SESSION V BEHAVIORAL ISSUES IN COMPUTER-GENERATED SPEECH

Mary Beth Rosson, *President*

Problems in evaluating the real-world usability of digital voice communication systems

ASTRID SCHMIDT-NIELSEN
Naval Research Laboratory, Washington, DC

Digital voice communication systems are being used in a wide variety of applications. Human observers are used to measure the performance of computer voice systems. We also need to determine how losses in speech quality and intelligibility affect the performance of the communicator. This paper considers several situational variables that can affect the usability and acceptance of digital voice communication systems: task constraints, user experience, and communication environment. User experience and communication environment influence the acceptability of degraded voice systems; greater experience and field environments lead to higher acceptability ratings. However, performance measures, such as task completion time, are still adversely affected by lower voice quality.

Digital voice communication systems are being used in a wide variety of applications. The systems vary from wideband (high-data-rate) systems, such as the 64-kbps (kilobits per second) PCM (pulse code modulation) used by the telephone system, to narrowband (low-data-rate) systems, such as the 2.4-kbps DoD (Department of Defense) standard LPC (linear predictive coding) algorithm. Wideband systems reproduce the speech signal more faithfully, but narrowband systems allow the same channel to carry a larger number of transmissions at the same time. Moreover, low data rates may be a necessity under conditions in which the channel will not support a higher data rate (e.g., some satellite communications). Reductions in the data rate are necessarily accompanied by some loss in speech quality and intelligibility, and specific applications may include other sources of degradation as well, such as high background noise or errors in the bit stream due to transmission channel conditions. These losses in voice quality have implications for the user that need to be evaluated in order to see how communication performance may be affected and to assess the relative merits of competing methods.

The measurement of the performance of a computer voice system, whether it be synthesis-by-rule or analysis-

synthesis, presents a particularly difficult problem because we must use human observers as measuring devices. The same attributes that are measured using human listeners have consequences for the performance of those using the voice system. This is different from evaluating something like chair size, with which one can measure the sizes of a group of people and the dimensions of the chair with a ruler. There is no mechanical measure of speech intelligibility. We have trouble getting machines to do more than the most limited recognition of human speech sounds, and with the present state of the art, they cannot determine which of two signals more accurately represents the intended message.

Tests of speech-output performance can be divided broadly into intelligibility tests, which provide objective right and wrong answers, and quality tests, which rely on the use of one or several opinion scales. The Diagnostic Rhyme Test (DRT) (Voiers, 1977) is an intelligibility test that has been widely used within the DoD because it has the advantage of providing highly reliable and repeatable scores that can be used to make comparisons even among systems that were tested at different times. Yet the potential users of the voice systems intensely dislike DRT scores because they have no reference frame by which to evaluate the numbers. They tend to prefer a "realistic" test even though such a test may be unrepeatable because the results will be confounded by a number

The author's mailing address is: Naval Research Laboratory, Code 7526, Washington, DC 20375.

of irrelevant variables, such as noise, distractions, and interruptions. The fact that human observers are used to obtain the speech scores has tended to obscure the fact that there are two separate issues to be addressed. One is the problem of accurately measuring the performance of the speech system, and the other is that of determining how specific losses in speech intelligibility or voice quality affect the performance of the communicator who tries to use the voice system in a real-world setting.

Although there is still room for improvement in test methods, it is also important to consider the consequences that reduced intelligibility and voice quality will have for communication accuracy and user acceptance under conditions that are likely to occur when the voice system is used.

This paper considers several situational variables that affect the usability and acceptance of digital voice communication systems. This research covers a variety of digital voice systems, some in current use and some under development. The focus is on the narrowband LPC algorithm at 2.4 kbps, because it is expected to be widely used. At 2.4 kbps, there is a noticeable loss in speech quality and intelligibility; the speech can be described as sounding "blurred" rather than clear. However, with practice, the system seems very usable even though it still takes some extra effort. The same effect seems to occur for speech synthesis by rule systems: They sound pretty awful at first, but with a little practice they become quite intelligible. The human listener is a wonderfully adaptable information processor.

The test procedure used in the experiments described here was the conversational test method. Conversational tests are especially useful for evaluating voice communication systems when voice quality is degraded, because such tests use an interactive two-way communication task, so that users can try to overcome the loss in voice quality. This also permits the distinction between compensable and noncompensable degradations. 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.

TASK CONSTRAINTS AND USER EXPERIENCE

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 of the United States, in 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. In one of our experiments, experienced and inexperienced users were compared using two communications tasks: one, highly structured and with a limited vocabulary, and the other, relatively unstructured and with an unconstrained vocabulary. The tests included a broad range of digital voice-processing systems. The purpose of the tests was both to determine how user experience and communication context influenced the rated acceptability of the voice systems, and how experience and task constraints affected the performance of the communicators talking over the voice channels.

Method

There were two conversational tests. The battleship game of the National Research Laboratory (NRL) test has a well-defined format with an effective vocabulary limited to only a few words, whereas the Free Conversation Test (Butler & Kiddle, 1969) has an open format and a potentially unlimited vocabulary. For the NRL Communicability Test (Schmidt-Nielsen & Everett, 1982), two people at a time complete a communication task based on the pencil-and-paper game of battleship (Figure 1). Each participant places two "ships" in a 5×5 grid, and players take turns "shooting" at each other by specifying cells in the grid (e.g., bravo-4). A ship is sunk when all three of its cells are hit; the game continues until one player wins by sinking both of the opponent's ships. This task provides the motivation for a natural exchange of information while also keeping vocabulary and task difficulty consistent from test to test. Four questions at the end of the form rate effort, voice (un)naturalness, need to speak carefully, and overall acceptability. Each person usually speaks several times over each voice system, and the order in which the voice systems are presented is balanced as far as possible to control for any possible effects of practice or fatigue. The task for the Free Conversation Test is a picture-comparison task (Figure 2). Each of the participants is given one of a pair of photographs taken a short time apart, and they discuss the pictures until they agree on which one came first. This test uses a single question with a 5-point scale for rating overall effort.

The inexperienced communicators were six students recruited through an advertisement in the University of Maryland newspaper. They were paid at an hourly rate for participating 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 communications 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.

The tests compared 10 voice links (Table 1), with a wide range of intelligibility being covered. The processors at 9.6 kbps and above were at least as good as an

NRL COMMUNICABILITY TEST

TALKER _____
 TALKING WITH _____

DATE _____
 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

Figure 1. Test form for the NRL Communicability Test.

Figure 2. Drawing of a typical pair of pictures for the Free Conversation Test.

ordinary telephone link, and the worst two processors were barely usable. Each subject was tested three times with each conversational test on each of the 10 voice systems using the equipment and test procedure described by Schmidt-Nielsen and Everett (1982). The order for the voice processors and conversational tests was counter-balanced for subject pairs.

Results

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

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. Table 2 summarizes the results. For convenience, the high-data-rate easy-to-use systems, the three LPC systems, and the two difficult systems have been grouped together and the results averaged. The experienced communicators spontaneously used phrase repetitions consistently, but the inexperienced group rarely did. Consequently the inexperienced group had more requests for repeats. The inexperienced communicators also had slightly more misunderstandings and errors per game than the experienced group. Even though most mistakes were eventually discovered and corrected, the smaller numbers of errors and misunderstandings by 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

Table 1
Voice Processing Systems Tested in the First Experiment

System	Data Rate	Abbreviation
Direct unprocessed channel		Unprocessed
Continuously variable slope delta modulation	32	CVSD 32
	16	CVSD 16
Residual excited linear predictive coding	9.6	RELPC 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

Table 2
Comparison of Talker Behaviors for the Constrained Task

	Experienced Communicators				Inexperienced Communicators			
	Unprocessed	Easy	LPC	Difficult	Unprocessed	Easy	LPC	Difficult
Spontaneously Repeated Phrases	5.67	9.67	12.00	25.56	0	0.03	0.04	0.33
Requests for Repeats	0	0.03	0.48	1.90	0.11	0.15	0.85	4.22
Errors and Misunderstandings	0	0.14	0.33	0.67	0.33	0.08	0.15	0.83
Time/Move (in seconds)	8.25	9.04	10.95	16.54	7.26	8.17	9.33	12.05

Note—The four easy systems, the three LPC systems, and the two most difficult systems have been grouped and averaged. The mean number of occurrences per game is shown for the voice systems in each group.

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 difficult 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.

As soon as the experienced communicators knew they were talking over a poor system, they put into use a number of standard communications conventions. They repeated phrases: “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.” 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 use of alternative vocabularies with more discriminable words is a well-known principle 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). Some of the intermediate strategies suggest that these users were not conscious of the general principles.

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 did. 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.

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. The talker behaviors are shown in Table 3. The changes were smaller than for the constrained task. Again, with the poorer systems the experienced group used more spontaneous repetitions, whereas the inexperienced group had more requests for repeats. The experienced group more often made constructive requests (“slow down” or “talk more quietly, you’re clipping” instead of “what?”). 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 counterbalanced 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

Table 3
Comparison of Talker Behaviors for the Unconstrained Task

	Experienced Communicators				Inexperienced Communicators			
	Unprocessed	Easy	LPC	Difficult	Unprocessed	Easy	LPC	Difficult
Spontaneously Repeated Phrases	0.22	0.22	0.74	2.39	0	0	0	0.11
Requests for Repeats	0	0.03	0.04	0.56	0	0.30	0.56	3.00
Total Game Time (in seconds)	203	214	239	255	87	108	115	152

Note—The four easy systems, the three LPC systems, and the two most difficult systems have been grouped and averaged. The mean number of occurrences per game is shown for the voice systems in each group.

the ratings. The subjects seemed to be able to rate the voice systems consistently and to divorce talking problems from solution difficulty.

It is clear that the quality of the voice systems affected the performance of the users on errors, requests for repeats, and task time, etc. 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. The objective behavioral measures were less stable as measures of processor performance than the sub-

Figure 3. Average ratings for the voice systems listed in Table 1. Brackets indicate groups within which differences are not statistically significant ($p < .05$, Newman-Keuls test).

jective ratings, but showed the trends that were consistent with the results of the ratings. This finding supports the validity of using ratings as a measure of processor performance.

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 Figure 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 ($p < .05$). 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 in which 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 finely graded distinctions among poor systems. The successful compensations made by this group may also have obscured processor differences.

TESTS IN AN OPERATIONAL ENVIRONMENT

Scores on laboratory tests may not be directly related to the usability of the communication system in a real-world environment. Frequently, users in real-world environments find a voice system to be quite acceptable even though it may have a relatively low intelligibility score. In order to evaluate the usability of a voice system based on the DoD standard LPC algorithm in an operational environment, communicability tests were conducted from a shipboard CIC room communicating with a shore station, and additional tests were conducted over radio links between two separate test rooms at the shore station. The principal comparison was of LPC with an existing narrowband voice system, and several other voice systems at other data rates were also included.

Method

Ship-to-shore tests. Tests using the NRL Communicability Test were conducted from the CIC room of the Navy destroyer U.S.S. *Peterson* to NAVELEX, Portsmouth, Virginia, and took place on the first 2 days of testing. A preliminary visit had been arranged to look over the CIC room and to discuss the test plans with the radio officer. The actual visit took place on another destroyer because the assigned ship was fogbound and had not returned on the day of the scheduled visit. The arrangement of the communications equipment was somewhat different on the two ships, and installation of the LPC system took longer

than anticipated. There was some difficulty in establishing and maintaining radio contact between the *Peterson* and Portsmouth. This was due to a variety of reasons, including the fact that there was apparently still some residual damage from a fire in the radio room several months earlier. During 2 days of testing, several links were set up and maintained long enough to complete at least one test set (i.e., three games, allowing all members of both teams to talk over and evaluate the system once). The order of testing was rotated between test sets so that each person talked to every member of the other team an equal number of times over the course of the entire series. The background noise level on the U.S.S. *Peterson* was measured at approximately 75 dB SPL on the "c" or flat scale. This included both talking and equipment noises. Noise levels would almost certainly be higher during normal activities at sea than they were in port, although we requested that normal activities continue as usual.

Shore tests. Some of the voice systems to be tested were not available aboard ship, and the remaining tests were conducted at Portsmouth during the last 2 days of testing. These tests were carried out by establishing HF and UHF radio links between the ship test room and shore test room at NAVELEX, Portsmouth. Because of time constraints, it was possible to complete only two test sets for each of the voice systems. This yielded a total of 12 evaluations per system, 2 for each of the six talkers. The order in which the voice systems were tested was reversed from the first to the second series of tests. Both rooms had a background of equipment and fan noises, as well as people walking in and out and talking. Noise measurements showed that the typical background noise level at the talkers' station in both rooms was around 70 dB SPL (on the "c" scale), and background talking ranged from 65 to 75 dB SPL.

Talkers. Since there were two separate test locations, two teams of three people each were used. These were composed of summer student employees and staff from the Naval Research Laboratory in Washington, DC, and from the Naval Electronics Systems Engineering Center in Portsmouth, Virginia. Each team received a short training series to familiarize them with the game rules and the rating forms.

Voice systems. The systems that were compared with LPC consisted of: a narrowband, 3-kHz, analog voice system used primarily on HF radio and wirelines; a narrowband digital channel vocoder used on satellite circuits, a wideband 16 kbps voice terminal using CVSD (continuously variable slope delta modulation) used primarily on UHF radio; and a wideband 19.2 kbps voice terminal using PCM also used on UHF radio. HF and UHF radio links without a voice-processing system were tested for comparison.

Results

The average scores for the tests in the two locations are shown in Figure 4. Scores on the ship-to-shore tests were generally lower than those for comparable systems

Figure 4. Ratings for the voice systems tested in field environments.

in the Portsmouth-to-Portsmouth tests. This result can probably be attributed to the difficulties in establishing and maintaining radio contact and to the poor quality of the radio link when it did work. The effect of such situation-specific variables is one of the things that makes it so difficult to interpret the results of field tests. There were also too few ratings collected from this location to draw any valid conclusions.

Communicability scores from the Portsmouth tests can be compared with scores for similar systems in laboratory tests conducted at NRL, shown in Figure 5. Note that the communicability scores depend on the communication context and on the quality of the other systems in the series. The LPC system was rated lower in the quiet laboratory environment than in the noisy field environment. Listeners were apparently more critical in the quiet laboratory, where they also had access to a high-quality unprocessed channel as a reference system. When talking in a noisy field environment, the users were more accepting of the voice systems even though DRT scores (Figure 6) suggest that the voice processors were actually less intelligible in background noise. DRT scores for LPC and CVSD with shipboard noise were considerably lower than scores for the same systems without noise. In marked contrast to DRT scores, communicability scores actually went up in the noisy environment. However, it is important to note that the rank ordering of the voice systems was consistent with the quality that might be expected on the basis of the data rate and technology of the individual processors—with higher data rates and newer systems giving better performance. User acceptance may vary with context, but a high-quality system is better than a poor one.

We know from experience with DRT tests (Sandy, 1984; Schmidt-Nielsen, 1985) that different background noises (e.g., jeep, aircraft, ship) affect both individual speakers and different types of voice processors in different ways. Some may perform better in one environment, and others may be better in another. It is important to take into account the conditions in the environment in which the system will be used.

The difficulties experienced in obtaining the shipboard data illustrate some of the pitfalls in conducting field tests. Similar difficulties were experienced by the U.S. Air Force in attempting to conduct real-time DRT intelligibility tests spoken from aircraft during actual flights and scored by listeners at a ground station (J. Tardelli, Arcon Corp., personal communication, October 18, 1984). In spite of elaborate preparations using a computerized list presentation and response system, very few usable data were collected. Problems included canceled flights, difficulties in establishing the HF communication link, and assorted software and equipment problems. The limited results that were obtained are difficult to interpret because there are no appropriate comparison conditions.

CONCLUSIONS

The results of these experiments suggest that people would be likely to be more accepting of reduced speech quality in a 'live' environment, complete with distractions, than they would be in a quiet laboratory. Communicators experienced in communicating over difficult channels are also more accepting of poor voice systems.

Figure 5. Ratings from laboratory tests for voice systems comparable to several of the systems in Figure 4.

Figure 6. Typical DRT scores for voice systems comparable to several of the systems in Figure 4.

This may be due partly to their ability to compensate for the degradations in speech quality. Communication task and user experience do not significantly affect the rank ordering of the judged acceptability of the voice systems, but they do affect the behavior of the communicators.

Although conducting tests in real-world conditions may provide a more realistic assessment of actual usability, there are also disadvantages. Among these are the following: Any one set of tests applies only to the specific application for which it was conducted and may have lit-

tle generality for other situations; the outcomes of the tests is likely to be affected by a variety of extraneous influences that can lead to questionable results (e.g., interruptions from people walking into the room, support equipment malfunctions, or changes in task loading for the operator); field tests are much more expensive and time consuming than laboratory tests. In some cases, it may be preferable to use simulated real-world conditions or to try to predict acceptability for a specific application from a combination of laboratory tests and an analysis of the application. In many cases, voice systems or algorithm improvements can be evaluated using standard intelligibility and quality tests, especially if they include multiple speakers and a range of background noises representative of the anticipated user environments. Limited supplementary communicability tests can be used to provide an indication of user acceptance in real-world environments.

REFERENCES

- BUTLER, L. W., & KIDDLE, L. (1969). *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* (Report No. 69014). Christchurch, England: Ministry of Technology, Signals Research and Development Establishment.
- SANDY, G. F. (1984). *Digital voice processor consortium final report* (Report No. MTR-84W00053-01, 02, 03). McLean, VA: The MITRE Corporation.
- SCHMIDT-NIELSEN, A., & EVERETT, S. S. (1982). A conversational test for comparing voice systems using working two-way communication links. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-30, 853-863.
- SCHMIDT-NIELSEN, A. (1985). *Tests of usability and intelligibility of LPC in different environments* (Report No. 8851). Washington, DC: Naval Research Laboratory.
- VOIERS, W. D. (1977). Diagnostic evaluation of speech intelligibility. In M. E. Hawley (Ed.), *Speech intelligibility and recognition*. Stroudsburg, PA: Dowden, Hutchinson, and Ross.
- WINER, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.