

AD-A126 109

VOICE-COMMUNICATION TESTING USING NAIVE AND EXPERIENCED 1/1
COMMUNICATORS(U) NAVAL RESEARCH LAB WASHINGTON DC
A SCHMIDT-NIELSEN 03 FEB 83 NRL-8655 SBI-AD-E000 533

UNCLASSIFIED

F/G 17/2

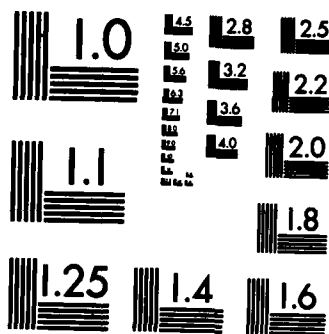
NL

END

FILED

XX

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

ade 000 533
(2)

NRL Report 8655

ADA 126109

Voice-Communication Testing Using Naive and Experienced Communicators

ASTRID SCHMIDT-NIELSEN

*Communication Systems Engineering Branch
Information Technology Division*

February 3, 1983

DTIC
MAR 15 1983
A



NAVAL RESEARCH LABORATORY
Washington, D.C.

83 08 15 065

Approved for public release; distribution unlimited.

DTIC FILE COPY

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NRL Report 8655	2. GOVT ACCESSION NO. AD-A126109	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) VOICE-COMMUNICATION TESTING USING NAIVE AND EXPERIENCED COMMUNICATORS		5. TYPE OF REPORT & PERIOD COVERED Interim report on a continuing NRL problem.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Astrid Schmidt-Nielsen		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Research Laboratory Washington, DC 20375		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61158N RR021-05-42 75-1596-0-2 and 75-0126-0-2
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217		12. REPORT DATE February 3, 1983
		13. NUMBER OF PAGES 17
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Voice testing Voice communications User experience Context effects		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Subjects with and without communication experience evaluated a variety of digital voice processors with data rates from 800 to 32,000 bits per second. Two conversational tests were used: one with a restricted vocabulary and one with an open vocabulary. Ratings of the voice processors were similar in all conditions, but the experienced communicators used more compensatory behaviors when talking over the poor voice systems. The inexperienced communicators developed similar compensations but used them less often. The results suggest that ratings by the (Continues)		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT (Continued)

subjects are reliable measures of the performance of voice processors. Objective measures of talking performance are less reliable but show the same trends. Because the experienced subjects were more tolerant of degradation, the ratings did not discriminate among systems as well as those of the inexperienced subjects. However, both groups showed more errors and longer test times with more degraded systems. Thus the best possible system should be chosen even for experienced users, but the choice may better be made with ratings by inexperienced test subjects, and the testing technique need not necessarily closely resemble the real-world application.

CONTENTS

INTRODUCTION	1
METHOD	2
Conversational Tests	2
Subjects	2
Voice Systems	5
Test Procedure	5
RESULTS	6
Ratings	6
Talker Behaviors	6
Constrained Task	6
Unconstrained Task	9
DISCUSSION AND CONCLUSIONS	9
ACKNOWLEDGMENTS	13
REFERENCES	13

[illegible]

VOICE-COMMUNICATION TESTING USING NAIVE AND EXPERIENCED COMMUNICATORS

INTRODUCTION

Conversational tests are especially useful for evaluating voice communication systems when the voice quality is degraded. Standard intelligibility tests such as the Diagnostic Rhyme Test (Voiers, 1977) or the Modified Rhyme Test (House, Williams, Hecker, and Kryter, 1965) are highly reliable measures of intelligibility, but they are one-way tests and do not give the talker an opportunity to adapt to the degraded communication situation. A conversational test (such as described in Schmidt-Nielsen and Everett, 1982) involves an interactive two-way communication task, so that users can try to overcome any loss in voice quality. Such a test permits distinction between compensable and noncompensable degradations.

In the real world, communication systems may be used in contexts that vary widely. A pilot landing an aircraft uses a highly standardized vocabulary and at any given time is expecting one of a limited set of messages. The pilot is also experienced with poor communication links and should be able to recognize and understand unexpected emergency messages in addition to standard messages, but an unexpected and irrelevant comment such as "happy birthday" may be met with confusion or a request to repeat. In contrast, the President talking on the telephone to a high-level advisor must be able to converse freely on a wide variety of topics, and often neither party will be experienced at communicating over a poor voice link.

Thus user experience and the communication context are important factors in the usability of a degraded voice link. Communication officers may be very satisfied with a military communication system that sounds almost unintelligible to a naive listener. A more commonplace example is the effect of context and experience on conversation over a CB radio.

The effects of context and of limiting the expected message set were demonstrated in the classic paper by Miller, Heise, and Lichten (1951), and similar results have been obtained by Voiers (1982) and other investigators. These investigators used different noise levels to degrade the speech and conducted intelligibility tests while varying the expected vocabulary size. They found that the smaller the size of the test vocabulary, the higher the intelligibility performance. Providing other forms of contextual support, such as sentences instead of isolated words, also improved intelligibility performance.

In the present experiment experienced and inexperienced users were tested using two communication tasks: one highly structured with a limited vocabulary and the other relatively unstructured with an unconstrained vocabulary. The tests included a broad range of digital voice-processing systems. The principal purpose of the tests was to determine how user experience and communication context influenced the subjects' ratings of the voice systems, and the secondary purpose was to determine how experience and task constraints affected the performance of the communicators when they talked over voice channels of varying difficulty.

A test that more closely reproduces the real-world application of interest to the potential user would seem to be more useful for evaluating voice systems. However, the high correlations

found among different tests, locations, and methods of measuring intelligibility or quality (Montague, 1960, Voiers, 1980, 1981, Goodman and Nash, 1982, and Schmidt-Nielsen and Everett, 1982) suggest that the various test methods may all be reasonably valid and that convenience and reliability may be better criteria for selecting a voice testing technique than how closely it resembles a real-world application.

METHOD

Conversational Tests

A conversational test usually has a task or problem that requires the participants to exchange information in a fairly natural way and a questionnaire or evaluation form to be filled out when the task is completed. All subjects participated in two conversational tests. The Free-Conversation Test (Butler and Kiddle, 1969) has an open format and a potentially unlimited vocabulary, whereas the NRL Communicability Test (Schmidt-Nielsen and Everett, 1982) has a well-defined format with an effective vocabulary limited to only a few words.

The task for the Free-Conversation Test is a picture-comparison task. Each participant is given one of a pair of photographs taken a short time apart, such as illustrated in Fig. 1, and the two participants discuss the pictures until they agree on which one came first. The picture pairs were selected from a large set of pairs of Polaroid pictures taken by the experimenters. A pretest with a separate set of subjects was used to select pairs that were roughly comparable in difficulty. Although these are not the same pictures that are used for the Free-Conversation Test in Britain, two colleagues who have had experience with the British test assure us that they are comparable in their nature and general difficulty. After completing the task, subjects rate the effort required to communicate using a 5-point scale from "complete relaxation possible; no effort required" to "extreme effort required; prolonged conversation impossible."

The NRL Communicability Test is based on a pencil-and-paper battleship game (Fig. 2). Each participant draws two ships in a game grid, and players take turns shooting at each other by specifying cells in the grid (saying, for example, "row 4"). After one player wins by sinking both of the opponent's ships, the players answer the four questions at the end of the form, which rate effort required to communicate, (un)naturalness, need to speak carefully, and overall acceptability.

Subjects

The inexperienced communicators were six students recruited through an advertisement in the University of Maryland newspaper. They were paid at an hourly rate for their participation in the test sessions.

The experienced communicators were six members of the NRL amateur radio club, who volunteered to participate in the tests. All had hundreds of hours of ham radio experience. In addition, four of the six were or had been in the U.S. Navy, and one of these was a former Navy communication officer. Even though it was not possible to obtain active military communicators, this group was representative of many of the important characteristics of experienced communicators.

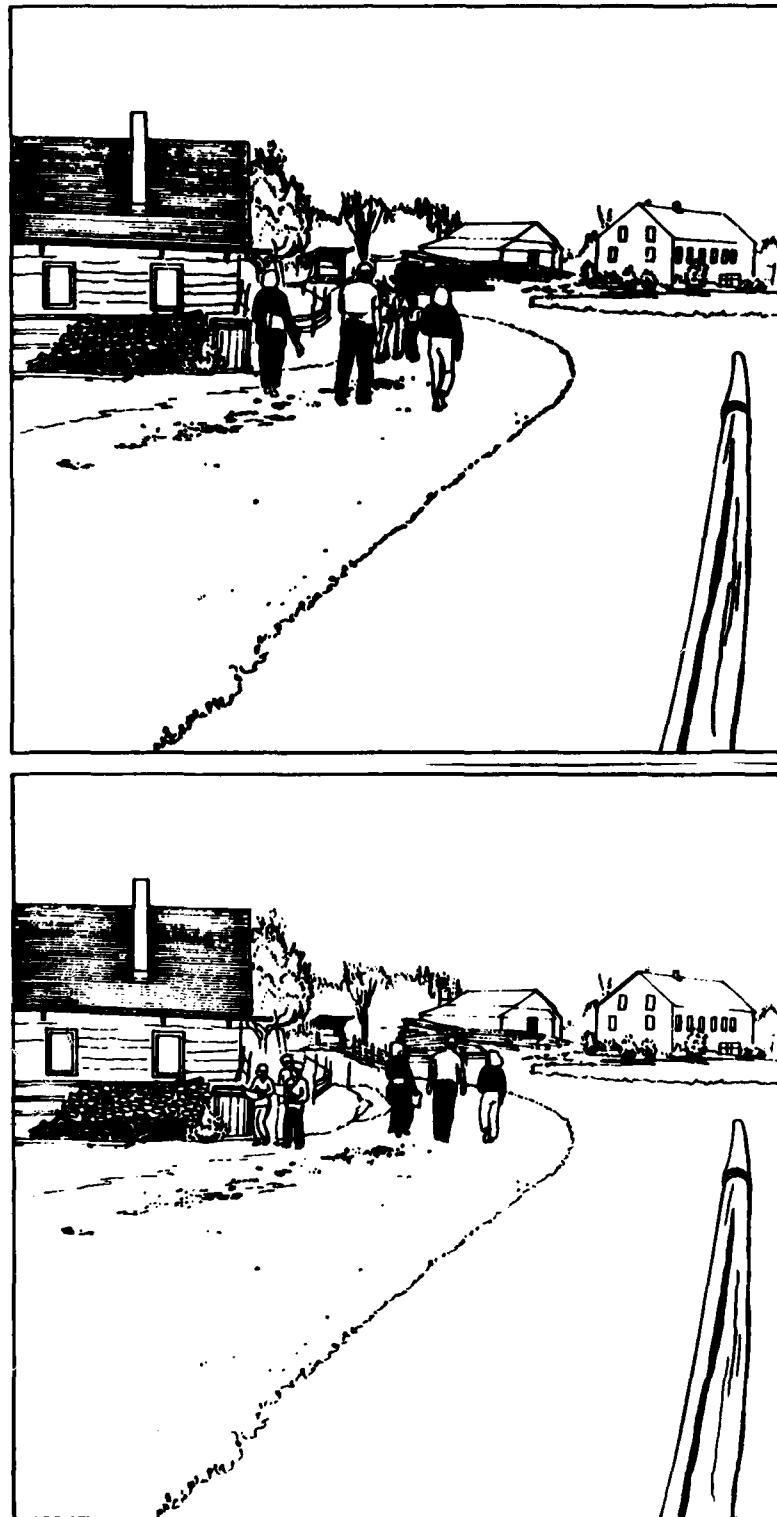
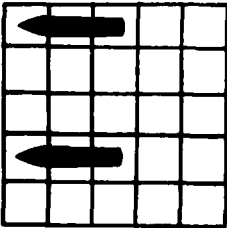
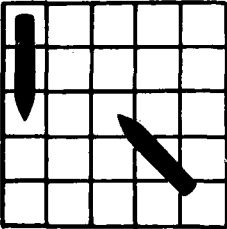
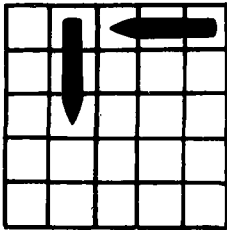


Fig. 1 - Drawing of a typical pair of photographs for the Free-Conversion test, with each subject being given one of the pairs. The subjects' task is to decide whose photograph came first.

**INSTRUCTIONS FOR THE
NRL COMMUNICATIONS TEST**

This short communications test for evaluating voice systems involves a game. The game is played on the two grids printed on the test form. You and your opponent both draw two ships in the left-hand grid such that each occupies three squares vertically, horizontally, or diagonally as in the following examples:

The object of the game is to be the first to sink both of the other person's ships. A ship is sunk when all three of its squares have been hit by the opponent's shots. You and your opponent shoot in turn, with each turn consisting of one shot. To shoot, you specify a cell in the grid (alpha 2, charlie 1, delta 4, etc.). Your opponent marks the specified cell and tells you whether it was a hit or a miss. Keep track of your shots in the right-hand grid (being sure to mark which ones are hits), and keep track of your opponent's shots at you in the left-hand grid.

The game begins when you and your opponent have placed your ships and agree to begin. After the game, please answer the questions at the bottom of the test form.

NRL COMMUNICATIONS TEST

TALKER _____ DATE _____
TALKING WITH _____ TEST # _____

Opponent's shots at you

A				
B				
C				
D				
E				

Your shots at opponent

A				
B				
C				
D				
E				

After the game, please answer the questions below. For each question, mark the space that best describes your opinion.

- EFFORT required to communicate

No special effort	Moderate effort	Extreme effort: normal conversation impossible
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
- UNNATURAL voice quality

Completely natural	Moderately distorted	Extremely unnatural
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
- Need to SPEAK CAREFULLY

Can talk normally and casually	Talk more carefully	Extreme care in talking and pronouncing
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
- Overall ACCEPTABILITY of the system

Excellent	Moderately acceptable	Unacceptable
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

(a) Instruction sheet (b) Test form

Fig. 2 - Instruction sheet and test form for the NRL communications test

Voice Systems

The tests compared ten voice links (Table 1), with a wide range of intelligibility being covered. The processors at 9.6 kb/s and above were at least as good as an ordinary telephone link, and the worst two processors were barely usable. The three LPC processors could be expected to be quite similar, but if any one might be slightly better than the others, it should be LPC(a) (the most recent version). Brand X has now been corrected by the manufacturer, but at the time it was useful as a difficult system.

Table 1 — Voice Processing Systems Tested

System	Data Rate (kb/s)	Abbreviation
Direct unprocessed channel	—	Unprocessed
Continuously-variable-slope delta modulation	32	CVSD 32
	16	CVSD 16
Residual excited linear predictive coding	9.6	REL P 9.6
Adaptive predictive coding	9.6	APC 9.6
Linear predictive coding: implementation a	2.4	LPC(a) 2.4
implementation b	2.4	LPC(b) 2.4
implementation c	2.4	LPC(c) 2.4
Faulty simulation of a digital vocoder	2.4	Brand X
An experimental system	0.8	Experimental 0.8

Test Procedure

Subjects were tested two at a time. One talking station was located in a sound booth, and the other in a separate room. Each station was equipped with a telephone-type handset with a Roanwell confidencer model-240-100002-653 dynamic microphone. The handsets were wired for push-to-talk by a system of relays simulating a half-duplex channel. Only one person at a time had use of the channel. The person who was listening heard the voice processed through the voice processor being tested, while the person who was talking heard only a normal unprocessed sidetone. A single processor in a loop-around mode could be used for both input and output, since the signal only had to go in one direction at a time. The control station, operated by the experimenter, could override the talker stations at any time, so the experimenter could interrupt, correct, give instructions, etc. The control station had a switching system for changing from one processor to another. A tape recorder at the control station recorded all of the sessions. Processed and unprocessed versions were recorded separately on the two channels of the tape.

The subjects were given three training trials on each of the two conversational tests. For each conversational test, they talked once each over the unprocessed channel, over the LPC(c) system at 2.4 kb/s, and over the 0.8-kb/s system. The training served both to familiarize the subjects with the two tests and to expose them to good and bad voice systems, thereby giving them a reference frame for using the rating scales.

Each subject was tested three times with each conversational test on each of the ten voice systems. Subjects came to the laboratory for 3-hour test sessions several times a week until testing was completed. Each test session consisted of three or four 30- to 45-minute testing periods (five voice systems per testing period) with 15 minutes of rest between testing periods. Subjects were not told which voice system they were using in any test. One conversational test was used for all five tests within a testing period, and the two conversational tests were alternated between testing periods. The order in which the voice processors were tested was counterbalanced for subject pairs, conversational tests, and testing periods (Schmidt-Nielsen and Everett, 1982.)

RESULTS

Ratings

Traditionally the scores 0 through 4 are assigned to the five categories of the Free-Conversation Test. These numbers were multiplied by 25 to make the scale more comparable to that of the NRL communicability test, which assigns the numbers 5, 20, 35, 50, 65, 80, and 95 to the seven categories. The mean ratings are shown in Fig. 3. The Newman-Keuls test (Winer, 1971) for pairwise differences was carried out for each set of scores. Brackets indicate categories within which differences were not statistically significant ($p < 0.05$). The overall rankings of the voice systems were very similar for both groups of subjects and for the two conversational tasks, and any minor differences fell within categories where the systems were not significantly different from one another. On the whole the inexperienced communicators gave better discrimination among voice systems; that is, there were more significant differences. The experienced communicators gave the poorer systems somewhat higher ratings than did the inexperienced communicators, and it seems to be this higher tolerance for degradation that accounts for their making less fine distinctions among poor systems. The successful compensations made by this group may also have obscured processor differences.

Correlations, Pearson's r , between the means for the two sets of subjects were 0.958 for the NRL test and 0.956 for the Free-Conversation Test. Correlations between the two tests were 0.959 for the inexperienced group and 0.986 for the experienced group. Thus for comparing voice systems the two sets of talkers and the two task types yield similar results. Hence test selection and subject population can be based on convenience and reliability and do not have to reflect the user's intended application.

Talker Behaviors

The voice systems can be divided into easy systems and difficult or poor systems. The processors with data rates of 9.6 kb/s and higher caused no difficulties in communicating, even though some were of higher quality than others. The five easy systems will be grouped together for comparisons. The processors at 2.4 kb/s and below clearly caused some difficulties in communicating accurately. Both the experienced and the inexperienced groups had to find ways of compensating for the degradation by the five poor systems in order to complete the communication task.

Constrained Task

The NRL Communicability Test with its limited vocabulary and structured format for information exchange accentuated the differences in the ways the two groups of talkers compensated when talking over the poor voice systems. With the constrained task, possible messages are

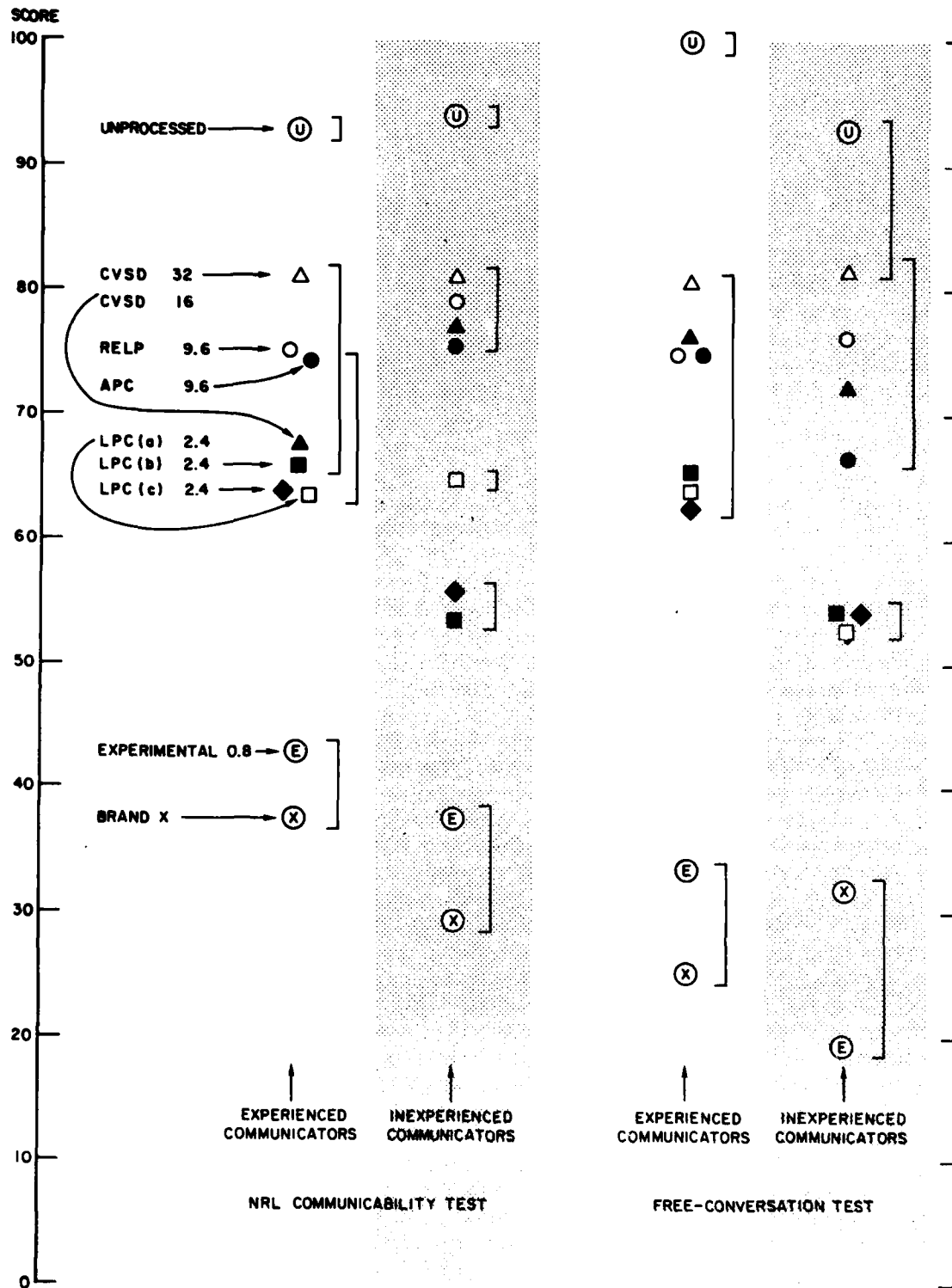


Fig. 3 - Ratings by the subjects of the ten voice systems listed in Table 1. Brackets indicate differences not statistically significant (Newman-Keuls test, $p < 0.05$).

limited, but each message must be understood exactly. Words that are easily confused must be correctly received before the game can continue. In the battleship game, the words "hit" and "miss" are both single-syllable words with the same center vowel and are almost impossible to tell apart when heard over some of the poorer systems.

As soon as the experienced communicators knew they were talking over a poor system, they put into use a readily available set of coping behaviors. They repeated almost every phrase (which they called double-talk): "My shot, my shot, delta one, delta one." They verified the information received before continuing: "I roger your delta one." They also talked more slowly and spoke important words with exaggerated articulation: "thu-ree." They made substitutions for words that might be easily misunderstood. Thus, "hit" and "miss" might become "affirmative" and "negative," and numbers as in "echo four" were often counted out: "Echo fo-wer, one, two, three, four." All of these conventions will be recognized by radio hams and anyone with communication experience. Even with these conventions available, the problems were not always resolved immediately. For example, one pair tried, "that is a miss, a Mary," which seemed fine until one of them scored "a hit, a Harry," not noticing immediately that "Harry" and "Mary" are even more similar than "hit" and "miss." They finally came up with "Honolulu" and "Mexico."

The inexperienced talkers did not compensate automatically as soon as they encountered a poor system. However, they eventually developed compensations similar to those of the experienced group. They also made substitutions for "hit" and "miss," although the word pairs they chose were different: "boom" and "miss," "yes" and "no," "hit" and "negative." They confirmed some messages but not as often as the experienced group. In general, they developed many of the same types of compensation, but they compensated no more than was necessary to complete the task, whereas the experienced group used a larger set of compensations and used them more of the time.

An observer unfamiliar with the test design and the purpose of the tests listened to the unprocessed channel of the tape recordings and counted phrase repetitions, requests for repeats, and errors and misunderstandings for each game. The time to complete each game was also measured and divided by the number of moves in the game. Figure 4 shows that the experienced communicators spontaneously used phrase repetitions consistently and that the inexperienced group rarely did. Consequently the inexperienced group had more requests for repeats. The inexperienced communicators similarly had more misunderstandings and errors per game than the experienced group. Even though most mistakes were eventually discovered and corrected, the smaller number of errors and misunderstandings on the part of the experienced communicators could be important in real life when a mistaken command could have serious consequences. The tendency for the experienced group to verify messages would also make it likely that they would catch errors sooner, but the variety of ways in which an error could affect the game made this elusive to quantify. Because of their extensive use of repetitions, the experienced group took somewhat longer for the entire process. Loss of voice quality increased the time and the errors for both groups, but especially in this situation the experienced group traded loss in time for higher accuracy.

Both the experienced and the naive communicators used compensatory behaviors, but the experienced group already had the compensations available to them and used them extensively in any degraded context. The naive group had to develop compensations and made only the minimum adjustments required to succeed in communicating over a given system. The same types of successful adjustments were used by both groups. They changed their speech pattern by talking more slowly and at the same time articulating more carefully. An unsuccessful adjustment that was more prevalent among the naive than among the experienced talkers was to talk more loudly. Since this is a fairly common reaction, the experienced group had probably learned to suppress this tendency.

Another shared compensation involved changes in vocabulary or language habits: repetition and alternative ways of saying things. The goal of using an alternative vocabulary is to make the messages more distinct from one another so that the correct interpretation is more easily perceived. Thus "boom" and "miss" are more distinct than "hit" and "miss" because the vowels are dissimilar, "negative" and "hit" are more distinct because they have a different number of syllables, and "Honolulu" and "Mexico" or even phrases are more distinct because they are longer.

These are well-known principles that can be seen in the vocabulary and standard phrases used in a variety of military and other constrained communication tasks (such as the phonetic alphabet and air-traffic-control vocabulary). Although good differentiation was achieved according to these principles, some of the intermediate strategies suggest that the subjects were not aware of any general principles.

Unconstrained Task

Although the Free-Conversation Test permits a larger and more unconstrained vocabulary, the talkers do have the shared context of two pictures depicting similar scenes, which helps them understand each other. There is also some redundancy in a description, so that it is not critical that every word be understood. Understanding part of the message or getting the gist is sufficient to be able to formulate an answer and keep the discussion going. There were fewer differences between the two groups for this task in that both groups tended to slow down and articulate more carefully. But again with the poorer systems the experienced group used more spontaneous repetitions while the inexperienced group had more requests for repeats (Fig. 5). The experienced group more often made constructive requests ("slow down" or "talk more quietly, you're clipping" instead of "what?").

The number of exchanges (each alternation of speakers counting as one exchange, so that if A talked twice and B three times, this counted as five exchanges) and the time to complete the task depended on both the difficulty of talking over the voice system and on how difficult a given picture pair was for the individual subjects. Figure 5 shows a trend toward more exchanges with more difficult voice systems. The experienced group again took more time to complete the task. In the case of the NRL test most of the differences in task completion time could be attributed to a greater use of repetitions and message verification. However, listening to the tape recordings of the Free-Conversation Test for the two groups gave the strong impression that the communicator group was conscientious about arriving at the correct choice of which picture came first, whereas the student group seemed satisfied merely to reach agreement. The time difference probably reveals less about coping strategy than about attitude toward the task.

The trends in task time due to voice processors were not as clear for the Free-Conversation Test as for the NRL test. Even though the picture pairs were pretested and were also counter-balanced for the different voice processors, there were differences in the difficulty that individual subject pairs had in reaching a solution. These differences affected objective measures such as solution time but did not noticeably influence the ratings. The subjects seemed to be able to rate the voice systems consistently and to divorce talking problems from solution difficulty.

DISCUSSION AND CONCLUSIONS

The primary effect of communicator experience was a set of compensatory behaviors readily available to put into use as soon as talking became difficult. The naive communicators developed

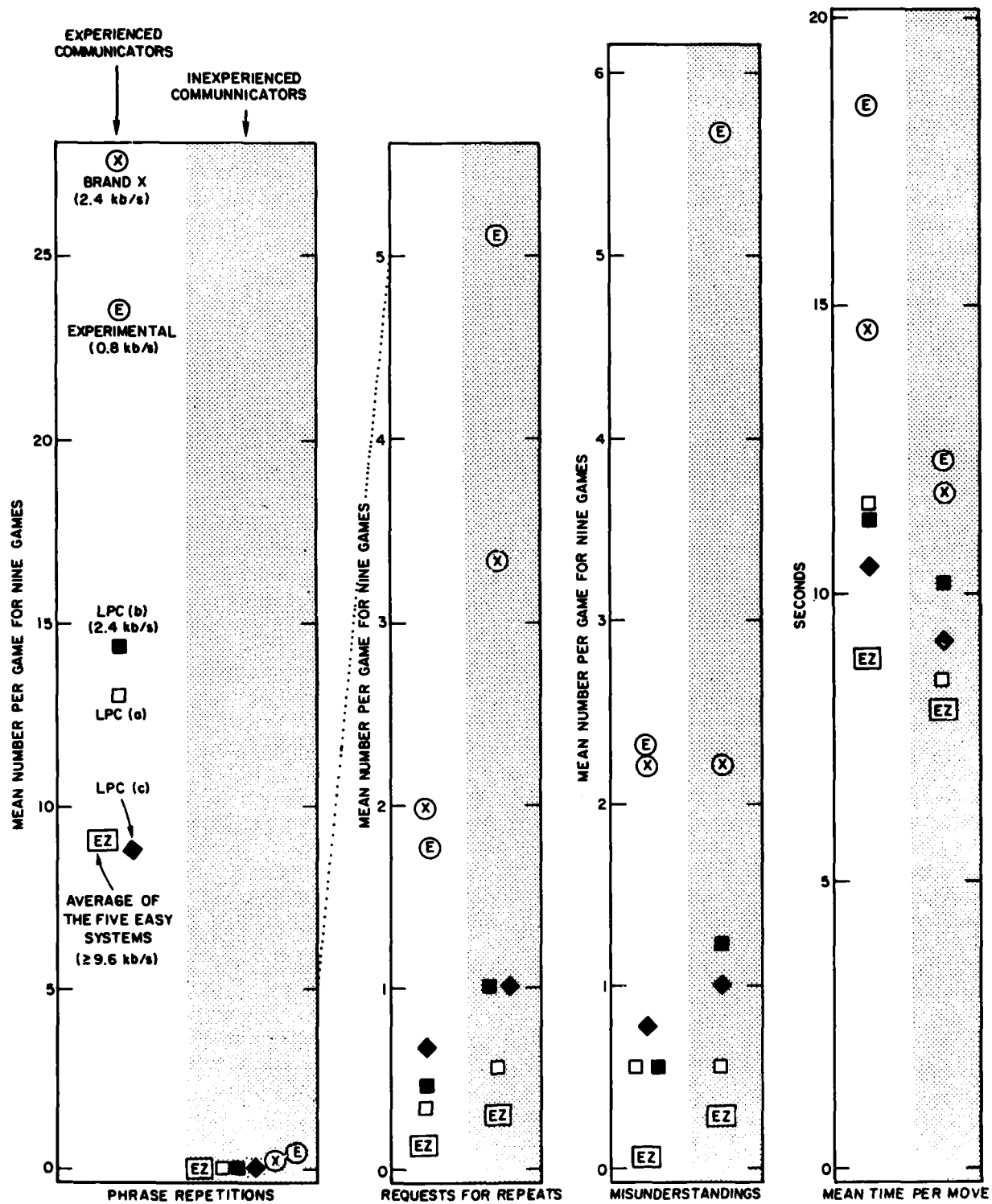


Fig. 4 - Results of objective measures of subjects talking on the ten voice processing systems for the NRL Communicability Test. Since each of six subjects (three pairs of subjects) in each of the two groups of communicators were tested three times on each system, there were nine tests on each system with each group. In each test the subjects' task was the constrained task of playing a battleship game (Fig. 2).

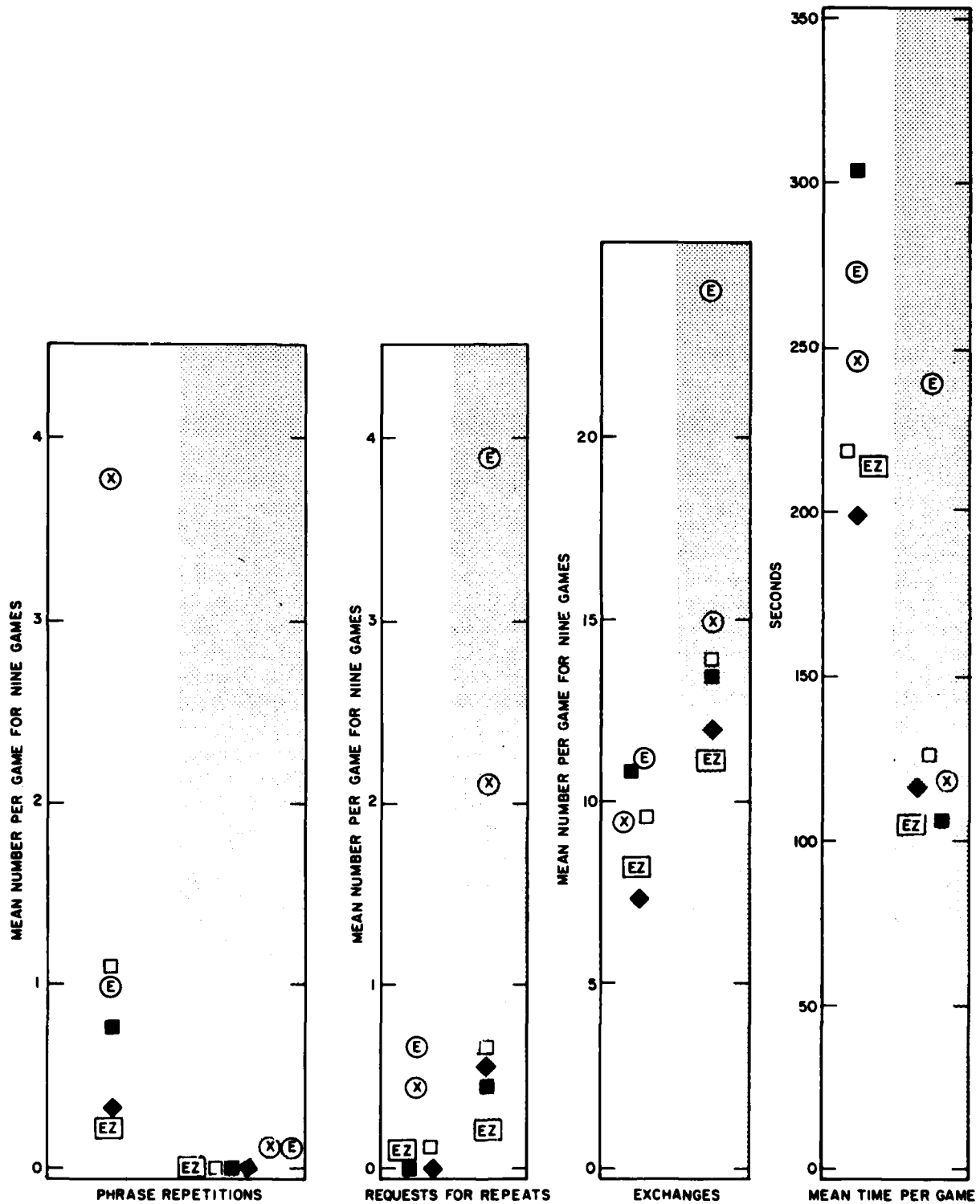


Fig. 5 - Results of object measures for the Free-Conversation Test. In each test the subjects' communication task was the game of deciding who had the first of a pair of photographs (Fig. 1).

the same compensations of slow and carefully articulated speech and vocabulary substitutions but used them only when normal talking failed. The additional compensations that were used principally by the experienced group were phrase repetition (double-talk), message verification, and constructive requests. In spite of or perhaps because of having and using more effective coping behaviors, the experienced group took longer to complete the communication tasks than did the inexperienced group. Since both groups generally completed the tasks successfully, it might seem paradoxical that the inexperienced group was faster. However, the experienced group communicated more accurately, as evidenced by fewer errors and not as many requests for repeats. The tasks used here were intended to get the subjects to exchange information long enough to evaluate the voice systems, and a longer time for task completion had no negative effects on the evaluation, although excessively long times could lead to boredom and also increase the cost of testing. To compare the relative merits of voice systems, a test need not mimic every aspect of real applications. So long as relative performance is consistent across a variety of situations, it is still possible to select the best system. There is no reason to believe that the experienced talkers would not perform well in situations where time was important, and there is good reason to believe they would communicate more accurately.

The subjects' ratings of the voice processors were similar for both communication tasks and for the two experience groups. Because the experienced subjects were more tolerant of degradation, their ratings did not discriminate among systems as well as those of the inexperienced subjects. Possible explanations for this are that the experienced group may simply have talked over enough poor voice links to find such links acceptable and that their extensive use of compensatory behaviors with all of the poorer systems may have obscured any real differences in the difficulty of communicating. Since the systems were ranked similarly by both groups, it would seem that more information can be gained by testing with relatively naive subjects. Even though the experienced group found degraded systems more acceptable and made fewer errors than the inexperienced group, losses in voice quality had consequences for performance for both groups in that both groups showed similar trends toward increased errors and longer task time. Thus it is important to select the best possible system even for experienced users, but the choice may perhaps be better made using naive test subjects.

Performance of voice processors has so far been considered primarily in terms of subjects' ratings. The measures that have been used to compare the performance of the users, namely, errors, requests for repeats, and task time, could also be used as more objective measures of processor performance. Unfortunately all of these measures are highly susceptible to variability from sources that are irrelevant to processor performance, such as jokes, inattention, and minor distractions. The subjects seemed to be able to discount these effects in making their ratings. Even though the seemingly objective measures were less stable as measures of processor performance than the subjective ratings, these measures all showed trends that were consistent with the results of the ratings. This finding supports the validity of using ratings as a measure of processor performance. For the two experience groups both conversational tests took longer as processor difficulty increased. Both tests also have variations in game length that are the result of causes other than processor differences. The variability in game length caused by differences in the number of moves on the NRL test can be removed, but differences in difficulty among picture pairs on the Free Conversation Test are almost impossible to eliminate, since they also differ for individual people. Therefore the time differences for the NRL test more clearly reflected actual differences in the difficulty of talking over the various processors than did the time differences for the Free-Conversation Test.

In using test scores to evaluate voice systems, the potential user should distinguish between the problem of comparing voice systems (where test reliability is an important issue) and that of predicting usability for a specific real-world context (where validity issues come to the fore). The

results reported here show that, although the behavior of the users differs with user experience and with the communication context, the actual scores for a group of voice systems were ranked very similarly in all of the conditions and show that the rankings agree with other performance measures. This suggests that one can select the test method that gives the most reliable results and the best discrimination among voice systems and expect the results to be valid for a variety of situations. The decision about how high a system should score for a specific application can then be based on environmental considerations such as user experience, the degree of contextual support, background noise, or other relevant factors.

ACKNOWLEDGMENTS

The author thanks Stephanie Everett for her extensive help in recruiting and testing the subjects and in sorting and analyzing the data, Howard Murphy for help in setting up and maintaining the test equipment, members of the NRL Radio Club for generously volunteering their time, and D. L. Horton of the Center for Language and Cognition, University of Maryland, for help in obtaining subjects.

REFERENCES

- Butler, L. W., and L. Kiddle, "The Rating of Delta Sigma Modulating Systems, With Constant Errors, Burst Errors and Tandem Links in a Free Conversation Test Using the Reference Speech Link," Signals Research and Development Establishment, Ministry of Technology, Report 69014, Feb. 1969.
- Goodman, D. J., and R. Nash, "Subjective Quality of the Same Speech Transmission Conditions in Seven Different Countries," IEEE Trans. on Communications COM-30, 642-654 (1982).
- House, S., E. Williams, L. L. Hecker, and K. D. Kryter, "Articulation-Testing Methods: Consonantal Differentiation With a Closed-Response Set," J. Acoust. Soc. Am. 37, 158-166 (1965).
- Miller, G. A., G. A. Heise, and W. Lichten, "The Intelligibility of Speech as a Function of the Context of the Test Materials," J. Experimental Psychology 41, 329-335 (1951).
- Montague, W. E., "A Comparison of Five Intelligibility Tests for Voice Communication Systems," U.S. Navy Electronics Laboratory (now Naval Ocean Systems Center), San Diego, NEL Report 977, June 1960.
- Schmidt-Nielsen, A., and S. S. Everett, "A Conversational Test for Comparing Voice Systems Using Working Two-Way Communication Links," NRL Report 8583, June 1982.
- Voiers, W. D., "Diagnostic Evaluation of Speech Intelligibility," in *Speech Intelligibility and Recognition*, M. E. Hawley, editor, Stroudsburg, Pa., Dowden, Hutchinson, and Ross, 1977.
- Voiers, W. D., "Interdependencies Among Measures of Speech Intelligibility and Speech Quality," in *Proceedings of the 1980 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Denver, Apr. 1980.
- Voiers, W. D., "Uses, Limitations, and Interrelations of Present-Day Intelligibility Tests," presented at the National Electronics Conference, Chicago, Oct. 1981.

SCHMIDT-NIELSEN

Voiers, W. D., "Some Thoughts on the Standardization of Psychological Measures of Speech Intelligibility and Quality," in *Proceedings of the Workshop on Standardization for Speech I/O Technology*, National Bureau of Standards, Gaithersburg, Md., Mar. 18-19, 1982.

Winer, B. J., *Statistical Principles in Experimental Design*, New York, McGraw-Hill, 1971.

4-8
DT