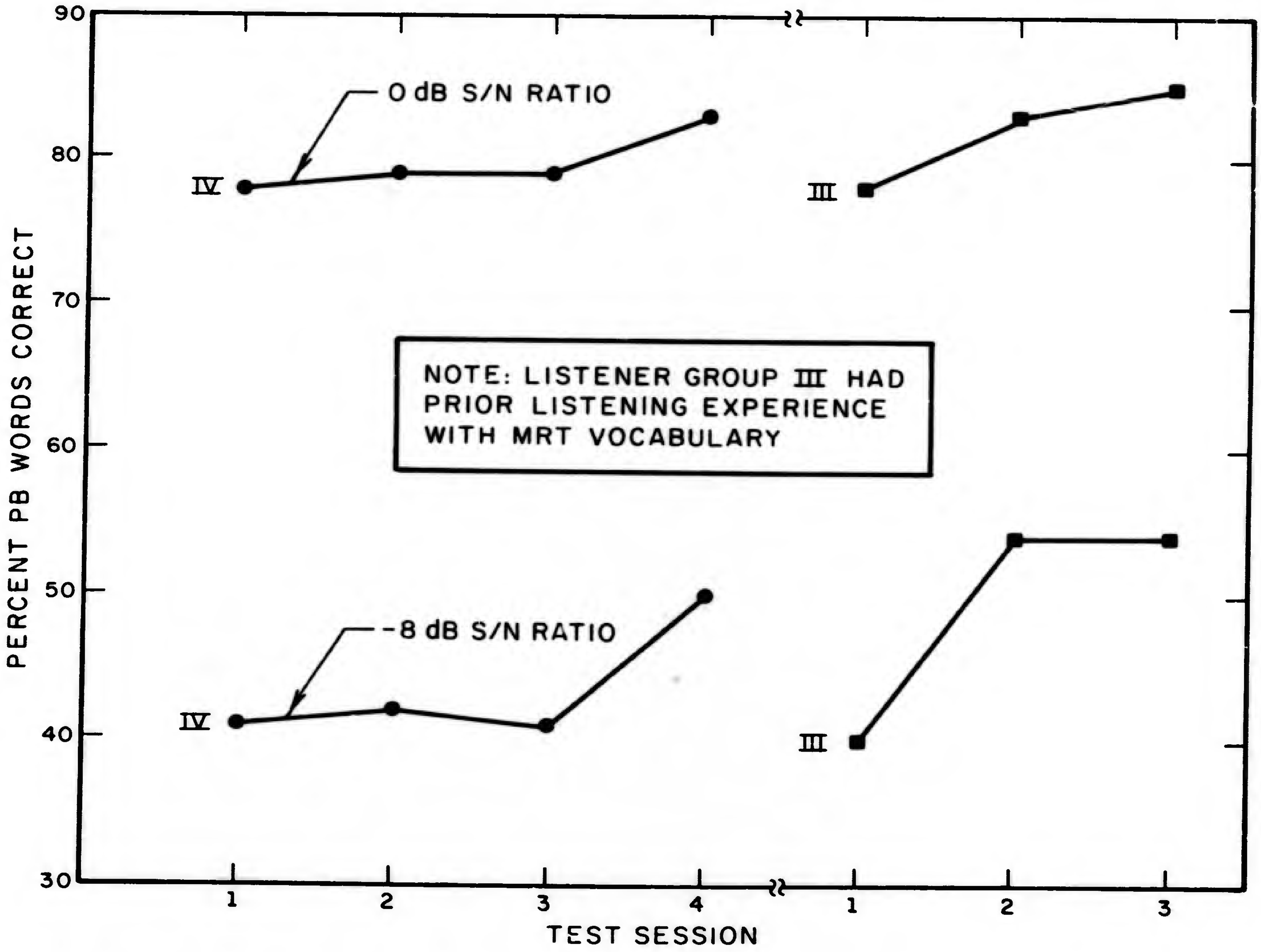


FIG. 8 PERCENT PB WORDS CORRECT, AS A FUNCTION OF TEST SESSION, FOR TWO GROUPS OF LISTENERS.

BLANK PAGE

SECTION IV

CONSONANT-VOWEL RATIO AND SPEAKER INTELLIGIBILITY

It has been recognized for some time that speakers as well as listeners are experimental variables in speech research. Speaker variability has been a particular problem in intelligibility testing, whether it be in clinical testing for speech reception or in the evaluation of communication systems. Even when speakers are selected on the basis of such category designations as "superior" and "experienced", and are considered to have no obvious speech idiosyncrasies, it is often found that tests recorded by such speakers yield significantly different intelligibility scores.

Some results of past research suggest that speaker intelligibility may be related to a physical measure of the speech signal, namely the consonant-vowel ratio. During the evaluation of the Modified Rhyme Test[5] it was found that the words of one speaker were not as well identified as those of the other speaker, and that his speech was characterized by a poorer consonant-vowel ratio. Fairbanks and Miron[3] found that, under various conditions of vocal effort, the consonant-vowel ratio within the syllable may change. They have suggested that the variations are systematic and large enough to have implications for intelligibility. Both Kryter[6] and Pickett[9] have emphasized the importance of vocal effort as a factor in the psychoacoustics of intelligibility.

Since voiced and voiceless speech sounds are generated by different mechanisms, it might be hypothesized that speaker intelligibility is influenced by differences in the production of these sounds and that these differences are reflected in the consonant-vowel ratio for voiceless consonants. The purpose of the present study was to explore further the relation between consonant-vowel ratio and speaker intelligibility.

To arrive at a group of speakers representing a suitable range of consonant-vowel ratios, recordings were made of 35 male college students reading the 50 monosyllabic words in List C of the Modified Rhyme Test. Graphic-level tracings were made of each recorded list and consonant-vowel ratios were obtained for eight words having the fricative /s/ in the initial position and for eight words having /s/ in the final position. The consonant /s/ was chosen for several reasons: Not only can it be easily differentiated from adjacent vowels in sound-pressure tracings, but it has a high frequency of occurrence in the language, and it is in the mid-range of consonant power. Two mean consonant-vowel ratios, corresponding to the initial and final positions, were thus determined for each of the 35 speakers. On the basis of several criteria which included similar rank order for initial and final /s/-vowel ratios and relatively small standard deviations, six speakers were selected to participate in the study. These speakers returned for a second recording session in which they recorded all six lists of the Modified Rhyme Test (MRT). Each speaker recorded different randomizations of the six lists.

To provide material for studying the influence of vocal effort on the consonant-vowel ratio, two additional speakers recorded lists at three levels of vocal effort. They recorded six lists employing normal vocal effort, three lists with decreased vocal effort, and three lists with increased vocal effort. The level of vocal effort was monitored with a sound-level meter located at the position of the recording microphone, 12 inches from the speaker's lips. Speech levels for the three vocal efforts, as measured on the "C" scale (fast deflection) of the sound-level meter, were: 69 dB, 78 dB, and 87 dB.

For each of the eight speakers using normal vocal effort, 54 words were used to obtain a mean /s/-vowel ratio for each consonant position. For the decreased and increased vocal efforts only 27 words were used. The words used to obtain these ratios may be seen in Appendix VI. Words ending in the clusters /ks/ and /st/ were included as words having /s/ in the final position. This was done to provide a sufficient number of words, and only after noting that the obtained level of /s/ did not differ significantly from the level of /s/ when it was not in one of these clusters. The mean /s/-vowel ratios and corresponding standard deviations calculated for each of the eight speakers are presented in Appendix VII.

All recordings obtained from the eight speakers were presented to listeners according to a test matrix that took into account possible order effects for speakers, test lists, and test conditions. Appendix VIII shows the test design for one test session. Four high-school seniors and four college freshmen served as listeners. The test conditions were produced by adding random noise to the speech signal and attenuating the speech to achieve signal-to-noise ratios of 0 dB and -8 dB. The noise spectrum level was uniform up to 500 Hz and decreased at a rate of 9 dB per octave above that frequency. Speech and noise were presented monaurally via Telephonics TDH-39 earphones in a sound-treated room. The mean intelligibility score for a given speaker, vocal effort, and signal-to-noise ratio was based on three different test lists. These scores, standard deviations calculated from listener scores for three test lists, and mean initial and final /s/-vowel ratios for each of the speakers are presented in Table 6.

Table 6. Mean percent intelligibility scores, standard deviations (shown in parenthesis) calculated from listener scores for three test lists, and mean initial and final /s/-vowel ratios, for each of the eight speakers.

Speaker	Intelligibility Score		Initial /s/-Vowel	Final /s/-Vowel
	0 dB S/N	-8 dB S/N		
1	79.2 (5.1)	53.3 (7.0)	15.6	22.3
2	90.7 (3.6)	68.8 (7.6)	6.8	14.3
3	85.8 (4.6)	64.3 (9.2)	11.8	20.1
4	89.0 (4.3)	74.0 (6.9)	7.5	11.7
5	81.7 (5.9)	67.3 (7.8)	6.2	13.2
6	82.4 (5.9)	59.0 (7.3)	14.3	19.2
7	89.3 (6.7)	69.5 (5.7)	13.2	12.6
8	90.5 (5.3)	75.2 (6.6)	14.2	17.4

Mean /s/-vowel ratios, both initial and final, and mean speaker intelligibility scores, were subjected to rank order correlations. A significant correlation (.02 level) was found between final /s/-vowel ratio and speaker intelligibility. In Figure 9 the eight speakers are divided into two groups, as shown by the two sets of brackets. The four most intelligible speakers, shown in the top set of brackets, exhibited the better consonant-vowel ratios, whereas the four least intelligible speakers, shown in the bottom set of brackets, exhibited the poorer consonant-vowel ratios. As would be expected, speaker differences were more evident at the -8 dB signal-to-noise ratio. There was a significant correlation between speaker intelligibility scores obtained at the 0 and -8 dB signal-to-noise ratios.

Based on the results of the experiment thus far, it was decided to examine the relation between consonant-vowel ratio and speaker intelligibility for other consonants besides /s/, some of which would represent a different class of consonants, but, like /s/, would easily permit the measurement of consonant-vowel ratios from graphic-level tracings. The consonants selected were the stops /t/ and /k/, the fricative /ʃ/, and the affricate /tʃ/. Instead of obtaining initial and final consonant-vowel ratios for all eight speakers, it was decided to merely look at two speakers from each of the two groups of speakers shown in Fig. 9. The speakers which were selected are represented by the dark circles in the figure. The number of words used to obtain these ratios was as follows: 54 words for the initial and final /t/-vowel ratios, 27 words for the initial /k/-vowel ratio, 36 words for the final /k/-vowel ratio, and three words each for the initial /ʃ/-vowel and final /tʃ/-vowel ratios. (See Appendix IX for the words used to obtain the C/V ratios.)

The consonant-vowel ratios for the additional consonants examined are presented in Appendix X. They are also shown in Figures 10 and 11, which illustrate graphically the relation between consonant-vowel ratio and speaker intelligibility. Included also, for purposes of comparison, are the initial and final /s/-vowel ratios and mean intelligibility scores for the four speakers. The abscissae of these two figures show, from left to right, the consonant-vowel ratios from best to worst. In viewing the figures it is seen that, except for the initial /t/-vowel and initial /k/-vowel ratios, the two more intelligible speakers always have the better (smaller negative number) consonant-vowel ratios. While the four speakers should be looked upon as two pairs of speakers, one pair representing a group of highly intelligible speakers and the other pair representing a group of less intelligible speakers, it should be noted that even for the four individual speakers there is a monotonic relation between final /t/-vowel ratio and intelligibility.

The mean intelligibility scores for each of the two speakers employing the three levels of vocal effort are shown in Fig. 12. Also shown are the initial and final /s/-vowel ratios for each speaker's three levels of vocal effort. It can be seen that for both speakers at the -8 dB signal-to-noise ratio, there are monotonic relations between level of vocal effort and /s/-vowel ratio and between /s/-vowel ratio and intelligibility. This is true for both the initial and final /s/-vowel ratios. At the 0 dB signal-to-noise ratio, intelligibility appears to be largely unaffected by level of vocal effort.

The results of this study demonstrate the importance of the consonant-vowel ratio as a factor in speaker intelligibility and

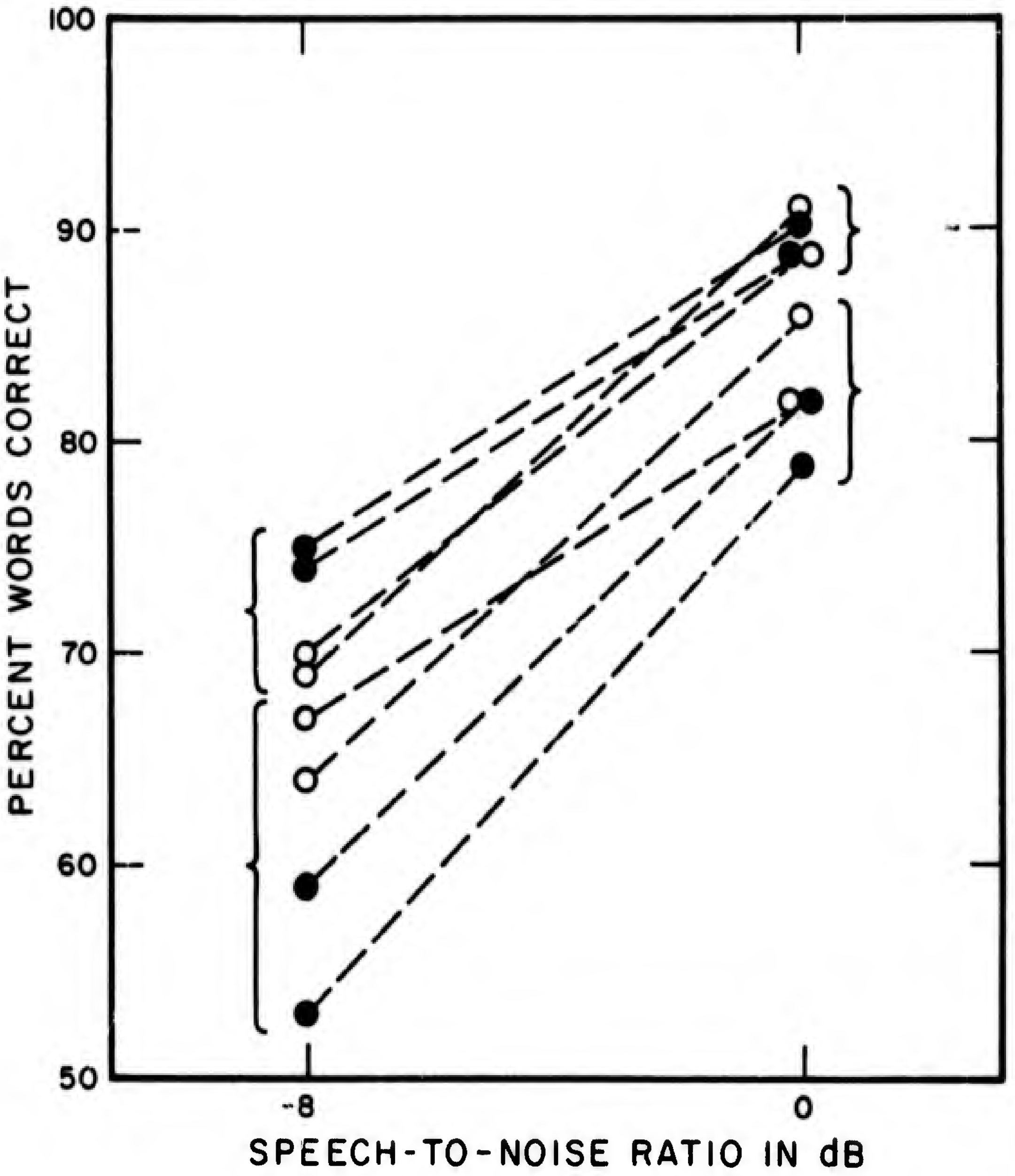


FIG. 9 MEAN INTELLIGIBILITY SCORES FOR EACH OF 8 SELECTED SPEAKERS AT TWO SIGNAL-TO-NOISE RATIOS. THE 4 SPEAKERS IN THE UPPER BRACKETS EXHIBITED THE BETTER /S/-VOWEL RATIO.

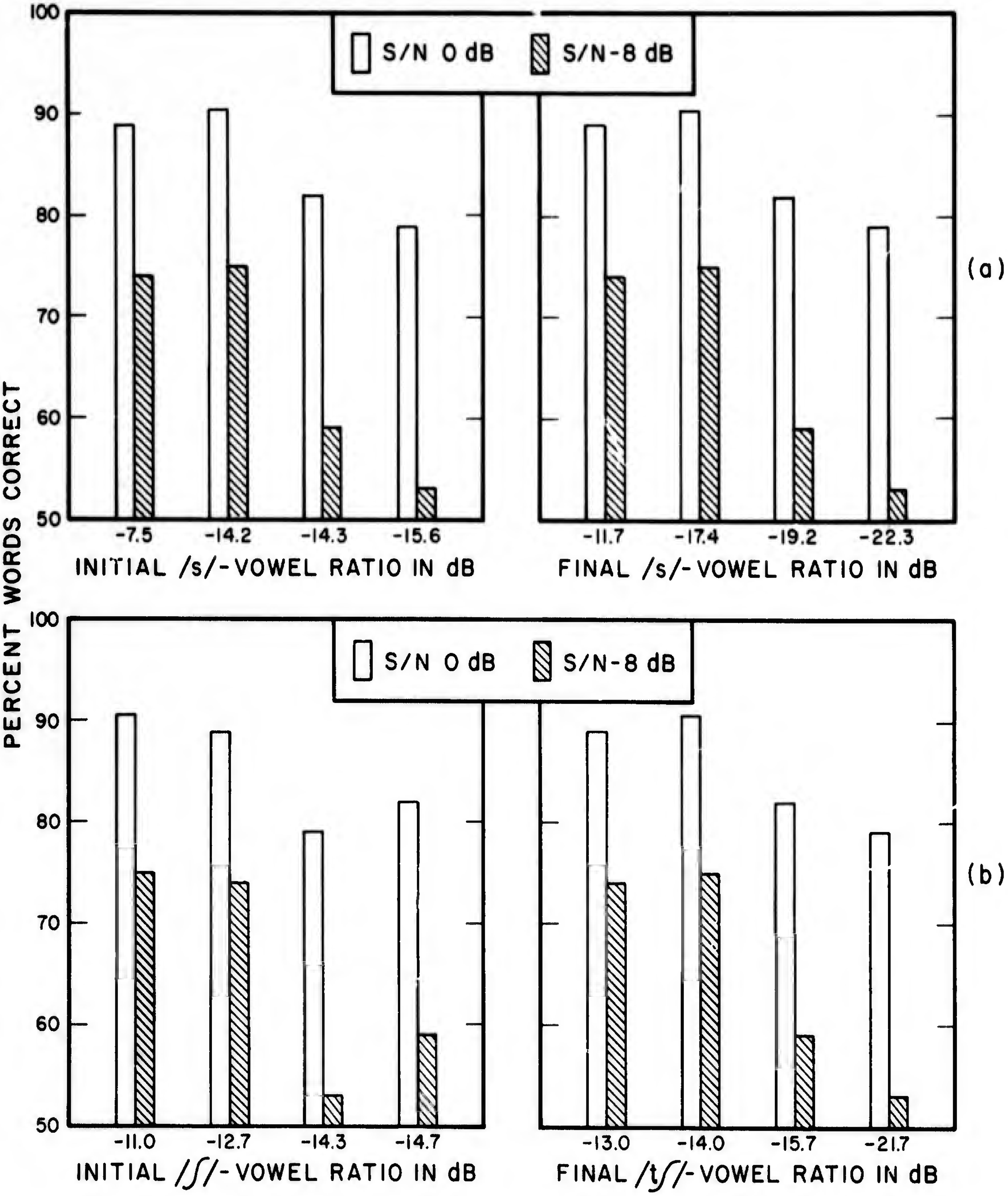


FIG. 10 MEAN INTELLIGIBILITY SCORES FOR 4 SELECTED SPEAKERS, EACH OF WHOM IS REPRESENTED BY HIS (a) /s/-VOWEL RATIO, AND (b) /ʃ/-VOWEL AND /tʃ/-VOWEL RATIO.

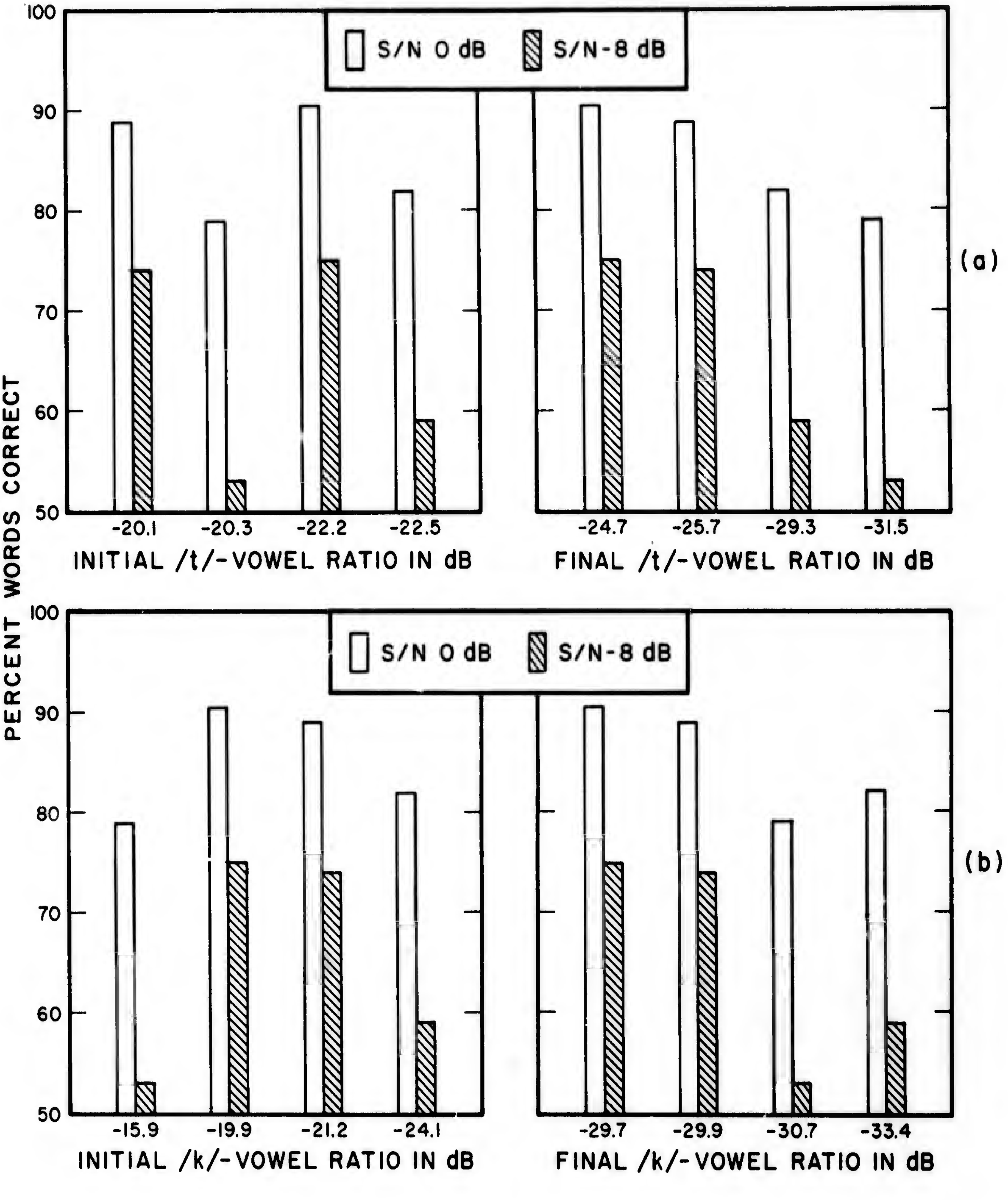
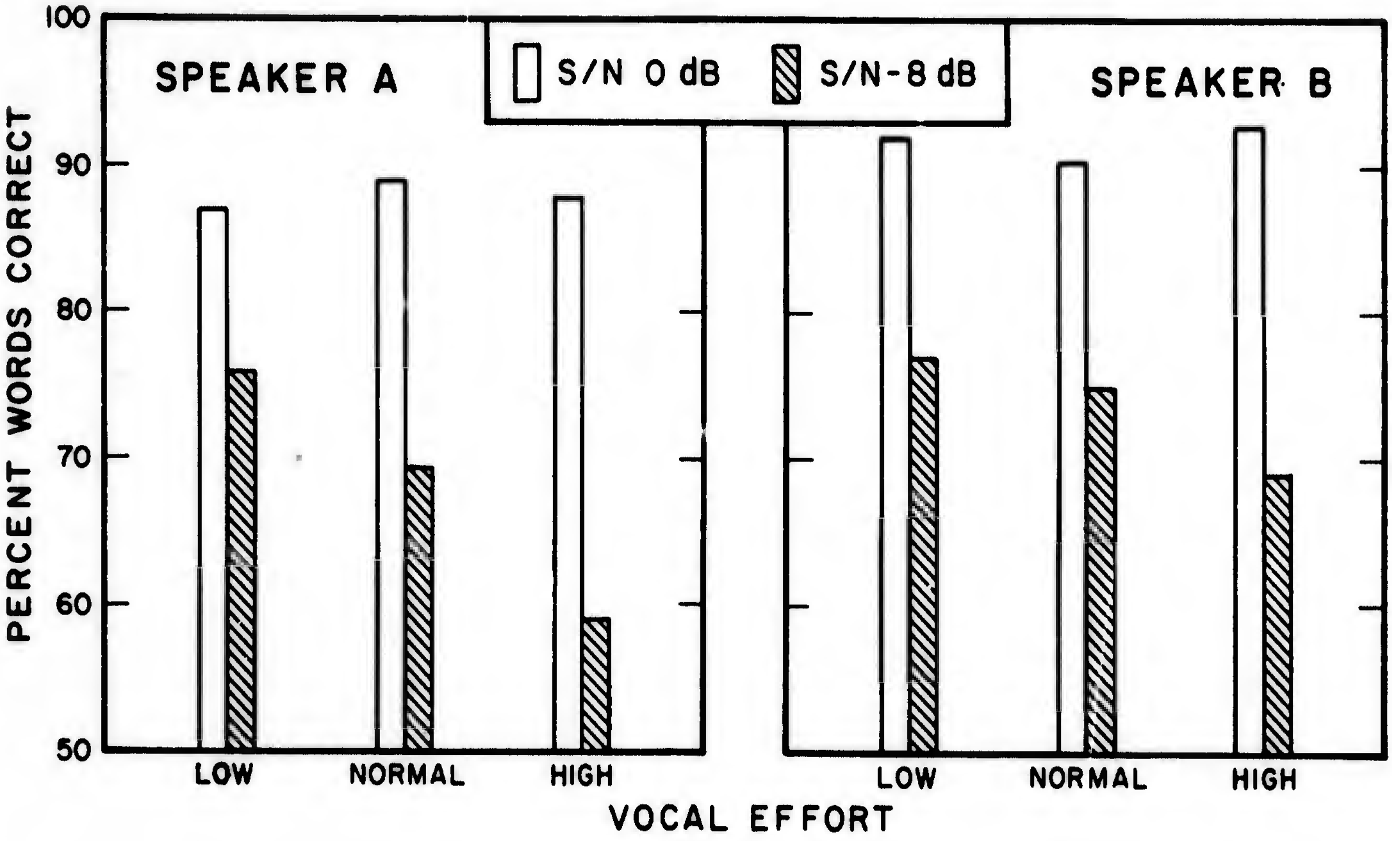


FIG. 11 MEAN INTELLIGIBILITY SCORES FOR 4 SELECTED SPEAKERS, EACH OF WHOM IS REPRESENTED BY HIS (a) /t/-VOWEL RATIO, AND (b) /k/-VOWEL RATIO.



INITIAL	-11.5	-13.2	-18.0	-10.0	-14.2	-15.3
FINAL	-9.3	-12.6	-19.0	-12.7	-17.4	-19.4

/s/-VOWEL RATIO IN dB

FIG. 12 MEAN INTELLIGIBILITY SCORES AND /s/-VOWEL RATIOS FOR EACH OF 2 SELECTED SPEAKERS, EMPLOYING 3 LEVELS OF VOCAL EFFORT.

confirm the effect of vocal effort on the consonant-vowel ratio. These findings have several practical implications. In the selection of speakers for recording speech materials for intelligibility tests, it may well be that consideration should be given to the intensity with which they generate final voiceless consonants. In training individuals to speak in noisy environments or over communication systems where the speech signal is likely to be degraded, individuals should be instructed in achieving a good consonant-vowel ratio. In such situations, individuals often have a tendency to raise their voice, thereby strongly emphasizing the vowels and neglecting the consonants. In recording materials for intelligibility tests, extreme care should be taken so that speakers employ a constant vocal effort.

REFERENCES

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

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

3. Fairbanks, G. and Miron, M. S., "Effects of Vocal Effort Upon the Consonant-Vowel Ratio Within the Syllable." J. Acoust. Soc. Am. 29, 621-626 (1957).

4. Hirsh, I. J., Reynolds, E. G., and Joseph, M., "Intelligibility of Different Speech Materials." J. Acoust. Soc. Am. 26, 530-538 (1954).

5. House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D., "Articulation Testing Methods: Consonantal Differentiation with a Closed-Response Set." J. Acoust. Soc. Am. 37, 158-166 (1965). Also, ESD-TDR-64-507.

6. Kryter, K. D., "On Predicting the Intelligibility of Speech from Acoustical Measures." J. Speech and Hearing Dis. 21, 208-217 (1956).

7. Kryter, K. D., and Whitman, E. C., "Some Comparisons Between Rhyme and PB-Word Intelligibility Tests." J. Acoust. Soc. Am. 37, 1146 (1965).

8. Nickerson, J. F., Miller, A. W., and Shyne, N. A., "A Comparison of Five Articulation Tests." Final Report No. RADC-TR-60-71, prepared under Contract AF30(602)-1818 for Rome Air Development Command, Air Research Development Command, U. S. Air Force, Griffiss AFB, N. Y. (March 1960).

9. Pickett, J. M., "Effect of Vocal Force on the Intelligibility of Speech Sounds." J. Acoust. Soc. Am. 28, 902-905 (1956).

10. Steele, R. W., and Cassel, L. E., "Effect of Transmission Errors on the Intelligibility of Vocoded Speech." IEEE Trans. on Communications Systems, CS-11, 118-123 (1963).

APPENDIX I

MEAN 10-SECOND COUNTER READINGS OBTAINED WHEN PROCESSING TEST MATERIALS WITH THE VOCODER. EACH ENTRY IN THE FOLLOWING TABLE IS A MEAN OF TEN COUNTER READINGS OBTAINED DURING THE PROCESSING OF ONE TEST LIST. THE INTENDED MEAN 10-SECOND COUNTER READINGS ARE SHOWN ON THE LEFT.

Condition	PB		RT		MRT		HTS	
	Speaker		Speaker		Speaker		Speaker	
	1	2	1	2	1	2	1	2
Nominal 2% Error Rate - 480	462	473	439	457	412	443	-	-
	436	451	433	464	433	447	-	-
	447	444	428	444	473	456	-	-
Nominal 5% Error Rate - 1200	1210	1210	1230	1210	1200	1280	1190	1240
	1190	1190	1180	1180	1210	1280	1200	1260
	1210	1200	1200	1250	1200	1240	1200	1260
Nominal 8% Error Rate - 1920	1941	1930	1938	1931	1930	1892	1928	1924
	1944	1940	1964	1923	1942	1910	1955	1943
	1964	1905	1944	1936	1912	1904	1954	1940

APPENDIX II

PORTION OF TEST DESIGN SHOWING ONE TEST SESSION* FOR EACH OF THE FOUR TEST MATERIALS

Test No.	HARVARD PB-WORDS			FAIRBANKS RHYME TEST			MODIFIED RHYME TEST			HARVARD TEST SENTENCES		
	Cond	List	Spkr	Cond	List	Spkr	Cond	List	Spkr	Cond	List	Spkr
1	1	1T	1	6	4-1	2	5	A5	1	5	49	2
2	11	8W	2	9	2	1	12	D2	1	7	62	2
3	3	10R	1	12	1-6	1	13	B3	2	15	15	1
4	14	9V	2	11	5-2	2	9	C5	1	11	64	2
5	12	3W	2	3	4	2	10	E6	1	14	35	1
6	2	5T	1	15	2-6	2	6	A1	1	4	38	1
7	9	11V	2	11	1-1	1	14	F3	2	5	31	1
8	14	6W	2	4	3-2	1	4	B4	1	15	44	2
9	4	18T	1	13	1-5	2	13	F5	2	11	23	1
10	1	13V	2	3	5-4	1	15	D2	2	10	52	2
11	13	11R	1	5	4-3	1	11	E5	1	7	18	1
12	10	16T	1	2	3-1	2	2	C3	1	4	61	2
13	3	13W	2	10	1-3	2	10	B4	2	10	19	1
14	15	7R	1	14	3-6	1	4	C5	2	3	58	2
15	5	19T	1	9	4-5	2	6	E3	2			
16	11	9W	2	2	2-4	1	15	A6	1			
17				15	3-3	1	3	F3	1			
18				12	4-2	2	11	A2	2			
19				4	5-6	2	9	E4	2			
20				6	2-1	1	2	D6	2			

*The overall test design for a given speech material was balanced over all test sessions for that material. The balance is not necessarily reflected in the test design for single test sessions.

APPENDIX III

MEAN LISTENER SCORES AND STANDARD DEVIATIONS (IN PARENTHESIS) OBTAINED WITH THE THREE DIFFERENT TYPES OF SPEECH DISTORTION

A. Speech With Additive Noise

Condition	PB		RT		MRT		HTS	
	Speaker		Speaker		Speaker		Speaker	
	1	2	1	2	1	2	1	2
+10 dB	97.9 (1.7)	96.5 (2.7)						
+5 dB	97.1 (2.0)	92.8 (4.4)	96.6 (2.3)	97.2 (2.3)	95.4 (2.4)	95.6 (2.6)		
0 dB	90.9 (5.8)	77.7 (8.7)	92.9 (4.5)	92.3 (3.1)	91.9 (3.8)	92.3 (3.6)	97.9 (2.1)	94.8 (2.7)
-3 dB	79.2 (7.1)	57.7 (9.5)	83.4 (5.3)	78.7 (6.3)	84.1 (5.7)	83.5 (5.0)	97.3 (2.7)	78.2 (13.6)
-5 dB	61.6 (10.6)	42.5 (8.8)	79.7 (6.2)	73.7 (7.7)	78.0 (5.1)	73.9 (7.2)	87.1 (8.2)	60.5 (11.7)
-8 dB			70.8 (9.8)	59.5 (9.5)	67.2 (6.8)	61.9 (9.4)	74.4 (14.6)	22.8 (14.2)
-10 dB							49.1 (21.6)	15.7 (10.0)

APPENDIX III (Continued)

B. Peak-Clipped Speech

Condition	PB		RT		MRT		HTS	
	Speaker		Speaker		Speaker		Speaker	
	1	2	1	2	1	2	1	2
0 dB	91.4 (4.9)	90.4 (5.3)						
4 dB	73.8 (6.4)	75.6 (8.7)	92.8 (3.3)	91.4 (4.1)	93.6 (2.9)	92.1 (3.8)		
16 dB	54.1 (8.8)	64.9 (6.3)	88.0 (4.4)	90.7 (4.2)	90.0 (4.2)	87.4 (4.0)	76.3 (8.2)	69.9 (8.5)
22 dB	50.2 (6.2)	50.5 (7.1)	85.5 (4.5)	86.8 (5.1)	88.9 (3.7)	88.5 (4.3)	62.4 (11.3)	66.3 (7.4)

APPENDIX III (Continued)

C. Vocoderized Speech

Condition	PB		RT		MRT		HTS	
	Speaker		Speaker		Speaker		Speaker	
	1	2	1	2	1	2	1	2
0% Error Rate	83.8 (5.8)	84.7 (5.2)	91.7 (3.6)	92.0 (4.2)	92.2 (4.1)	89.3 (3.0)		
2% "	76.2 (5.9)	77.8 (7.3)	87.7 (4.0)	89.1 (4.5)	83.7 (5.7)	85.9 (4.4)		
5% "	65.7 (4.9)	62.6 (9.7)	82.1 (5.6)	82.5 (5.9)	83.7 (4.1)	78.9 (4.8)	91.1 (4.8)	75.2 (12.5)
8% "	53.3 (11.4)	52.1 (7.8)	75.3 (7.7)	74.0 (5.1)	77.0 (5.4)	73.5 (5.5)	82.7 (6.0)	68.9 (14.5)

BLANK PAGE

APPENDIX IV

SAMPLE MRT RESPONSE FORM

NAME________________________ DATE________________ TEST NO.____________

FORM 4X SCORE____________

1	sun	nun	gun
	run	bun	fun
2	kit	kick	kin
	kid	kill	king
3	bust	just	rust
	dust	gust	must
4	pill	pick	pip
	pit	pin	pig
5	ban	back	bat
	bad	bass	bath
6	rent	went	tent
	bent	dent	sent
7	pad	pass	path
	pack	pan	pat
8	bill	fill	till
	will	hill	kill
9	gang	hang	fang
	bang	rang	sang
10	sun	sud	sup
	sub	sung	sum
11	pave	pale	pay
	page	pane	pace
12	safe	save	sake
	sale	sane	same
13	tang	tab	tack
	tam	tap	tan
14	gale	male	tale
	pale	sale	bale
15	test	nest	best
	west	rest	vest
16	pub	pus	puck
	pun	puff	pup
17	pop	shop	hop
	cop	top	mop
18	name	fame	tame
	came	game	same
19	sin	sill	sit
	sip	sing	sick
20	sip	rip	tip
	lip	hip	dip
21	may	gay	pay
	day	say	way
22	sin	win	fin
	din	tin	pin
23	soil	toil	oil
	foil	coil	boil
24	cuff	cuss	cub
	cup	cut	cud
25	wig	rig	fig
	pig	big	dig
26	sap	sag	sad
	sass	sack	sat
27	tick	wick	pick
	kick	lick	sick
28	lot	not	hot
	got	pot	tot
29	park	mark	hark
	dark	lark	bark
30	seen	seed	seek
	seem	seethe	seep
31	dun	dug	dub
	duck	dud	dung
32	beach	beam	beak
	bead	beat	bean
33	did	din	dip
	dim	dig	dill
34	led	shed	red
	wed	fed	bed
35	peas	peal	peach
	peat	peak	peace
36	tease	teak	tear
	teal	teach	team
37	map	mat	math
	mad	mass	man
38	came	cape	cane
	case	cave	cake
39	keel	feel	peel
	reel	heel	eel
40	gold	hold	sold
	told	fold	cold
41	paw	jaw	saw
	thaw	law	raw
42	race	ray	rake
	rate	rave	raze
43	bit	sit	hit
	wit	fit	kit
44	fizz	fill	fib
	fin	fit	fig
45	lame	lane	lace
	late	lake	lay
46	bus	buff	bug
	buck	but	bun
47	cook	book	hook
	shook	look	took
48	hen	ten	then
	den	men	pen
49	meat	feat	heat
	neat	beat	seat
50	heal	heap	heath
	heave	hear	heat

APPENDIX V

PORTION OF TEST DESIGN SHOWING ONE TEST SESSION FOR EACH OF THE FOUR LISTENER GROUPS

Test No.	Group I		Group II		Group III		Group IV	
	MRT List	S/N Ratio in dB	MRT List	S/N Ratio in dB	MRT List	S/N Ratio in dB	PB List	S/N Ratio in dB
1	A1*	0	E3	0	A1	0	1V	0
2	E4	-8	F2	-8	E2	-8	2V	-8
3	C3	0	D1	0	C3	0	3V	0
4	F1	-8	C4	-8	B4	-8	4V	-8
5	B2	0	B2	0	D2	0	5V	0
6	D3	-8	A1	-8	F1	-8	6V	-8
7	F4	0	D2	0	E3	0	2T	0
8	E3	-8	F3	-8	A4	-8	1T	-8
9	C1	0	C2	0	F2	0	4T	0
10	D2	-8	A4	-8	D3	-8	3T	-8
11	B3	0	E1	0	B1	0	6T	0
12	A2	-8	B3	-8	C4	-8	5T	-8

*Letters denote test lists and numbers denote randomizations.

APPENDIX VI

MRT WORDS USED TO OBTAIN CONSONANT-VOWEL RATIOS

A. Initial /s/

List A	List B	List C	List D	List E	List F
six*	six*	six*	six*	six*	six*
six	six	six	six	six	six
six	six	six	six	six	six
six	six	six	six	six	six
six	six	six	six	six	six
sake	sass	sag	seat	sup	sat
sad	same	say	sane	seep	saw
sold	seem	seethe	sack	sap	sung
sud	sum	sub	sun	sale	sent

B. Final /s/

List A	List B	List C	List D	List E	List F
six*	six*	six*	six*	six*	six*
six	six	six	six	six	six
six	six	six	six	six	six
six	six	six	six	six	six
six	six	six	six	six	six
peace	sass	race	test	mass	cuss
bust	bus	rust	bass	pus	dust
pass	just	pace	must	gust	case
nest	vest	west	lace	best	rest

*Spoken as a test item number, the word "six" was used to help provide a sufficient number of words having the desired initial and final consonants.

APPENDIX VII

MEAN /s/-VOWEL RATIOS AND STANDARD DEVIATIONS FOR THE EIGHT SELECTED SPEAKERS

A. Initial /s/-Vowel Ratio

Spkr. No.	List A	List B	List C	List D	List E	List F	$\overline{X}$	S.D.
1	17.3	14.0	15.2	14.8	16.0	16.3	15.6	2.29
2	7.1	7.2	6.9	5.9	7.2	6.3	6.8	2.54
3	13.2	9.0	12.6	11.9	13.2	10.9	11.8	2.52
4	8.3	7.6	8.0	7.1	6.8	7.0	7.5	2.06
5	6.8	6.4	6.4	4.1	6.1	7.6	6.2	2.14
6	15.0	14.3	14.8	12.9	14.0	14.8	14.3	2.62
7	14.6	12.8	13.6	12.9	11.7	13.0	13.2	2.62
8	14.8	13.6	14.0	13.2	13.7	16.0	14.2	2.25

B. Final /s/-Vowel Ratio

Spkr. No.	List A	List B	List C	List D	List E	List F	$\overline{X}$	S.D.
1	23.0	20.8	21.8	21.1	24.0	23.1	22.3	2.65
2	13.1	16.1	14.3	14.9	12.4	15.1	14.3	3.72
3	20.3	17.8	21.2	20.0	22.1	19.2	20.1	3.09
4	10.9	12.9	12.6	11.2	11.4	10.9	11.7	2.74
5	12.8	12.9	13.8	11.4	14.4	13.6	13.2	2.22
6	19.0	20.9	20.4	17.9	18.2	18.8	19.2	3.03
7	12.9	13.1	13.1	12.8	11.4	12.4	12.6	2.13
8	16.7	17.9	17.0	17.0	17.2	18.4	17.4	2.22

APPENDIX VIII

TEST DESIGN OF ONE OF THE TEST SESSIONS[*] CONDUCTED TO OBTAIN SPEAKER INTELLIGIBILITY SCORES

Test No.	MRT List	Speaker No.	S/N Condition
1	A1	1	0
2	D4	4	-8
3	E5	7	0
4	F2	2	-8
5	D5	5	0
6	E2	8	-8
7	B6	6	0
8	C3	3	-8
9	A2	2	0
10	B1	1	-8
11	D6	6	0
12	A5	5	-8
13	B3	7[**]	0
14	C4	8	-8
15	F6	7	0
16	B4	4	-8
17	A3	3	0
18	B6	8[**]	0

[*] The overall test design was balanced over all test sessions. The balance is not necessarily reflected in the test design for single test sessions.

[**] Low vocal effort.

BLANK PAGE

APPENDIX IX

MRT WORDS USED TO OBTAIN CONSONANT-VOWEL RATIOS

A. Initial /t/

List A	List B	List C	List D	List E	List F
tab	tan	tam	tang	tack	tap
tent	took	tale	test	toil	tick
ten	tear	tip	till	tame	tot
top	told	tin	teal	team	teak
teach	two	tease	two	two	two
two*	two*	two*	two*	two*	two*
two	two	two	two	two	two
two	two	two	two	two	two
ten*	ten*	ten*	ten*	ten*	ten*

B. Final /t/

List A	List B	List C	List D	List E	List F
bust	just	rust	rate	gust	dust
kit	bit	fit	must	wit	pit
bat	bent	went	sit	rent	hit
tent	feat	heat	dent	beat	sent
meat	pat	beat	seat	heat	neat
nest	vest	west	test	best	sat
cut	sit	not	pot	lot	rest
hot	got	but	kit	peat	tot
late	mat	eight*	eight*	eight*	fit

*Spoken as test item numbers, the words "two", "ten", and "eight" were used to help provide a sufficient number of words having the desired initial and final consonants.

APPENDIX IX (Continued)

C. Initial /k/

List A	List B	List C	List D	List E	List F
kick	king	kill	cook	kin	kill
kick	cape	kid	kit	cave	cold
kit	coil	cane	cop	cud	case
came	cub	came	cake	keel	cuss
cut		cuff	cup		

D. Final /k/

List A	List B	List C	List D	List E	List F
kick	lick	sick	pick	wick	tick
book	took	shook	puck	hook	look
sake	rake	peak	cook	sick	teak
kick	lake	pack	sack	buck	seek
duck	pick	back	cake	tack	bark
hark	dark	mark	lark	park	beak

E. Initial /ʃ/

shed

shook

shop

F. Final /tʃ/

teach

beach

peach

APPENDIX X

MEAN CONSONANT-VOWEL RATIOS AND STANDARD DEVIATIONS FOR FOUR SELECTED SPEAKERS

A. Initial /t/-Vowel Ratio

Spkr. No.	List A	List B	List C	List D	List E	List F	$\bar{X}$	S.D.
1	21.6	19.7	19.4	19.7	20.1	21.4	20.3	2.39
4	19.9	20.3	19.7	19.8	20.8	20.0	20.1	2.92
6	24.0	21.7	20.6	21.7	23.8	22.9	22.5	3.20
8	21.2	24.1	21.9	22.3	20.6	23.2	22.2	2.38

B. Final /t/-Vowel Ratio

Spkr. No.	List A	List B	List C	List D	List E	List F	$\bar{X}$	S.D.
1	33.3	27.1	30.2	31.4	32.8	34.3	31.5	4.01
4	27.9	25.3	24.2	23.9	25.0	27.8	25.7	5.21
6	31.9	29.6	29.0	28.7	27.0	29.7	29.3	5.35
8	23.4	25.1	24.2	24.8	24.7	26.2	24.7	2.92

APPENDIX X (Continued)

C. Initial /k/-Vowel Ratio

Spkr. No.	List A	List B	List C	List D	List E	List F	$\overline{X}$	S.D.
1	13.8	16.8	17.0	14.0	15.3	18.3	15.9	1.80
4	19.2	20.3	20.0	23.4	21.5	23.0	21.2	1.70
6	25.8	24.0	22.2	24.2	23.3	25.0	24.1	1.26
8	18.6	21.3	21.0	19.8	19.5	19.3	19.9	1.04

D. Final /k/-Vowel Ratio

Spkr. No.	List A	List B	List C	List D	List E	List F	$\overline{X}$	S.D.
1	31.7	28.2	29.8	31.8	30.0	32.7	30.7	1.66
4	32.5	27.5	28.7	32.0	29.8	28.7	29.9	1.99
6	34.0	31.7	32.0	34.2	36.7	31.5	33.4	2.02
8	26.3	30.2	28.3	20.2	30.5	32.5	29.7	2.12

APPENDIX X (Continued)

E. Initial /ʃ/-Vowel Ratio

Spkr. No.	List B	List C	List F	$\overline{X}$	S.D.
1	13	15	15	14.3	1.15
4	12	11	15	12.7	2.08
6	13	16	15	14.7	1.53
8	9	10	14	11.0	2.65

F. Final /tʃ/-Vowel Ratio

Spkr. No.	List A	List B	List F	$\overline{X}$	S.D.
1	22	21	22	21.7	.58
4	10	18	11	13.0	4.36
6	15	17	15	15.7	1.15
8	14	13	15	14.0	1.00

BLANK PAGE

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)*

Bolt Beranek and Newman, Inc.
Cambridge, Massachusetts

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

Intelligibility Test Methods and Procedures for the Evaluation of Speech Communication Systems

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

Final Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Carl E. Williams, Michael H. L Hecker, Kenneth N. Stevens, Barbara Woods

6. REPORT DATE

December 1966

7a. TOTAL NO. OF PAGES

72

7b. NO. OF REFS

10

8a. CONTRACT OR GRANT NO.

AF19(628)-5659

b. PROJECT NO.

2808

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

ESD-TR-66-677

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

BBN Report No. 1442

10. DISTRIBUTION STATEMENT

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Decision Sciences Laboratory
Electronic Systems Division, AFSC
USAF, L.G. Hanscom Field
Bedford, Massachusetts 01731

13. ABSTRACT

In further exploring the Modified Rhyme Test (MRT), a recently developed intelligibility test designed for the evaluation of speech communication systems under operational military conditions, research has been conducted in the following areas: (a) the relation between MRT scores and other intelligibility test scores for various types and levels of speech distortion; (b) the influence of the closed-response format and listening experience on MRT scores; and (c) speaker intelligibility and the selection of speakers for recording the test lists. The present report describes the work undertaken in each of these areas. The ultimate objective of the work is the development of valid procedures for the efficient evaluation of speech communication systems.

The major experimental results demonstrate that (1) the relation between scores obtained with different intelligibility test materials is not unique but depends considerably on the type of speech distortion employed, (2) neither the closed-response format nor prior listening experience appreciably affects MRT scores, and (3) less intelligible speakers tend to be those whose voiceless consonants are generated with lower intensity, particularly in word-final position.

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Psychoacoustics						
Speech						
Speech Transmission						
Verbal Behavior						
Human Engineering						