SPEECH PERFORMANCE TESTS FOR VOICE PROCESSORS

Z. Uzdy
The Aerospace Corporation
El Segundo, California

ABSTRACT

Voice processors, including those using linear predictive coding (LPC) algorithms, are becoming more prevalent in communication systems. Several intelligibility and speaker recognition tests were used to evaluate two currently available LPC voice processors. Both text-dependent and text-independent techniques were developed. The tests are aimed at potential users, yielding quantitative data that permit performance comparisons both of systems and of human speakers and listeners.

INTRODUCTION

Measurement of the speech performance of voice processors is a well-established activity that has produced numerous publications. The usefulness of these tests, however, has been limited to the degree that the test “instruments” have been human listeners whose judgments are subjective. The measured performance characteristics of speech processors have included the intelligibility, quality, and acceptability of the speech they produce and their ability to produce speech that is recognizable as originating with specific persons. Intelligibility was one of the earliest characteristics to be measured; one of the most widely accepted tests of intelligibility is the Diagnostic Rhyme Test (DRT). More recent emphasis has been on speaker recognition, particularly by automatic speaker recognition systems. Tests to produce quality and acceptability ratings have been carried out to supplement intelligibility test results, which have not always predicted user acceptance. Most of these performance tests are based on paired comparisons. In the DRT, the listener is asked to choose between pairs of words. In the speaker recognition tests, the choice is between pairs of speakers; in the quality tests, it is between paired voice systems.

A problem with most of these intelligibility and speaker recognition tests is that they assess performance in laboratory environments; they do not predict the performance to be expected in work settings. This paper describes simple tests that can provide a basis for the evaluation of voice processors by users. To test intelligibility, the DRT is supplemented by additional text-dependent tests. One group of tests is used to compare
voice processor performance and individual ability in a speaker-recognition task. Care is taken to minimize bias and to have only one variable per test. The tests described have several desirable features. They all yield quantitative data for comparison as opposed to purely subjective evaluation. Specially selected test subjects are not required; in fact, it is desirable that the test team include the potential users. The speaker recognition tests are carried out in real time using telephone-like handsets, simulating actual operating conditions.

This investigation was prompted by the need to compare the speech performance of two commercially available LPC voice processors, designated A and B for this paper. Both voice processors were capable of operating at a rate of 2400 bits per second (bps); one of them could also perform at 4800 bps (switch selectable). The potential users are small groups of engineering personnel in remote satellite tracking stations. However, the test methods described can be applied to other voice systems (e.g. channel vocoders), and to other user groups (doctors, military personnel, etc.), by appropriate choice of test materials.

TEST CONFIGURATION

The basic hardware setup for the tests is shown in Figure 1. Each test requires a pair of identical voice processors (labeled 1 and 2 in the figure). One voice processor analyzes the speech waveform and sends it in digital form to the second voice processor, which resynthesizes it into voice. Connecting the voice processors as shown in Figure 1 is referred to as back-to-back connection; this is commonly employed in laboratory tests. Since the data format is digital, the length of the transmission medium between the two units is not significant, and back-to-back testing is generally representative of long-link speech performance, even at high bit-error rates. The telephone handsets attached to the voice processors contain high-quality dynamic microphones. For the intelligibility tests, high-quality tape recorders were used to input the stimulus materials and record the output. The output tapes were later played back to the listeners via high-quality headphones.

INTELLIGIBILITY TESTS

Three tests of intelligibility were conducted--one using the isolated word pairs of the Diagnostic Rhyme Test, one using the DRT words embedded in meaningful sentences, and one using a technical paragraph with key words missing. The experimental scheme for all tests required the subjects to listen to tapes of utterances produced by the voice processors and then to utterances (spoken by the same speakers) recorded directly from high-quality microphones.
The Diagnostic Rhyme Test

The Diagnostic Rhyme Test (DRT) is a two-choice test of consonant discriminability widely used for testing voice systems. It yields a gross indicant of speech intelligibility; additional data can be extracted relating to specific aspects of the speaker, listener, or system under test.

The DRT utilizes a basic list of 192 word--96 rhyming pairs whose initial consonants differ in terms of a single phonemic attribute\(^{(2)}\). A word from each pair is presented twice during a testing session. In a given instance, the listener’s task is to indicate which member of the rhyming pair was spoken. A correct choice indicates that the listener has, in effect comprehended the speaker’s intent as to the state of one of six essentially binary perpetual attributes of English consonant phonemes. An incorrect choice indicates that the speaker, listener, or system under test has failed to distinguish the source state of the attribute. When, for example, the stimulus word is “zeal” and the choices open to the listener are “zeal” and “seal,” a correct response indicates that the listener has comprehended (and the system under test has transmitted) the speaker’s intent with regard to the attribute, voicing. Correspondingly, an incorrect response indicates that the system under test has failed to transmit a sufficient number of the acoustical features which distinguish the voiced consonants from their unvoiced cognates. Depending on the word pair involved, each item thus serves to test for the perception of one of the following elementary phonemic attributes:

- Voicing
- Nasality
- Sustention
- Sibilation
- Graveness
- Compactness

The performance of a voice processor with respect to phonemic attributes is of more interest to the equipment designer than to the user. During normal conversation, a speaker is unable to eliminate words that might confuse the voice processor. However, the DRT is a stringent test, and the manufacturer’s specified gross score does provide an initial baseline for comparing the performance of one voice processor with that of another. The gross DRT score can be calculated by deducting the number of incorrect responses from the number of correct responses (thus adjusting the score to correct for guessing):

\[
D = \frac{(R - W)}{T} \times 100 \quad \% \quad (1)
\]
where

\[ D = \text{Percent correct discriminations} \]
\[ R = \text{Number of correct responses} \]
\[ W = \text{Number of incorrect responses} \]
\[ T = \text{Total number of test word pairs} \]

The gross scores for all members of a test group are averaged to obtain the overall DRT rating.

The aim of our DRT tests was to evaluate the voice processors using the type of personnel likely to use them, rather than the specially selected subjects (those who score high on audiometric tests) who are normally used. The test team consisted of six engineers and technicians (all males) spanning a wide age group. Their first exposure to the DRT was to select, from three candidates, the most intelligible voice to be used as the stimulus voice for the tests. (This voice produced a 96\% average DRT score.) Two high-quality tape recorders were used to make the recordings. All tapes were evaluated in a normal room (without sound proofing), using high-quality headphones. When the tapes were played for the evaluators, they heard first the voice processor tape then the master tape. This meant that they evaluated the poorer-quality tape first and that no bias was introduced through memorizing. The evaluation was also limited to one tape a day so that the listeners would not become fatigued or bored.

**Text-Dependent Tests**

In order to determine whether intelligibility could be improved by placing the DRT words in a meaningful context (rather than isolated rhyming pairs), sentence tests were devised. Because the missing words were unique to the test, these tests were for one-time use and were therefore limited for execution on only one voice processor.

The sentences contained a subset of the DRT word list either as rhyming pairs--for example, “The tennis players met at the net”--or simply as words detached from the pairs. The printed sentences were shown to the listener with the underlined DRT words left blank. The tape recording of the sentence was then played for the listener, who was given about four seconds to write in each missing word. The number of missing words in the sentences varied between two and three. No word list of possible choices was provided, the listeners having to rely entirely on their own perception.

A paragraph test was also developed; it was similar in concept to the sentence test except that the material was in the technical field of the listeners. The test paragraph consisted of an introductory description of the basic operating principles of the voice processor, followed by simple instructions to unpack and operate the unit. Key words, comprising
about 20% of the total, were left blank. The emphasis was on key words essential to the meaning and not on word-for-word dictation. This test provided a measure of how familiarity with a given field affects intelligibility over and above casual material like that used in the previous sentence tests.

The sentence and paragraph material were presented to the listeners via tape, as for the DRT. The listeners had no previous exposure to these types of tests, so detailed instructions were provided. Both tests had about 30 missing words in their texts. The score (no correction for guessing was necessary here) was determined simply by

$$D = \left( \frac{R}{T} \right) \times 100 \quad (\%)$$

where

- $D =$ Percent correct discriminations
- $R =$ Number of correct responses
- $T =$ Total number of missing words

The results of the three intelligibility tests are shown in Figure 2. In each case, the voice processor’s performance is compared with that of the master tape. The manufacturer’s specified DRT rating of the voice processor is also included.

**SPEAKER RECOGNITION TESTS**

Most potential users of voice processors are accustomed to telephone quality voice. A data rate of 2400 bps, however, does not permit full telephone fidelity. The aim of the speaker recognition tests was to measure how effectively existing speaker recognition skills can be adapted to a voice-processor system. The tests were of two types: text dependent and text independent. The text-dependent test used the same fixed phrases repeatedly; not varying the test materials allowed us to compare the voice fidelity of the two processors without introducing semantic variables. The text-independent test used new material each time, but with the same voice processor; this test was more characteristic of continuous speech.

In a security-conscious environment, or in private conversation, the people involved usually know each other and are familiar with each others’ voices. Thus, when conducting speaker recognition tests with live subjects, it is reasonable to use a team of coworkers. To start a test, one team member would act as the listener, and the remaining six would form the speaker pool. The listener would go into a room isolated from the speakers. One handset of the voice processor was available to the listener, the other to the speaker pool.
**Text-Dependent Test**

The material for the text-dependent test consisted of six conversational utterances of increasing length. The phrases were, in order of presentation:

1. Hello
2. Nice day
3. How are you
4. What is the time
5. The weather here is fine
6. Have you recognized my voice yet.

Presenting the phrases in this manner provided material graded by length. Often a speaker could be recognized from a distinctive “hello;” otherwise, additional vocal information (e.g., rhythm) supplied by the continuous flow of the longer phrases was available. To preclude bias, the six speakers were always presented in random order, and multiple appearances by the same speaker were possible.

The speaker began by reading the first phrase. The listener repeated each phrase back to the speaker before the next phrase was spoken. This feedback enabled the listener to control the pace of the speaker to suit his rate of comprehension. Further, since the speaker knew the identity of the listener in each test, he was learning to recognize the listener’s speech pattern at the same time. When the listener recognized the speaker, he wrote the speaker’s name beside the last phrase uttered. After the last speaker, the listener rejoined the speaker pool and a new listener was selected. This procedure was repeated until every team member assumed the role of listener, completing one test session. No more than two test sessions were conducted per day (morning and afternoon)—again, to minimize fatigue and boredom among the participants.

The latter part of the test series data is shown in Figure 3. The bars indicate the probability of speaker recognition by the number of the phrase spoken. The amount by which the bars fall short of 100% is due to mistaken identity or failure to recognize the speaker by the end of the last phrase. The percentages are obtained by expressing all the phrase 1 identifications, phrase 2 identifications, etc., scored by the test team as a fraction of the total phrases available per test. The percentages are then designated probabilities.

**Text-Independent Test**

The text-independent test used paragraph material as the stimulus. This test was a logical extension of the previous phrase test. At this point, the participants had considerable exposure to the voice processors using the same phrases, and the test scores showed signs
of saturation. Thus, this series of tests provided a measure of how successfully the participants could transfer their acquired skills to new, more difficult material simulating continuous speech.

Four test sessions were held. The test procedure was similar to that used in the phrase tests, except that a different paragraph was passed out at the beginning of each test session. The paragraphs were about 60 words long and contained a variety of excerpts from technical journals. To begin a test, the speaker started reading the paragraph. As soon as the listener recognized the speaker, he told him to stop reading. At this point, the speaker circled the word in the paragraph where he stopped, and the listener wrote down the name of the person he believed to be the speaker. The response time of the speakers to the stop command was sufficiently fast so that the scores were accurate to within 1 or 2 words.

The test data were processed using the matrix shown in Table 1. Each participant appears both as a speaker and as a listener. This results in a matrix with the diagonal blank (because a participant could not be both listener and speaker simultaneously). The participant scoring the lowest average in both speaker recognition and voice recognizability is a good choice within a group for handling voice processor tasks requiring speaker recognition.

The final results are shown in Figure 4. The bars represent the range of the mean scores for each participant for the test series. The shaded bars represent the speaker recognition score, the white bars the ease of voice recognizability.

CONCLUSIONS

A number of techniques for evaluating the speech performance of voice systems have been described. The results of the intelligibility tests (Figure 2) show a number of significant features. It is apparent that the test team scored less than the manufacturer’s specified voice processor DRT value. This was expected because of the less than ideal test subjects and conditions. The test with DRT words in sentences did not significantly improve the average team score on the voice processor. Evidently, for most of the test subjects, the presence of meaningful context was cancelled by the absence of a word list from which to choose. (However, one individual did score 100%.) The technical paragraph test with missing key words shows an entirely different and unexpected result: all the participants scored as high on the voice processor as on the master tape. Our limited testing indicates that familiarity with the terminology balances the negative effects of the imperfect voice processor audio fidelity. A possible explanation is that the choice of words that can logically fit a technical passage is more limited, and there are no intentionally confusing words. Further tests to verify these allegations could involve exposing nontechnical people
to technical materials as well as testing other technical groups, such as medical personnel with medical terminology.

In the speaker-recognition tests, the resolution capability of the phrase tests for ranking voice systems is apparent from the team scores in Figure 3. The superiority of the telephone is clearly indicated; speaker recognition was almost certain (94% probability) after the first two phrases, compared to three phrases for the best voice processor performance. The two groups of 2400 bps scores also show voice processor B significantly ahead by the end of the third phrase. The six consecutive tests for voice processor A, three at 2400 bps and three at 4800 bps, do not indicate any advantage in using the higher bit rate, at least not for male voices. The speaker-recognition test scores using continuous text (Figure 4) show considerable variation between individuals. Such a result is not unusual for human subjects and has been documented previously. Participant 2 obtained the lowest and most consistent score for the test series. The test has determined him to be the best candidate for voice processor tasks in a security-conscious environment. It is also apparent that a short verbal preamble between conversing parties is desirable before confidential information is discussed.

ACKNOWLEDGEMENTS

The author is grateful to the Aerospace Corporation technical staff members and associates who contributed their time and effort as test subjects for this study, and to Mr. H. Tomita for presenting this paper.

REFERENCES

Table 1. Typical Result Matrix for Paragraph Test

<table>
<thead>
<tr>
<th></th>
<th>Speaker 1</th>
<th>Speaker 2</th>
<th>Speaker 3</th>
<th>Speaker 4</th>
<th>Speaker 5</th>
<th>Speaker 6</th>
<th>Speaker 7</th>
<th>Average (Voice Recognizability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener 1</td>
<td>-</td>
<td>6</td>
<td>-</td>
<td>21</td>
<td>9</td>
<td>20,28*</td>
<td>21</td>
<td>11.0</td>
</tr>
<tr>
<td>Listener 2</td>
<td>4</td>
<td>-</td>
<td>6</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>13</td>
<td>9.4</td>
</tr>
<tr>
<td>Listener 3</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>21</td>
<td>9</td>
<td>20</td>
<td>20</td>
<td>9.0</td>
</tr>
<tr>
<td>Listener 4</td>
<td>12</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>12</td>
<td>12,12</td>
<td>13</td>
<td>12.0</td>
</tr>
<tr>
<td>Listener 5</td>
<td>7</td>
<td>6</td>
<td>9,8</td>
<td>5</td>
<td>Error</td>
<td>-</td>
<td>-</td>
<td>7.0</td>
</tr>
<tr>
<td>Listener 6</td>
<td>20</td>
<td>12</td>
<td>13</td>
<td>12</td>
<td>-</td>
<td>12</td>
<td>28</td>
<td>16.2</td>
</tr>
<tr>
<td>Listener 7**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Average (Voice Recognizability)</td>
<td>11.0</td>
<td>9.4</td>
<td>9.0</td>
<td>15.0</td>
<td>9.5</td>
<td>16.0</td>
<td>19.0</td>
<td></td>
</tr>
</tbody>
</table>

*Two values indicate two appearances for that speaker.

**Listener 7's scores on speaker recognition tests were biased by prior absence and were discarded as unrepresentative.
Figure 1. Voice Processor Test Configuration

Figure 2. Intelligibility Tests - Team Scores

Symbols:
- ▲ - Specified DRT Score
- M - Master Tape
- V - Voice Processor B
Figure 3. Speaker Recognition Team Scores - Phrase Tests

Figure 4. Voice/Speaker Recognition Individual Scores - Paragraph Test