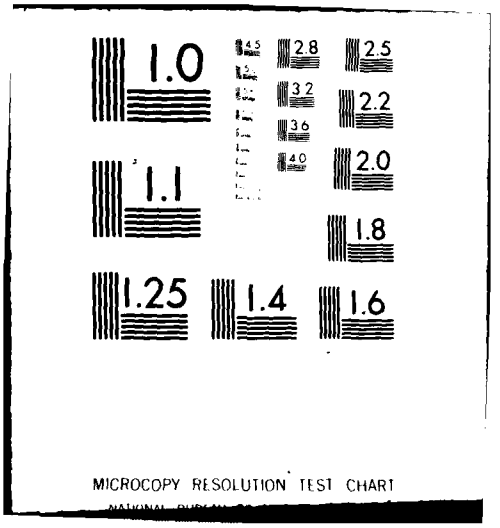
AD-A116 054 NAVAL RESEARCH LAB WASHINGTON DC F/G 17/2
A CONVERSATIONAL TEST FOR COMPARING VOICE SYSTEMS USING WORKING--ETC(U)
JUN 82 A SCHMIDT-NIELSEN, S S EVERETT
UNCLASSIFIED NRL-8583 NL

END
DATE
FILMED
7-82
DTIC


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6
MICROCOPY RESOLUTION TEST CHART

AD A116054

(2)

NRL Report 8583

A Conversational Test for Comparing Voice Systems Using Working Two-Way Communication Links

A. SCHMIDT-NIELSEN AND S. S. EVERETT

Communication Systems Engineering Branch
Information Technology Division

June 11, 1982

DTIC
ELECTE
JUN 25 1982
S D
B

DTIC FILE COPY

NAVAL RESEARCH LABORATORY
Washington, D.C.

Approved for public release; distribution unlimited.

82 06 24 039

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NRL Report 8583	2. GOVT ACCESSION NO. AD-A116 054	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A CONVERSATIONAL TEST FOR COMPARING VOICE SYSTEMS USING WORKING TWO-WAY COMMUNICATION LINKS		5. TYPE OF REPORT & PERIOD COVERED Interim report on a continuing NRL problem. 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) A. Schmidt-Nielsen and S. S. Everett		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 NRL Problem 75-0129-0-2 PE 61153N Proj. RR021-05-42
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE June 11, 1982 13. NUMBER OF PAGES 26
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 naturalness
Perceptual tests
Voice communicability test methods

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

A conversational test using live two-way communications provides a measure of the actual usability of voice systems, especially when voice quality is degraded. A conversational test developed at NRL was compared with two other communicability tests in a series of experiments using a variety of digital voice processors with data rates from 800 to 32,000 bps. All three tests ranked the voice processors very similarly, but they did not discriminate equally well among different processors. Other advantages and disadvantages of conversational test methods are discussed.

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)

CONTENTS

INTRODUCTION 1

BACKGROUND 2

RESEARCH WITH THE NRL TEST 4

- Test Facility 4
- Experiment I 5
- Experiment II 6
- Experiment III 11

SUMMARY AND CONCLUSIONS 16

ACKNOWLEDGMENTS 17

REFERENCES 17

APPENDIX — The NRL Communicability Test 18

DTIC COPY INSPECTED 2

Accession For
NTIS GRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A

A CONVERSATIONAL TEST FOR COMPARING VOICE SYSTEMS USING WORKING TWO-WAY COMMUNICATION LINKS

INTRODUCTION

The development of voice testing techniques has arisen from two different areas of investigation: the clinical need for diagnosis of hearing loss and the engineering need to determine the effectiveness of voice communication equipment such as vocoders, telephones, and radio links. Many of the test methods developed for one of these needs have been adapted to the others. Although some tests may be suitable both for clinical and for equipment evaluation, the requirements for the two uses of voice tests differ in may ways. The clinical test uses a speech signal known to be of high quality to evaluate the ability of individuals to discriminate speech sounds correctly. The equipment evaluation test presents a speech signal of unknown quality to several listeners (whose hearing abilities will not all be identical) and uses their responses to determine the signal quality.

Although some objective measures can be used to predict voice quality for certain highly specified types of degradation (e.g., signal-to-noise ratio for noisy environments), there is at present no general method for determining the quality or intelligibility of a speech signal except by asking the human listener, the intended receiver of the signal. People understand ordinary speech so easily and automatically that they often "hear" the meaning without noticing the exact words or consciously analyzing the phonemes that compose the words. They can also recognize a phoneme such as /b/ whether it occurs at the beginning or end of a word or is spoken by a man, woman, or child. A voice communication system must transmit to the listener the right acoustic information to preserve this understanding of speech sounds.

Voice testing of communication equipment and systems serves three primary purposes: selection, evaluation, and development. When there is a choice of several competing systems (e.g., different manufacturers), tests are needed to determine which one has the best quality and intelligibility. When any changes or "improvements" are made, tests are used to decide whether intelligibility is actually better or worse. Finally, tests of existing systems can be used to determine weaknesses and to decide how future systems might be improved.

The two most important characteristics of a test are reliability (the degree to which scores are repeatable) and validity (the extent to which it measures what it is supposed to measure). It is possible for a test to be highly reliable without being valid (i.e., the scores are meaningless). Validity cannot be achieved unless the test is also reliable. The validity of a test can be no greater than the square of the correlation of the test with itself. This square is the variance due to what is being measured, and the remaining variance is measurement error. It should be obvious that a test cannot have a large error if it is to measure anything well. In the context of voice testing there is to some extent a tradeoff between reliability and validity. To obtain a high reliability the test itself becomes quite artificial and unlike a natural communication situation. This is the case with intelligibility tests where many of the sources of undesirable variability in scores are carefully controlled, but the listener's task of discriminating between phonemes in rhyming words is totally unlike ordinary two-way communications. On the other hand, conversational tests offer a more realistic task but are necessarily more subject to variability from

Manuscript submitted January 21, 1982.

test to test. In addition, reliability must not be obtained at the expense of sensitivity. A test that produces repeatable scores but does not discriminate among systems is of no more use than a test that measures irrelevant characteristics of the system.

Other characteristics contributing to the usefulness and desirability of a test include diagnostic capability and ease and cost of testing. A test that provides specific information about strengths and weaknesses of a system is more useful than one that provides only a single score. A test that is cumbersome or time consuming to administer or one that requires extensive training or complex equipment and test facilities is unlikely to find widespread acceptance if other methods are available.

Most voice tests can be grouped into three major types, although a few may be said to bridge these categories:

- *Intelligibility or phoneme tests* assess the ability to hear or discriminate among individual speech sounds. The test materials are usually words or syllables, spoken either in isolation or in a carrier phrase, and the response can be either a written word or letter or a multiple choice selection. Word recognition tests are also included in this category. The score is based on the number of correct discriminations.

- *Quality or rating tests* are used to obtain opinion measures and assess acceptability rather than intelligibility per se. The test materials consist of one or more sentences which are rated by the listeners on various rating scales. Some tests use only a single scale and others use as many as twenty.

- *Conversational or communicability tests* assess the usability of a system using a two-way communication task. This allows the users to interact and to adapt to the requirements of the system (talk louder, talk slower, ask for repeats, etc.). On completion of the task the usability of the system is rated on one or more scales.

Scores on all three types of tests will to a large extent be correlated with one another in that many of the same characteristics will lead to high or low performance on all three test types. For example, very low intelligibility will invariably reduce acceptability and usability. However, these tests are by no means perfectly correlated, and to the extent that they measure different aspects of overall communicability, an adequate assessment should be based on the scores of several different types of tests. A particular processor might produce highly intelligible speech but sounds so harsh or grating that it may be highly annoying and therefore unacceptable even though it is perfectly understandable. A more pleasant sounding voice sample may also on occasion be less intelligible than an unpleasant one. There will also be situations in which only an interactive test can determine whether the deficiencies of a particular system can be overcome by learning compensatory behaviors.

Several excellent measures of intelligibility and quality are available, e.g., Modified Rhyme Test (MRT) [1], Diagnostic Rhyme Test (DRT) [2], Diagnostic Acceptability Measure (DAM) [3]. In situations where voice quality can be expected to be seriously degraded, as for example in low data rate and very low data rate digital voice communications, it becomes increasingly important to evaluate the communicability or actual usability of the voice system in addition to obtaining intelligibility scores.

BACKGROUND

The general format of a conversational test consists of a communication task requiring an exchange of information between the participants, followed by an evaluation of the ease or difficulty of using the voice system.

A variety of tasks could potentially be used as the basis for communication between the participants. The most obvious factor in task selection is the requirement of a two-way exchange of information. The exchange should be reasonably natural and interesting enough to keep the participants motivated. To control sources of variability due to factors other than system differences, the difficulty of the task itself should vary as little as possible from one test to the next. In this regard, it is also highly desirable for the task to be relatively insensitive to differences in intelligence. Since repeated testing of the same subjects is another useful way of reducing undesirable variance, it is very desirable if the task is something that can be used repeatedly with the same people, rather than, say, a puzzle whose answer is known once it is solved.

An important contribution in conversational test methods is the Free Conversation Test (FCT) [4] used extensively in Britain. Subjects are tested in pairs. Each subject is given one of a pair of photographs taken a short time apart. Their task is to discuss the photographs in order to determine which one came first. At the end of the conversation, they rate the amount of effort required to converse using a single scale with five levels of effort. Usually 12 systems are tested using 12 pairs of subjects, each pair conversing once over each of the systems. The order in which the systems are presented to each pair of subjects is determined by a Latin square design. The results are analyzed using analysis of variance.

The photograph comparison task is very good for motivating a two-way exchange of information, but once the pictures have been seen, they cannot be reused with the same people. This means that in order to do a large number of tests one must either have a large library of photographs or a regular source of new participants. A more serious problem is that even with very careful selection of photographs, some pairs will inevitably be easier to solve than others. The difficulty of task solution also seems to be highly dependent on the intelligence of the participants.

The Diagnostic Communicability Test (DCT) [5] was developed by Dynastat, Inc. and is based on a stock trading game. The set of stocks assigned to each person varies from game to game, so that the same task can be reused indefinitely. This makes it possible to train and maintain a test crew with relatively stable performance, which increases the comparability of tests conducted at different times. The rules for trading are highly structured, and once the game is learned, task difficulty does not vary from game to game and is relatively unaffected by differences in intelligence.

The test uses a crew of five trained participants who play the game for about 5 min after which each player rates the system on a questionnaire having 15 rating scales. The choice of the number of participants (5) and the rules for trading have been optimized to make maximum use of the communication channel and the participants' time in evaluating the system. The need for five participants does require conferencing capability in setting up the tests and limits the situations in which the test can be used. It is possible to conduct the test with fewer than five participants, but the stock game becomes uninteresting with only three people and insufferably boring with two. The use of multiple rating scales provides more information about the performance of the system than a single scale would. Communicability scales include such attributes as difficulty in hearing, understanding, and recognizing other talkers as well as background interference. Compensatory behaviors such as talking more carefully, louder, or slower are assessed; and personal reactions include effort, irritation, fatigue, and acceptability.

The NRL Communicability Test was designed to be used in a variety of situations ranging from informal equipment demonstrations to formal evaluation procedures. As in the FCT, participants are tested two at a time, but the communication task is a short version of the pencil-and-paper game "battleship." Since players place their own "ships" for each game, the test requires only a supply of test forms, and it can be used any number of times. A crew of trained subjects can be maintained, as with the DCT. The task is easily learned and can be used with either naive or practiced participants. There are four rating scales to be filled out after the game is completed. A detailed description of the test and recommended test procedures can be found in the appendix.

The NRL test combines the advantages of previous conversational tests and eliminates the major drawbacks of each test. The battleship game as a conversational task is reasonably interesting, and it can also be reused with the same subjects, whereas the picture comparison task for the FCT requires a new set of pictures for each new test. The picture task also varies in difficulty from test to test whereas the battleship game does not. The use of only two subjects at a time eliminates the need for conferencing and makes the NRL test more versatile than the DCT. The simpler test procedure and shorter questionnaire also eliminates the extensive training required for the DCT and makes the NRL test easier to use for informal assessments as well as more rigorous comparisons.

RESEARCH WITH THE NRL TEST

A series of experiments was conducted to compare the NRL test with other conversational tests. The conversational tests were all conducted using the NRL test facility.

Test Facility

The present test facility (Fig. 1) has a control station for the experimenter, and can accommodate up to five talker stations for the test participants. The five talker stations are isolated from one another by being located in separate rooms or in a sound booth. Each station has a telephone-type handset with a Roanwell Confidencer Model 240-100002-653 dynamic microphone. The handsets are wired for push-to-talk and are controlled by a system of relays simulating a half-duplex channel. Only one person at a time has use of the channel. The other stations hear the voice processed through the voice processor being tested while the talker hears only a normal unprocessed, undelayed sidetone. The half-duplex setup permits the use of a single processor in loop-around mode for both input and output since the signal only has to go in one direction at a time. The control station, operated by the experimenter, can override the talker stations at any time. The control station also has a switching system for changing from one processor to another, which can accomodate up to 12 different processors. Test sessions can be tape recorded for future analysis of the conversations.

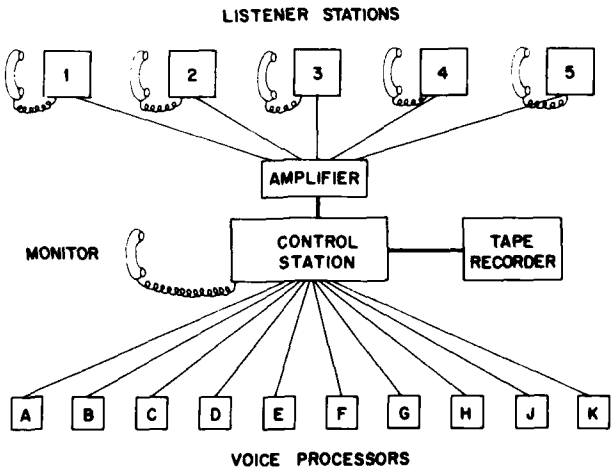


Fig. 1 — The NRL conversational test facility

Experiment I

The first experiment compared the NRL test and the DCT by using a small crew of subjects trained on both tests. The final version of the NRL test is slightly different from the one used in these experiments. The players originally placed one ship each in the battleship game; they now use two ships as a result of the outcome of this research. The questions on the first version were the same four that are shown in the appendix, and there were seven answer categories per question. Only the end points for each question were labeled, however, and the labels used were "Very low" and "Very high" for all questions.

Method

Four voice conditions or systems—three digital voice processors and a clear unprocessed voice channel—were tested. The voice processors were a 9600 bps residual excited linear predictive coder (RELP), a 2400 bps linear predictive coder (LPC), and a developmental 800 bps system. Five NRL employees, two females and three males, were trained on both tests and served as experimental subjects. On the DCT series, all five subjects participated in each test. The entire series consisted of four tests on each voice processor. The order in which the four systems were tested varied for each sequence according to a Latin square design. For the NRL test series, subjects were tested in pairs, each subject talking once over each system with each of the other subjects. This resulted in four tests per subject per system as in the DCT series. Again, the order in which the processors were tested for each subject pair was carefully counterbalanced over the entire test series.

Results

Each question on each of the two tests was analyzed separately by using a two-way analysis of variance with processors as a fixed effect, subjects as a random effect, and test repetitions as replications. The Newman-Keuls test [6] for comparisons among means was used for comparing the processors. In addition, ω^2 (the proportion of variance accounted for by an experimental effect [6, pp. 425-430]) was computed for the effects of processors, subjects, and the processor by subject interaction. For comparison purposes only the four questions common to both tests are discussed here.

Table 1 gives the results of the Newman-Keuls tests. The overall analysis of variance is not shown as it is the outcome of the comparison of mean scores for processors that is of interest for these tests. Except for the "Speak Carefully" question on the DCT, the voice systems were all significantly different from one another on all questions. In general, the NRL test provided somewhat greater separation among systems than the DCT. The voice systems that were tested all used very different data rates and were noticeably different in quality. The two conversational tests reflected the expected differences. This outcome illustrates that conversational tests do indeed reflect differences in the usability of different voice systems; the experiments that follow test their ability to distinguish among more closely competing voice processors. It is of interest to note that subjects were able to complete the communication task for both tests with the 800 bps system, even though the initial reaction to this system is that it sounds terrible.

Table 2 shows variance proportions. The proportion of the total variance accounted for by processor differences is very high for both tests. This may in part be a result of the fact that the processors were all so obviously different in quality. On all questions, the NRL test had a greater proportion of variance due to processors than the DCT.

The results of this experiment show that the NRL test is promising as a measure of the usability of voice communication systems. The following experiments give a more detailed comparison of conversational test methods.

Table 1 — Mean Scores on Each Question for Four Voice Systems Used in Experiment I. Brackets indicate groups within which no significant differences were found using the Newman-Keuls test ($p \leqslant .05$).

Question	NRL Test		DCT	
	Processor	Score	Processor	Score
Effort	Clear	93.5]	Clear	88.0]
	9600 bps	69.5]	9600 bps	70.8]
	2400 bps	35.8]	2400 bps	47.8]
	800 bps	15.5]	800 bps	17.8]
Unnatural Quality	Clear	92.0]	Clear	94.0]
	9600 bps	66.5]	9600 bps	71.2]
	2400 bps	38.0]	2400 bps	46.8]
	800 bps	13.2]	800 bps	17.0]
Speak Carefully	Clear	91.2]	Clear	59.5]
	9600 bps	71.0]	9600 bps	48.8]
	2400 bps	35.8]	2400 bps	39.2]
	800 bps	19.0]	800 bps	15.8]
Acceptability	Clear	93.5]	Clear	95.0]
	9600 bps	71.8]	9600 bps	79.2]
	2400 bps	40.2]	2400 bps	54.5]
	800 bps	21.5]	800 bps	28.2]

Table 2 — Variance Proportions (ω^2) for Each Test Question for Experiment I

Question	Variance Proportion		
	Processors	Subjects	Processors X Subjects
NRL Effort	0.73	0.05	0.03
NRL Unnatural Quality	0.77	0.02	0.00
NRL Speak Carefully	0.80	0.01	0.02
NRL Acceptability	0.71	0.07	0.00
DCT Effort	0.62	0.13	0.08
DCT Unnatural Quality	0.69	0.12	0.07
DCT Speak Carefully	0.58	0.07	006
DCT Acceptability	0.65	0.08	0.09

Experiment II

The next experiments were designed to compare the performance of the NRL test, the Free Conversation Test, and the Diagnostic Communicability Test. Ten voice systems were tested by use of

each of the three conversational tests. In Experiment II, all three tests were conducted with eight subjects and four trials per subject on each voice processor, a procedure that would be recommended for the DCT. Experiment III was conducted according to the procedure recommended for the FCT, using 20 subjects tested one time each on each processor. The voice processors tested covered a broad range of quality and data rates, and there were also groups of processors that could be expected to be very similar in performance. The sensitivity of each of the tests in discriminating among processors could be compared for each of the test procedures. At the same time, the amount of agreement in the results of the three tests served as a measure of the validity of the various conversational tests and procedures.

Table 3 — Voice Systems Tested in Experiments II and III

System	Type	Data Rate (bps)
A	Unprocessed voice	
B	CVSD[a]	32,000
C	CVSD	16,000
D	RELP[b]	9,600
E	APC[c]	9,600
F	LPC[d] version A	2,400
G	LPC, version B	2,400
H	LPC, version C	2,400
J	System H, Mode 2	2,400
K	Developmental	800

[a]Continuously Variable Slope Delta modulation
[b]Residual Excited Linear Predictive coder
[c]Adaptive Predictive coder
[d]Linear Predictive coder

Method

Table 3 lists the voice systems that were tested. Eight paid volunteers, recruited through an advertisement in the University of Maryland student newspaper, participated in the tests. Subjects were tested in two groups of four. Each subject was tested four times on each voice system on each of the three conversational tests. Subjects came in for 4 hours 3 days a week until testing was completed. Each 4-hour session consisted of several 30 to 45 min test sets with about 15 min of rest between test sets. Only one of the conversational tests was used in any given set, and the three tests were alternated within test sessions so that any effects of practice or fatigue were balanced across tests. Within each test type, the order in which the voice systems were presented was counterbalanced according to the procedures described in the appendix, with each 30 to 45 min test set being sufficient for five voice systems or half of one test order. Within each group of four, all four subjects participated in the DCT tests, while the FCT and NRL tests were conducted with pairs of subjects (partners were changed for different test orders).

Results

An analysis of variance was carried out for each question on each test, and the Newman-Keuls test was used for comparisons among means. One subject was unable to complete the fourth test order for the DCT. The average of her three other DCT scores for each system was used in lieu of a fourth score. Variance proportions were also computed as in Experiment I. In addition to the analyses of individual questions, the four questions on the NRL test were averaged for each answer sheet, and the same analyses were carried out. The four questions on the DCT that are comparable to those on the NRL test were also averaged and analyzed. These averages were used for comparing the three tests as a whole. Average DCT scores for all questions except "Voice Interference," "Pitch," and "Success of Compensatory Behaviors" were computed and analyzed as well. The results of this analysis were almost identical to those for four questions and accounted for somewhat less of the total variance, so the four question analysis was used for comparison purposes.

Table 4 shows the results of the Newman-Keuls procedure for the one FCT question, the individual NRL test questions, and the four comparable DCT questions.* In Table 4, the average scores on the NRL test and the DCT are compared with the FCT. From the average scores, it can be seen that on the whole the three tests agree very well. The voice systems were ranked very similarly on all three tests, and reversals occurred only within groupings where there were no significant differences. Although some of the differences among processors did not always reach statistical significance on all three tests, the overall statistical groupings were very similar, and all three tests differentiated quite well among the various voice systems. Other data (e.g., DRT scores, see Table 5) suggest that processors C, D, and E are indeed very similar in intelligibility and that processor B is only slightly better than these. Processors F, G, and H are different implementations of the same LPC algorithm, and would be expected to be quite similar. Processor J, with a data rate of 2400 bps, was supposed to simulate a digital vocoder but was clearly not functioning correctly and was next to unusable. It was significantly poorer than even the 800 bit processor K on the NRL test, but this ranking was reversed on the FCT. The reason for this is not clear, but it could be the result of a constrained vocabulary on the NRL battleship task and DCT compared to an unconstrained vocabulary on the FCT picture task.

Individual test questions gave very similar results to the average scores (see Table 6). Since all of the questions were highly correlated (as can be seen below) this is not surprising. In general NRL test and FCT questions seem to give better discrimination among systems than DCT questions. The remaining DCT questions were also analyzed, and all except the "Difficulty Understanding" question (which was almost identical to the "Unnatural Quality" question) gave poorer discrimination than the questions that are shown here.

Table 7 gives the proportion of experimental variance accounted for by the effect of processors, subjects, and the processor by subject interaction, both for individual questions and for average scores.

In general, one can expect a test to be a better measure of differences among voice processors if a larger proportion of the total variance is attributable to processor differences rather than to individual subject differences or to idiosyncratic subject by processor effects. However, the test design using repeated measures on the same subjects does permit individual subject effects to be partitioned out, and it is more important that the subject by processor interaction be small. Since the interaction term is the appropriate denominator for the F ratio for the processor effect in the Mixed Model, the significance tests will be negatively biased if there is a large subject by processor interaction. The analyses of variance showed no significant interaction effect for the FCT or the NRL test, and this is reflected in the

*For convenience in comparing numerical values, the FCT scores (originally 0-4) have been multiplied by 25 to make the scale comparable to the other tests, and those DCT questions for which high scores were assigned to the "bad" end of the scale have been reversed ($x' = 100 - x$) for comparability with other scales.

Table 4 — Average Scores for Eight Subjects and Four Trials for Three Conversational Tests in Experiment II. Brackets indicate groupings within which no statistically significant differences were found using the Newman-Keuls procedure ($p \leqslant .05$).

NRL		FCT		DCT	
Processor	Score	Processor	Score	Processor	Score
A	93.24	A	92.97	A	84.02
B	80.82	B	80.47	B	75.53
C	76.84	D	77.34	E	68.59
D	75.23	C	70.31	D	67.98
E	75.20	E	64.84	C	66.75
F	58.56	H	53.91	F	48.09
H	52.93	F	52.34	H	47.34
G	50.57	G	50.78	G	42.63
K	37.81	J	28.91	K	32.52
J	29.26	K	20.31	J	31.42

Table 5 — DRT Scores for Those Processors for Which Intelligibility Scores Were Available

Processor	DRT Score
A	97.2
B	92.9
C	89.1
D	89.8
E	91.4
F	87.2
K	72.2

Table 6 — Scores for Eight Subjects and Four Trials on Selected Questions for Three Conversation Tests in Experiment II. Brackets indicate groupings within which no statistically significant differences were found using the Newman-Keuls procedure ($p \leqslant .05$).

Question	Test					
	NRL		FCT		DCT	
	Processor	Score	Processor	Score	Processor	Score
Effort	A	94.06	A	92.07	A	94.53
	B	83.28	B	80.47	B	83.84
	C	77.66	D	77.34	E	75.78
	E	77.66	C	70.31	D	75.72
	D	77.19	E	64.84	C	71.41
	F	60.31	H	53.91	H	56.03
	H	55.63	F	52.34	F	55.69
	G	54.22	G	50.78	G	47.75
	K	40.16	J	28.91	J	39.28
	J	29.38	K	20.31	K	38.44
Unnatural Quality	A	92.66			A	95.38
	B	80.94			B	85.94
	C	78.13			E	74.31
	E	74.84			C	74.06
	D	72.50			D	72.75
	F	56.56			F	47.09
	H	47.66			H	42.34
	G	46.25			G	39.22
	K	36.41			J	30.00
	J	27.97			K	28.97
Need to Speak Carefully	A	92.19			A	52.09
	B	80.47			B	48.38
	C	78.13			D	48.38
	E	77.66			E	48.28
	D	76.72			C	46.47
	F	61.72			G	39.13
	H	57.97			H	39.06
	G	55.16			F	38.84
	K	42.03			K	30.94
	J	33.13			J	27.19
Acceptability	A	94.06			A	94.06
	B	79.06			B	83.97
	C	74.84			E	76.00
	D	74.53			D	75.09
	E	72.03			C	75.06
	F	56.56			H	51.94
	H	50.94			F	50.72
	G	49.06			G	44.44
	K	33.13			K	31.72
	J	26.09			J	29.22

Table 7 — Variance Proportions (ω^2) for Test Questions and Averages for Eight Subjects and Four Trials in Experiment II

Question	Variance Proportion		
	Processors	Subjects	Processors X Subjects
FCT Effort	0.56	0.13	0.03
NRL Effort	0.56	0.12	0.02
NRL Unnatural Quality	0.53	0.18	0.03
NRL Speak Carefully	0.42	0.23	0.01
NRL Acceptability	0.54	0.13	0.01
DCT Background Interference	0.16	0.35	0.10
DCT Voice Interference	0.08	0.55	0.11
DCT Unnatural Quality[a]	0.49	0.23	0.03
DCT Difficulty Hearing	0.21	0.41	0.09
DCT Difficulty Understanding	0.45	0.25	0.05
DCT Difficulty Recognizing	0.34	0.30	0.08
DCT Speak Slower/Faster	0.33	0.08	0.05
DCT Speak Louder/Softer	0.07	0.20	0.13
DCT Lower/Raise Pitch	0.00	0.19	0.06
DCT Speak Carefully[a]	0.48	0.08	0.02
DCT Success of Compensation	0.00	0.66	0.07
DCT Irritation	0.17	0.39	0.08
DCT Fatigue	0.15	0.44	0.08
DCT Effort[a]	0.36	0.34	0.06
DCT Acceptability[a]	0.51	0.19	0.01
NRL (Average of 4 questions)	0.56	0.16	0.02
DCT (Average of 4 questions)	0.52	0.23	0.03

[a]DCT questions equivalent to questions on NRL test

small variance proportions given for the interaction in Table 7. A number of DCT questions did have significant interaction effects and correspondingly higher variance proportions. On the whole, both the FCT and the NRL test had greater processor effects than the DCT and smaller interaction effects, but when DCT scores were averaged over the four equivalent NRL questions, the DCT compared more favorably with the other two tests. Some of the questions asked on the DCT may be difficult for the subjects to evaluate consistently. The question of correlations among questions is dealt with in a later section. For the NRL test the average score is probably the best overall evaluation.

Experiment III

Method

The test procedure was the same as in the preceding experiment, except that 12 additional subjects were tested once on each processor on each of the three tests. One subject was unable to complete testing, and one set of DCT scores is missing from the data. The first set of scores on each test from the eight subjects in the preceding experiment were also used giving a total of 20 subjects for the NRL test and FCT and 19 subjects for the DCT. The order in which the processors were presented was appropriately counterbalanced for each test.

Results

Analyses of variance, Newman-Keuls tests, and variance proportions were computed as in the preceding experiment. Tables 8 and 9 show the results of the Newman-Keuls test. Since each subject was tested only once on each processor, there is no separate interaction term in the analysis of variance as it is confounded with the error term. Table 10 shows the variance proportions attributable to processor and subject effects. The results for the FCT and the NRL test are very similar to the preceding experiment, with the FCT discriminating slightly better. The DCT, on the other hand, gives poorer discrimination among processors and has a smaller proportion of the total variance attributable to the effect of processors. Since the DCT had larger interaction effects in the preceding experiment, this may have had a deleterious effect with only one test per subject where error and interaction are confounded.

Table 8 — Average Scores for 20(19) Subjects and One Trial for Three Conversational Tests in Experiment III. Brackets indicate groupings within which no statistically significant differences were found using the Newman-Keuls procedure ($p \leqslant .05$).

NRL		FCT		DCT	
Processor	Score	Processor	Score	Processor	Score
A	93.31	A	97.50	A	79.54
D	74.19	B	81.25	B	72.50
C	74.00	C	75.00	D	70.33
B	71.94	D	73.75	C	67.63
E	68.56	E	68.75	E	65.86
F	53.38	H	55.00	F	45.66
H	46.25	F	50.00	H	42.43
G	42.69	G	45.00	J	41.65
K	36.88	J	31.25	G	37.50
J	28.63	K	20.00	K	33.22

The NRL test and the FCT performed somewhat better as conversational tests than the DCT in both experiments. The NRL test seems to be somewhat better with multiple measures on a smaller set of subjects, and the FCT seems to be better with single tests on a larger number of subjects, but these small differences may just be due to normal variation.

At the end of the testing, the subjects were given a questionnaire asking them to rate the three tests or to comment on them. Results are shown in Table 11 for those subjects who gave numerical ratings. Although they thought the FCT was the most interesting test to take, and the NRL test almost as interesting, they found the questions on the NRL test and the DCT better for evaluating the processors. All tests gave sufficient talking time for an adequate evaluation except on the rare occasions on the NRL test when a player's ship was sunk immediately. This has been corrected in the present version by assigning two ships to each player, thus lengthening the minimum time to complete a game. Because the grid is small, the average game time and maximum game time are not very much increased by this change.

Correlations, Pearson's r, were computed by matching individual subject scores for each question with every other question on all three tests. Table 12 shows all correlations, and Table 13 gives correlations for those questions that were analyzed in more detail above. The questions on the NRL test and

Table 9 — Scores for 20(19) Subjects and One Trial on Selected Questions for Three Conversational Tests in Experiment III. Brackets indicate groupings within which no statistically significant differences were found using the Newman-Keuls procedure ($p \leqslant .05$).

Question	Test					
	NRL		FCT		DCT	
	Processor	Score	Processor	Score	Processor	Score
Effort	A	94.25	A	97.50	A	87.10
	D	77.75	B	81.25	B	79.47
	C	77.00	C	75.00	D	77.89
	B	74.75	D	73.75	E	71.84
	E	71.00	E	68.75	C	71.32
	F	54.50	H	55.00	F	47.89
	H	51.50	F	50.00	J	46.84
	G	47.75	G	45.00	H	45.26
	K	40.25	J	31.25	G	38.16
	J	30.50	K	20.00	K	37.10
Unnatural Quality	A	93.50			A	91.84
	B	74.00			B	83.68
	C	73.25			E	77.63
	D	73.25			D	77.63
	E	65.75			C	75.53
	F	53.75			F	47.89
	H	43.25			J	43.68
	G	38.00			H	42.63
	K	37.25			G	39.74
	J	28.25			K	31.05
Need to Speak Carefully	A	93.50			A	49.74
	D	74.75			D	46.84
	C	73.25			C	45.26
	E	71.75			E	45.26
	B	70.25			B	44.74
	F	53.75			F	36.84
	H	50.00			H	35.53
	G	43.25			G	32.63
	K	38.75			K	32.37
	J	29.00			J	32.10
Acceptability	A	92.00			A	89.47
	C	72.50			B	82.11
	D	71.00			D	78.95
	B	68.75			C	78.42
	E	65.75			E	68.68
	F	51.50			F	50.00
	G	41.75			H	46.32
	H	40.25			J	43.95
	K	31.25			G	39.47
	J	26.75			K	32.37

Table 10 — Variance Proportions (ω^2) for Test Questions and Averages for 20(19) Subjects and One Trial (Experiment III)

Question	Variance Proportion	
	Processors	Subjects
FCT Effort	0.59	0.10
NRL Effort	0.49	0.18
NRL Unnatural Quality	0.50	0.15
NRL Speak Carefully	0.45	0.20
NRL Acceptability	0.50	0.17
DCT Background Interference	0.12	0.36
DCT Voice Interference	0.08	0.56
DCT Unnatural Quality[a]	0.47	0.18
DCT Difficulty Hearing	0.17	0.38
DCT Difficulty Understanding	0.47	0.21
DCT Difficulty Recognizing	0.32	0.29
DCT Speak Slower/Faster	0.23	0.18
DCT Speak Louder/Softer	0.09	0.21
DCT Lower/Raise Pitch	0.00	0.31
DCT Speak Carefully[a]	0.26	0.23
DCT Success of Compensation	0.02	0.59
DCT Irritation	0.23	0.34
DCT Fatigue	0.25	0.34
DCT Effort[a]	0.33	0.33
DCT Acceptability[a]	0.43	0.28
NRL (Average of 4 questions)	0.53	0.18
DCT (Average of 4 questions)[a]	0.46	0.27

[a]DCT questions equivalent to questions on NRL test

Table 11 — Subjects' Opinions of Test Adequacy (Rated from 1 to 10, with 1 Low) for Three Conversational Tests

Question	NRL (n=16)	FCT (n=16)	DCT (n=14)
Kind of Questions Asked	7.9	6.1	8.0
Number of Questions	8.3	4.4	6.9
Answer Categories (format)	8.6	6.2	7.8
Sufficient Talking Time	8.0	9.1	9.3
Game Interest	7.6	8.1	5.5

Table 12 — Correlations for Communicability Test Questions

		NRL 1	NRL 2	NRL 3	NRL 4	DCT 1	DCT 2	DCT 3	DCT 4	DCT 5	DCT 6	DCT 7	DCT 8	DCT 9	DCT 10	DCT 11	DCT 12	DCT 13	DCT 14	DCT 15
FCT	Effort	.674	.656	.614	.670	-.302	.315	.615	.327	.610	.564	.410	.239	-.015	.441	.191	.451	.468	.567	.590
NRL 1	Effort		.924	.897	.876	.347	.268	.654	.373	.610	.537	.381	.208	.056	.492	.123	.483	.438	.535	.635
2	Unnatural Quality			.866	.867	.382	.316	.681	.407	.637	.576	.361	.253	.029	.463	.106	.487	.447	.543	.632
3	Speak Carefully				.881	.297	.167	.605	.305	.556	.494	.344	.170	.057	.488	.034	.455	.426	.498	.583
4	Acceptability					.343	.251	.660	.357	.624	.561	.373	.169	.113	.453	.136	.448	.438	.545	.664
DCT 1	Background Interference						.575	.581	.443	.566	.498	.410	.387	.032	.374	.179	.501	.554	.528	.548
2	Voice Interference							.482	.725	.501	.523	.431	.428	-.108	.362	.142	.365	.378	.469	.512
3	Unnatural Quality								.585	.918	.821	.551	.383	.071	.661	.187	.738	.717	.835	.845
4	Difficulty Hearing									.616	.620	.544	.462	-.064	.465	.241	.504	.548	.605	.597
5	Difficulty Understanding										.807	.586	.401	.070	.663	.206	.759	.758	.886	.845
6	Difficulty Recognizing											.556	.348	.040	.625	.241	.673	.646	.798	.775
7	Speak Slower/Faster												.404	.160	.654	.108	.569	.564	.594	.599
8	Speak Louder/Softer													.139	.384	.003	.488	.509	.408	.418
9	Lower Raise Pitch														.102	-.027	.103	.112	.080	.019
10	Speak Carefully															.083	.728	.653	.724	.721
11	Success of Compensation																.123	.171	.247	.205
12	Irritation																	.941	.873	.792
13	Fatigue																		.854	.761
14	Effort																			.864
15	Acceptability																			

Table 13 — Correlations for Selected Question

	NRL 1	NRL 2	NRL 3	NRL 4	DCT 14	DCT 3	DCT 10	DCT 15
FCT effort	.67	.66	.61	.67	.57	.62	.44	.59
NRL 1 effort		.94	.90	.88	.54	.65	.49	.64
NRL 2 unnatural			.87	.87	.54	.62	.46	.63
NRL 3 speak carefully				.88	.50	.60	.49	.58
NRL 4 acceptability					.54	.66	.45	.66
DCT 14 effort						.84	.72	.86
DCT 3 unnatural							.66	.84
DCT 10 speak carefully								.72
DCT 15 acceptability								

the corresponding DCT questions are highly correlated with the FCT. All of the NRL questions are very highly correlated with one another, suggesting that they measure essentially the same thing. The high correlations among questions on the same test are also influenced by the tendency for subjects to mark every question high when they like a system and low when they do not. On the DCT, many of the correlations are not as high. The questions that show low correlations with the other two tests tend to be the same ones that have poor discrimination among systems and small variance components due to processors. This suggests that while these questions may measure something not measured by the other questions, whatever they measure is also not very important for distinguishing among the voice processors.

SUMMARY AND CONCLUSIONS

Three conversational tests for the evaluation of voice communication systems were compared. The rank ordering of the voice processors that were tested and the general statistical grouping of the processors was quite similar for all three tests. The overall results were reasonable when related to the data rates of the various voice systems and agreed well with what might be expected given the DRT intelligibility scores for the same processors. These results indicate that conversational tests can be a useful method for comparing voice systems.

On the whole, the NRL test and the FCT provided somewhat better discrimination among systems than did the DCT. The DCT, with its requirement for conferencing capability to accommodate the four or five subjects needed for each test may also be more difficult to administer readily. The major advantage of the DCT is that with more questions on the rating form, it provides more detailed information about the voice systems. However, the questions that provide the best discrimination among systems are highly correlated with one another and probably measure much the same thing. The questions that have a lower correlation with the "best" questions may provide information about different aspects of the voice systems, but they provide less discrimination among systems. The questions on the NRL test are all highly correlated with one another and probably do not provide much more information than the single question on the FCT. However, the subjects felt that they could give a more complete rating with the NRL test than with the single FCT question. Averaging the four NRL test questions gives excellent discrimination among systems, and scores on the individual questions provide at least some additional information about the systems. Subjects tended to prefer the answer format of the NRL test with seven categories from which to choose. They found that the five categories on the FCT gave them too few choices, and the continuous scale on the DCT gave them too many.

The NRL test and the DCT can both be used as many times as desired with the same subjects, which is a distinct advantage if the potential pool of subjects is small. The FCT requires constant renewal of either the picture pairs or the subject pool. In laboratories where frequent testing is

required, it might be advantageous to hire a semipermanent crew of subjects, and for this the NRL test or the DCT would be more useful. The FCT also represents a more unconstrained conversational environment in that the vocabulary s more extensive and the information exchange is considerably less structured than the other two tests. Communication requirements in real-world situations also vary from highly structured, limited vocabulary contexts to unstructured, unlimited vocabulary contexts. It is interesting that both the structured and unstructured tests gave very similar results. This suggests that a single test can be used to evaluate the voice systems, and that the requirements of the communication context should be used to set standards for the type of system to be selected, i.e., a constrained context with a very small vocabulary can tolerate more degradation than an unconstrained context.

A two-way communicability test is a measure of voice system usability. The NRL Communicability Test is versatile and requires little training. It provides at least as good discrimination among voice systems as other conversational tests and is easier to administer.

ACKNOWLEDGMENTS

This work was supported by ONR 6.1 research funds (RR021-05-42). The authors thank Thomas Tremain of NSA and George Kang and Larry Fransen of NRL for making many of the voice processors available, Howard Murphy of NRL for technical support in developing and maintaining the NRL test facility, and W. D. Voiers of Dynastat, Inc. for developmental research using the DCT and for many fruitful discussions on communicability testing.

REFERENCES

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

2. W.D. Voiers, "Diagnostic Evaluation of Speech Intelligibility," in *Speech Intelligibility and Recognition*, M.E. Hawley, ed., Stroudsburg, Pa: Dowden, Hutchinson and Ross, 1977.

3. W.D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems," in *Conf. Rec., 1977 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, May 1977, pp. 204-207.

4. L.W. Butler 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, Christchurch, Hants., Report No. 69014, Feb. 1969.

5. W.D. Voiers and M.H. Clark, "Exploratory Research on the Feasibility of a Practical and Realistic Test of Speech Communicability," Dept. of the Navy, Navy Electronic Systems Command, Wash., D.C., Final Report on Contract No. N0039-77-C-0111, April 1978.

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

Appendix

THE NRL COMMUNICABILITY TEST

The NRL Communicability Test (Fig. A1) was designed as a two-way conversational test for evaluating the usability of voice communication systems. The test uses a short version of the pencil-and-paper game battleship as the communication task. In this game, players place "ships" on a grid and then attempt to sink one another's ships by taking turns "shooting" at specified squares on the grid. Figure A2 gives the playing instructions. After completing the communication task, the participants fill out the question at the bottom of the form as their evaluation of the voice system over which they were talking.

The test is relatively short—usually about 5 min—and can be used for demonstrations or informal evaluations as well as for more rigorous comparisons of voice processing systems. The controlled test procedures described below are recommended when the test is to be used as an evaluation tool for comparing voice communication systems.

Subjects—The subjects who participate in the tests should be reasonably naive about the voice systems to be tested. Clearly they should not be people involved in the design or development of these systems. Whether or not subjects who have been in previous tests should be tested repeatedly is a more difficult question. It has been our experience that over time subjects become more tolerant of the poorer systems and tend to give them higher ratings than they did initially. On the other hand, it can also be argued that a trained and experienced group of subjects may be more consistent in their responses. The optimum number of tests for good resolution without excessive testing is either six subjects tested four times each on every voice system or eight subjects tested three time each. Alternatively, if eight subjects are tested four times each, this makes it possible to eliminate up to two subjects if they are clearly performing erratically. If only a few subjects are available, four subjects tested six times can also be used.

Training and Reference Systems—To familiarize subjects with the test procedure and with the type of voice systems to be expected, six training trials are recommended before the start of testing. The voice systems used for training should preferably span a broad quality range. We have found that using one ideal system, one very low-quality system, and one of moderate quality with two training tests on each of these, gives subjects a good reference frame for the subsequent test series. These same three reference systems are also included in every test series and help to provide some comparability with tests conducted at different times. A laboratory that conducts tests regularly could use a larger set of reference systems and standardize scores based on the scores given the reference systems.

Test Design and Procedure—Subjects are tested in pairs. If possible, subjects should change talking partners for each new test set. Each pair is tested once on each of the voice systems. The order in which the voice processors are presented should be different for each test set. Table A1 gives one possible set of assignments of talking partners for different numbers of participants. Table A2 gives orders for testing voice systems that are as close to balanced as is possible for the number of tests and number of voice systems to be tested (i.e., the series for 12 systems constitutes a Latin square design, and two Latin squares are used for the six-system series).

Subjects are isolated from one another, either in sound booths or in separate quiet rooms. Each testing session should last no more than 30 to 45 min, and subjects should be permitted about 15 min

NRL COMMUNICABILITY TEST

TALKER _____ DATE _____

TALKING WITH _____ TEST # _____

	1	2	3	4	5
A					
B					
C					
D					
E					

Opponent's shots at you

A = Alfa
B = Bravo
C = Charlie
D = Delta
E = Echo

	1	2	3	4	5
A					
B					
C					
D					
E					

Your shots at opponent

* * * * * * * * * * * * *

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

1. EFFORT required to communicate

No special effort — Moderate effort — Extreme effort: normal conversation impossible

|____| |____| |____| |____| |____| |____| |____|

2. UNNATURAL voice quality

Completely natural — Moderately distorted — Extremely unnatural

|____| |____| |____| |____| |____| |____| |____|

3. Need to SPEAK CAREFULLY

Can talk normally and casually — Talk more carefully — Extreme care in talking and pronouncing

|____| |____| |____| |____| |____| |____| |____|

4. Overall ACCEPTABILITY of the system

Excellent — Moderately acceptable — Unacceptable

|____| |____| |____| |____| |____| |____| |____|

Fig. A1 — Test form and answer sheet

NRL COMMUNICABILITY TEST

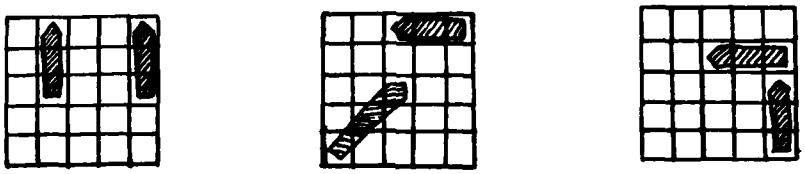
A short communicability test for evaluating voice systems.

* * * * * * * * * * * * *

Rules of Play:

The game is played on the two grids printed on the test sheet. You and your opponent both have two ships, each of which takes up 3 squares vertically, horizontally or diagonally.

Examples of ship placement:



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 3 of its squares have been hit.

Turns alternate, each turn consisting of one "shot". To shoot, you specify a cell in the grid (Alfa-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.

Place your ships in the left-hand grid, and tell your opponent when you are ready to begin.

After the game, please answer the questions at the bottom of the page.

Fig. A2 — Instructions for the communication task

Table A1 — Suggested Combinations of Participants

Test Set	Number of Participants		
	4	6	8
1	AB	AB	AB
2	AC	CD	AC
3	CD	EF	CD
4	BD	AE	BD
5	BC	BC	BC
6	AD	DF	AD
7	AB	AF	EF
8	AC	BD	EG
9	CD	CE	GH
10	BD	AC	FH
11	BC	BF	FG
12	AD	DE	EH

rest between sessions. This means that about five to eight systems can be tested in a single session, and if there are more systems than this in a test set, each set should be broken into two sessions.

Scoring and Data Analysis—Numerical values are assigned to each subject's responses using the chart in Table A3. Although each question can be analyzed separately, it is usually more convenient to use the average score for all four questions on each response sheet for the statistical analysis.

Since each subject is tested more than once with each voice system, a two-way analysis of variance with the repeated tests serving as replications can be carried out. The appropriate model is the Mixed Model with voice systems as a fixed effect and subjects as a random effect. Table A4 shows the analysis. The differences between individual systems are tested using multiple comparison techniques, for example, the Newman-Keuls procedure.[41]

$$\text{critical difference} = q\alpha\sqrt{\frac{\text{MS error}}{n}}.$$

In this case the MS error for processors is MS interaction, the appropriate "error" term for the F-test; n is the number of scores going into each treatment mean (i.e., number of subjects times number of tests per subject); and $q\alpha$ is the value of the Studentized Range statistic for the desired significance level, α, at the degrees of freedom for the denominator in the relevant F-ratio, and the distance between ordered means (the system means ordered from highest to lowest).

Comparisons among voice systems should generally be confined to systems tested in the same test series. One cannot expect exactly the same scores in a new test series with a different set of subjects, even though rankings and relative scores ought to be very similar.

[41] B. J. Winer, *Statistical Principles in Experimental Design*, New York: McGraw-Hill Book Company, 1971.

Table A2 — Suggested Order of Presentation of Systems to Be Tested

Test Set	4 Systems	5 Systems	6 Systems	7 Systems
1.	3 1 2 4	2 1 5 3 4	1 4 3 6 5 2	7 6 2 1 5 4 3
2.	1 4 3 2	5 3 4 2 1	5 1 6 4 2 3	4 7 6 3 2 1 5
3.	2 3 4 1	4 2 1 5 3	2 5 4 1 3 6	1 4 7 5 6 3 2
4.	4 2 1 3	1 5 3 4 2	3 2 1 5 6 4	3 1 4 2 7 5 6
5.	4 2 3 1	3 4 2 1 5	6 3 5 2 4 1	5 3 1 6 4 2 7
6.	2 1 4 3	1 3 5 2 4	4 6 2 3 1 5	2 5 3 7 1 6 4
7.	3 4 1 2	5 4 2 3 1	2 4 6 5 3 1	6 2 5 4 3 7 1
8.	1 3 2 4	2 1 3 4 5	5 1 3 4 2 6	3 4 5 1 2 6 7
9.	2 4 1 3	3 5 4 1 2	4 6 2 1 5 3	5 1 2 3 6 7 4
10.	4 3 2 1	4 2 1 5 3	1 3 5 6 4 2	2 3 6 5 7 4 1
11.	1 2 3 4	4 3 5 1 2	6 2 4 3 1 5	6 5 7 2 4 1 3
12.	3 1 4 2	3 1 4 2 5	3 5 1 2 6 4	7 2 4 6 1 3 5

Test Set	8 Systems	9 Systems	10 Systems
1.	1 2 5 8 4 3 6 7	1 6 2 7 4 8 5 9 3	4 7 10 5 1 2 9 6 3 8
2.	6 3 7 1 8 5 2 4	4 3 6 9 5 1 7 2 8	5 6 3 7 8 4 1 10 9 2
3.	2 5 4 6 1 7 3 8	5 8 3 2 7 4 9 6 1	7 10 9 6 2 5 8 3 1 4
4.	3 7 8 2 6 4 5 1	7 1 8 6 9 5 2 3 4	6 3 1 10 4 7 2 9 8 5
5.	5 4 1 3 2 8 7 6	9 4 1 3 2 7 6 8 5	10 9 8 3 5 6 4 1 2 7
6.	7 8 6 5 3 1 4 2	2 5 4 8 6 9 3 1 7	3 1 2 9 7 10 5 8 4 6
7.	4 1 2 7 5 6 8 3	6 7 5 1 3 2 8 4 9	9 8 4 1 6 3 7 2 5 10
8.	8 6 3 4 7 2 1 5	3 9 7 4 8 6 1 5 2	1 2 5 8 10 9 6 4 7 3
9.	7 6 3 4 8 5 2 1	8 2 9 5 1 3 4 7 6	8 4 7 2 3 1 10 5 6 9
10.	4 2 5 8 1 7 3 6	3 9 5 8 4 7 2 6 1	2 5 6 4 9 8 3 7 10 1
11.	8 3 7 1 6 4 5 2	8 2 7 1 5 9 6 3 4	8 3 6 9 2 1 5 10 7 4
12.	1 5 4 6 2 8 7 3	1 6 9 4 7 2 3 8 5	2 9 10 1 4 8 7 3 6 5

Test Set	11 Systems	12 Systems
1.	2 6 11 9 3 5 1 8 7 4 10	10 12 8 11 9 6 7 3 5 4 2 1
2.	5 7 10 8 11 9 3 6 4 1 2	4 1 9 3 10 2 5 12 11 6 7 8
3.	9 4 2 6 10 8 11 7 1 3 5	6 8 10 12 4 7 11 1 3 2 5 9
4.	8 1 5 7 2 6 10 4 3 11 9	2 9 4 1 6 5 3 8 12 7 11 10
5.	6 3 9 4 5 7 2 1 11 10 8	7 10 6 8 2 11 12 9 1 5 3 4
6.	7 11 8 1 9 4 5 3 10 2 6	5 4 2 9 7 3 1 10 8 11 12 6
7.	4 10 6 3 8 1 9 11 2 5 7	11 6 7 10 5 12 8 4 9 3 1 2
8.	1 2 7 11 6 3 8 10 5 9 4	3 2 5 4 11 1 9 6 10 12 8 7
9.	3 5 4 10 7 11 6 2 9 8 1	12 7 11 6 3 8 10 2 4 1 9 5
10.	11 9 1 2 4 10 7 5 8 6 3	1 5 3 2 12 9 4 7 6 8 10 11
11.	10 8 3 5 1 7 4 9 6 7 11	8 11 12 7 1 10 6 5 2 9 4 3
12.	6 2 10 3 5 4 9 1 8 11 7	9 3 1 5 8 4 2 11 7 10 6 12

Table A3 — Numerical Values for Scoring Each Response Category

1. EFFORT required to communicate

No special effort			Moderate effort			Extreme effort: normal conversation impossible
95	80	65	50	35	20	5

2. UNNATURAL voice quality

Completely natural			Moderately distorted			Extremely unnatural
95	80	65	50	35	20	5

3. Need to SPEAK CAREFULLY

Can talk normally and casually			Talk more carefully			Extreme care in talking and pronouncing
95	80	65	50	35	20	5

4. Overall ACCEPTABILITY of the system

Excellent			Moderately acceptable			Unacceptable
95	80	65	50	35	20	5

Table A4 — Two-Way Analysis of Variance with One Fixed Effect and One Random Effect

Source	Sum of Squares (SS)	df	Mean Square (MS)	E (MS)	F
Systems	$\sum_{j=1}^{a}(\sum_{k=1}^{b}\sum_{i=1}^{n} Y_{ijk})^2/bn - (\sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{n} Y_{ijk})^2/N$	a-1	$SS_{systems}/(a-1)$	$\sigma_e^2 + n\sigma_{\alpha\beta}^2 + bn\sigma_\alpha^2$	$MS_{systems}/MS_{interaction}$
Subjects	$\sum_{k=1}^{b}(\sum_{j=1}^{a}\sum_{i=1}^{n} Y_{ijk})^2/an - (\sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{n} Y_{ijk})^2/N$	b-1	$SS_{subjects}/(b-1)$	$\sigma_e^2 + an\sigma_\beta^2$	$(MS_{subjects}/MS_{error})$
Interaction	$SS_{total} - SS_{systems} - SS_{subjects} - SS_{error}$	(a-1)(b-1)	$SS_{interaction}/(a-1)(b-1)$	$\sigma_e^2 + n\sigma_{\alpha\beta}^2$	$(MS_{interaction}/MS_{error})$
Error	$\sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{n} Y_{ijk}^2 - \sum_{j=1}^{a}\sum_{k=1}^{b}(\sum_{i=1}^{n} Y_{ijk})^2/n$	$ab(n-1)$	$SS_{error}/ab(n-1)$	σ_e^2	
Total	$\sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{n} Y_{ijk}^2 - (\sum_{j=1}^{a}\sum_{k=1}^{b}\sum_{i=1}^{n} Y_{ijk})^2/N$	N-1			

a = number of systems
b = number of talkers
n = number of observations per talker-system combination

N = Total observations = a b n

DAT
FILM