Prediction and Statistical Evaluation of Speech Recognition Test Scores

Gerald A. Studebaker*
Ginger A. Gray*
William E. Branch*

Abstract

A speech test evaluation and presentation system is described. The test presentation sub-

system has the flexibility and speed of live-voice testing while using recorded test materials.
The speech test evaluation subsystem compares an individual subject's test performance

on a monosyllabic word test with that of an average person with the same hearing loss. The

elements needed to make such evaluations are discussed. Also, a trial of the procedure is
described. The primary purpose of the trial was to obtain data that would provide a basis for

statistical probability statements about individual monosyllabic word test results obtained in

clinical settings. Data were collected from three audiology clinics in three different types of

settings. Except for a few cases with highly asymmetric speech scores, all nonconductive

hearing losses were included. Subject ages ranged from 8 to 92 years. Importance-weighted

average pure-tone hearing losses ranged from 0.4 to 97.6 dB HL. Fifty-word recognition

scores and audiograms for 2609 ears were included in the main analysis. Twenty-five-word

recognition scores and audiograms for another 932 ears from one clinic were used in a sub-

sidary analysis. Results indicated that distributions of absolute speech recognition scores

in hearing-impaired samples are highly skewed. However, after transformation of the scores

into rationalized arcsine units (rau), the differences between individual subject scores and

scores predicted from the audiogram were reasonably well described by the normal distrib-

ution. The standard deviation of this distribution of differences, for the data combined across

the three audiology clinics, was approximately 13 rau.

Key Words: Articulation index, audibility index, frequency importance function, intensity

importance function, speech intelligibility index, speech intelligibility tests, speech recogni-
tion tests, statistical evaluation of speech intelligibility test scores rationalized arcsine units

Abbreviations: AI = articulation index, CD = compact disc, FIF = frequency importance

function, HLD = hearing loss desensitization, IIF = intensity importance function, IWHL =
importance-weighted hearing loss, MSHC = Memphis Speech and Hearing Center, MVAH =
Memphis Veterans Administration Hospital, NU-6 = Northwestern University Auditory Test
No. 6, PL = presentation level, rau = rationalized arcsine units, SII = speech intelligibility
index, SL = sensation level, SLD = speech level desensitization, STEPS = Speech Test
Evaluation and Presentation System, SRT = speech recognition threshold, STI = speech
transmission index, TF = transfer function, UMMS = University of Maryland Medical System,
WAI = weighted audibility index

Speech "discrimination" tests, or speech recognition tests, have been a routine part
of audiologic evaluations for nearly 50 years (Newby, 1958; Harris, 1965). In spite of this
long history and the large number of papers devoted to their development, assessment, and
utilization, these tests are still frequently applied and evaluated in ways that have changed little
since the 1950s (Newby, 1958; Martin et al, 1998). The test items are often presented via live voice
and/or with shortened test lists (Martin et al, 1998), whereas the assessment of results is usually
based on little more than the clinical impressions of the audiologist. At present, the most
commonly used speech recognition test is the CID W-22, developed in the 1950s (Hirsh et al, 1952),
although recent trends suggest that the Northwestern University Auditory Test No. 6 (NU-6)
(Tillman and Carhart, 1966), developed in the 1960s, will soon overtake it (Martin et al., 1998). Clearly, there is substantial inertia in this area. An analysis of the reasons for this may be instructive.

The widespread clinical use of live-voice presentations and/or testing with an abbreviated list has continued in spite of the plethora of evidence that these practices produce results with low reliability (Jerger, 1998). The roles of test speed and presentation rate adaptability in promoting this course of action seem obvious. That is, live-voice testing is both faster and more adaptable to the response rates of different subjects than recorded test materials, and shortened lists save valuable time.

But, there is another, less obvious factor that encourages these shortcuts. Many clinicians feel that the expenditure of the additional time to produce a more reliable speech recognition score is not worthwhile because, in any case, speech score values are interpretable only in a vague manner. That is, in the absence of better tools, clinicians can have only a general impression of what the score should be, given the hearing loss, or whether the obtained score differs enough from expectations to have diagnostic significance. Given the fuzzy nature of the interpretation aspect of the procedure, a clinician quite logically may feel that there is little justification for spending extra time obtaining a more reliable score.

If the above analysis is correct, then it is obvious what needs to be done in order to improve clinical practice. First, clinicians need a test presentation system that uses recorded materials and is as fast and adaptable as live-voice testing. Second, they need a system that permits abbreviated test lengths that are invoked automatically under particular circumstances defined in advance. With computers and compact disc(CD) players, meeting these first two requirements is relatively simple, and these elements are included in the speech test system that is described in this paper.

But there is a third thing needed. It is perhaps the most important and it is the most difficult to achieve. This is the need of the clinician for an easy-to-use, objective, and quantitative evaluation of the results they obtain. This aspect of the problem is the primary focus of this paper.

To determine what might be done to help clinicians obtain greater meaning from their speech recognition test results, we considered how an ideal experienced clinician would evaluate these results. First, we concluded, the clinician would decide what score was expected, considering the subject’s pure-tone hearing loss and the conditions of test presentation (test material, presentation level, etc.). Then, the clinician would note whether the subject scored higher or lower than expected and whether the difference was large enough to have clinical significance.

There are two principal reasons why practicing clinicians have difficulty implementing this scenario well. First, they do not have an adequate basis for knowing what score to expect, given the particular hearing loss and listening conditions. Second, clinicians do not have available an adequate basis for assessing the likelihood that the observed departure from expectations could have occurred by chance. Some clinicians, undoubtedly, are better than others at making such judgments, and skills can be developed through experience, but the number of free variables, and the sensitivity of the results to these variables, makes these assessments a daunting challenge for anyone.

Overall, our goals for this project were to develop and evaluate a procedure designed for clinical use that would statistically evaluate individual speech recognition test results in relation to expected results. This procedure would also provide the speed and adaptability of live-voice presentations yet offer the reliability of recorded speech presentations. To achieve these goals, an integrated system was developed consisting of an audiometer, computer, CD player, and supporting software. The software calculated an expected score based on the conditions of the test and the subject’s audiogram, compared the observed and expected values, and provided a graphic interface for test administration control and display of the results. Together, these components were referred to as STEPS (Speech Test Evaluation and Presentation System).

The first part of this paper describes the elements and rationale of the STEPS system while the second part reports on a data collection project using the system that was carried out in three different clinical settings. The purposes of this project were (1) to test the STEPS system under real clinical conditions and (2) to provide data on the statistical distributions of the differences between predicted and observed speech scores for a large number and broad range of audiology clinic clients.

**RATIONALE**

**Why Predict Speech Recognition Scores?**

A major goal of STEPS is to predict and evaluate absolute speech recognition scores, not
just articulation index (AI) values or relative speech scores. Prediction of absolute score values is a substantially more demanding goal than predictions of relative scores but one that, if achieved, offers very considerable advantages. Absolute score predictions enable descriptive and probabilistic statistical comparisons of an individual's speech test results either with the average of other persons under the same conditions of audibility or with the expected average score of persons with the same pure-tone hearing loss. Such comparisons may increase the diagnostic value of the tests and perhaps be useful in planning individual rehabilitation programs for hearing-impaired persons. These applications are not fundamentally different from what clinicians have done on a subjective basis for many years.

A second type of application for STEPS results is to use a subject's performance on a speech test—in relation to an expected score for the average person with the same hearing loss under like listening conditions—to make quantitative predictions about the likely performance of that individual under other listening conditions or with other types of speech. An example of this would be to predict the degree of listening success for an individual in real-world listening conditions based, in part, on his or her relative performance on a standardized clinical speech test.

A third kind of application for predictions of speech recognition scores is the rapid estimation of optimal amplification characteristics for a particular person for each of a range of listening conditions (Hou and Thornton, 1994). Although such estimations can be made based simply on an index value, predicted speech scores provide a better basis for assessing the practical importance of the factors being considered.

A fourth type of application is one more normally used in research settings. This application is to use speech score predictions to estimate the effect of audibility on performance so the investigator can focus more clearly on factors other than audibility that might affect speech recognition test performance across individuals.

What Is Needed?

Assuming, then, that predictions of absolute speech recognition scores have potential value, what is needed in order to make and use such predictions in the evaluation process? Following are some of the more important aspects.

Theoretical Structure

The first requirement is a theoretical structure that provides the basis for calculating a prediction. Such a structure may be provided by articulation theory (Beranek, 1947; French and Steinberg, 1947; Fletcher and Galt, 1950) and its progeny, including the AI (ANSI, 1969), the speech transmission index (STI) (Steeneken and Houtgast, 1980), the speech intelligibility index (SII) (ANSI, 1997), and the weighted audibility index (WAI) (Studebaker, 1991).

The theoretical structure used to predict speech recognition scores in this project is a modification and extension of the articulation theory of French and Steinberg (1947) called audibility theory (Studebaker, 1991). Its index is referred to as the WAI to distinguish it from the classical AI. The WAI has been described only briefly in published form (Studebaker, 1991) but is described in detail in several unpublished papers (Studebaker et al, 1991; Studebaker and Sherbecoe, 1994; Studebaker, 1992, 1994) available from the first author.

Modifications of the classical AI included in the WAI primarily consist of how the effects of masking spread are calculated and in the frequency importance-weighting function and speech peak values that are used (Studebaker et al, 1993). The extensions consist mainly of corrections for the effects of speech level and threshold elevation (see below).

More significant than these modifications and extensions, however, is incorporation of the WAI into a larger scheme that takes into consideration the effects of specific listening conditions (e.g., earphone type) and test materials (e.g., NU-6) on score. These modifications in both the form and the concept of the classical AI make possible comparisons on an absolute score basis between score predictions and actual performance.

Speech Test with Known Characteristics

The second thing needed to make absolute speech recognition score predictions possible is a speech recognition test with well-known characteristics under conditions of reduced audibility. Specifically, it is necessary to know the test's frequency importance function (FIF), its intensity importance function (IIF), and its transfer function (TF).

Frequency Importance Function. The FIF (also called the band importance function) is a weighting function that reflects the fact that
some frequency regions in the speech spectrum are more important than others. The literature suggests that the FIF varies with speech type (Pavlovic, 1994). However, the more important question is whether such variations have practical consequence.

An analysis by Studebaker and Sherbecoe (1993) suggested that under many ordinary listening conditions, with ordinary types of speech, the FIF used in the calculation has little effect on predictions. Nevertheless, significant effects are likely under those listening conditions in which audibility varies substantially with frequency, such as with hearing losses and/or hearing aids, or when sharply filtered conditions are used in the experimental laboratory. As these are representative of the conditions frequently encountered by audiologists, it is important that the method used to predict absolute speech scores should include a FIF specific to the speech materials used to test the subjects. The FIF used in this study was the one found appropriate for the Auditec recordings of the NU-6 by Studebaker et al., (1993).

**Intensity Importance Function.** The IIF consists of three components: dynamic range, speech peaks, and importance weighting. The AI, STI, speech intelligibility index (SII), and the WAI assume that the dynamic range of speech is 30 dB. French and Steinberg (1947) assumed it to be 36 dB and Fletcher and Galt (1950) assumed it to be 68 dB. Speech peaks values (i.e., the root-mean-square levels exceeded by the speech 1% of the time) are normally taken to be either 12 (Beranek, 1947; Kryter, 1962; ANSI, 1969) or 15 dB (Steeneken and Houtgast, 1980; ANSI, 1997). Most assume that every decibel within the dynamic range contributes equally to the index value and that decibels of signal-to-noise ratio outside that range contribute nothing. There are exceptions. French and Steinberg (1947) assumed that the upper 12 dB of the dynamic range counted less per decibel than the lower level parts, whereas Fletcher and Galt (1950) assumed a complex function that assigned the greatest importance to levels near the middle of the dynamic range.

Each of these issues is currently under investigation and alternative assumptions may be suggested in the future. However, in this study, STEPS was set to use 15-dB speech peaks, 30-dB dynamic range, and a rectangular IIF (i.e., one where all levels within the dynamic range have equal value). This was done on the grounds that these are the assumptions most widely used at the present time. Also, they are the same assumptions as used in the new SII (ANSI, 1997) and in the STI (Steeneken and Houtgast, 1980).

**Transfer Function.** The TF (the function that relates an index value to an actual score) is perhaps the single most important influence on absolute score predictions. It has long been widely recognized that this function is different for each type of speech (ANSI, 1969) or for each particular version (e.g., recording, talker) of the same type of speech (Studebaker and Sherbecoe, 1991; Studebaker et al., 1993). Thus, the requirement for a TF that is valid for the specific test material employed cannot be overemphasized. The TF used in this study was one found by Studebaker et al (1993) to be appropriate for the NU-6 test used in this study.

**Other Factors**

A third requirement for making absolute score predictions is knowledge of the effect of any factors, other than simple audibility loss, that may significantly influence a person’s recognition performance. Examples include things such as speech level, extent of hearing loss effects beyond audibility loss, and subject age.

**Speech Level Desensitization (SLD).** When speech audibility is held constant, higher than normal speech levels have a negative effect on speech recognition (French and Steinberg, 1947; Fletcher and Galt, 1950). The detrimental effects of higher than normal speech levels may be greater when speech is heard in the presence of a noise than in quiet (Studebaker et al., 1999). However, in the case of the present investigation, the speech test materials were always presented in quiet. Therefore, in this instance, a table of correction values based on the data of French and Steinberg (1947) was applied. This table produced correction values very similar to those of the “level distortion factor” included in the SII standard (ANSI, 1997) calculation.

**Hearing Loss Desensitization (HLD).** To better predict the speech recognition performance of hearing-impaired subjects, it is

---

For speech sound pressure levels above 65 dB, the WAI was reduced by 0.00425 per dB from 66 to 84 dB, 0.00605 per dB from 85 to 104 dB, and 0.0062 per dB above 105 dB.
necessary to include an estimate of HLD. The HLD is intended to correct speech score predictions for the fact that, given the same audibility, persons with more hearing loss will, on average, understand speech less well than persons with less hearing loss. Pavlovic et al (1986) labeled this the “desensitization” factor. Since then, to distinguish it from the SLD, just described, we call this factor hearing loss desensitization.

The HLD, like the SLD, was created as a multiplier of the WAI. It originally ranged from 1.0, in the case of thresholds of 15 dB HL or less, to zero for thresholds above 94 dB HL. In the absence of data, Pavlovic et al (1986) drew a straight line between those two points. Later, to obtain a better estimate of this function, we used results from the files of about 100 adult clients of the Memphis Speech and Hearing Center (MSHC) as follows. First, each observed speech score was converted into a WAI using equation 1:

\[ I_a = -Q \log (1 - S^{1/2}) \]  

where S is the score as a proportion, Q and N are fitting constants from the transfer function, and \( I_a \) is the apparent WAI, that is, the index value needed by the average subject to produce the observed score. Each apparent WAI was turned into a ratio (i.e., a proficiency factor) by dividing it by the WAI calculated from the physical descriptions of the presentation conditions and hearing loss for each case. These ratios were plotted as a function of the average pure-tone hearing loss at 500, 1000, and 2000 Hz and fit with a line using a variety of mathematical functions using TableCurve 2D (Jandel Scientific, 1994). A very good fit was obtained with equation 2 with X set at 104.13793:

\[ HLD = 1 - (PTA/X)^3 \]  

This equation and constant were used to calculate the HLD function for this project.

**Age.** Another factor to consider in a speech recognition prediction scheme is the age of the subject. Be aware that the question posed here is not whether older subjects have more hearing loss, and thereby poorer speech recognition scores, than younger subjects. Rather, it is whether older subjects do more poorly on speech tests when speech audibility and the extent of hearing loss are the same as for younger subjects.

Whether to use an age correction in a prediction scheme involves a philosophical question of the “Who do you compare to?” type. This issue is discussed in more general terms in the next section. In the case of age, the specific question is whether it is more meaningful to compare a subject’s performance to that of an average subject of the same age and/or hearing loss or to that of younger subjects with the same hearing loss or normal hearing (see below).

Studebaker et al (1997) studied the effect of age on monosyllabic word test scores in an investigation in which all of the subjects had normal hearing within the frequency passband of the speech test material. The present study offered an opportunity to estimate the effects of age independent of audibility differences in a larger and more diverse subject sample. Therefore, we did not include an age correction in our predictions for this project. Results concerning this issue are presented later.

### Statistical Issues

**Who To Compare To?** To compare obtained and predicted scores meaningfully, it is necessary to evaluate the implications of each plausible decision about who the comparison group should be. The decision depends primarily on the purpose of the comparison being made. For example, an obtained score could be compared to either an average normal-hearing subject or an average subject with the same pure-tone hearing loss as the subject.

Both comparisons are defensible, depending on the question being answered. The former seems better suited to assessments of hearing impairment, whereas the latter may be better for the purposes of diagnosis and prognosis. These latter types of applications are more numerous and, in most clinical applications, would seem more useful because comparisons with hearing-impaired persons having the same hearing loss serve to identify the hearing-impaired subject as unusual among those with the same hearing loss, not simply that the subject does not have normal hearing. Comparisons with the speech scores of normal hearers are advocated by Halpin et al (1994) and Hou and Thornton (1994) but seem redundant in most clinical applications as it is already known whether the subject has normal hearing from the pure-tone audiogram. Thus, in the present project, all comparisons were between the subject’s actual performance and the performance predicted for an average person with the same...
pure-tone hearing loss as the subject. Inclusion of the HLD factor in STEPS is consistent with this approach.

**Statistical Evaluation of Performance Differences.** A major problem in making statistical comparisons of speech recognition scores arises from the fact that the scores are percentages. Percentages have the unfortunate statistical characteristic that the standard deviations of groups of such scores are strongly correlated with their mean value (Studebaker et al, 1995). This relationship violates one of the most important assumptions underlying the use of parametric statistics: the homogeneity of variance.

The arcsine transform has been used for many years to overcome this problem (Bartlett, 1947; Shore et al, 1960). The advantage of the arcsine transform is that the standard deviations of groups of such scores change little with their mean value. It is, therefore, better suited for comparisons between performance scores than percentages. Unfortunately, the arcsine transform has the problem that its numeric values bear no apparent relation to conventional performance scores. This makes them difficult to interpret. The rationalized arcsine transform (Studebaker, 1985) was devised to overcome this problem. The units of this transform are referred to as rationalized arcsine units (rau). Rau have the look and feel of percentage scores and are nearly numerically equal to percentages between 15 percent and 85 percent. Therefore, they can be normally interpreted like percentages. Because they are linear transforms of the arcsine transform, they retain the statistical characteristic of homogeneity of variance and are thus more suitable for evaluating relative speech recognition scores than percentage scores. For a more extensive discussion of these issues, see Studebaker et al (1995).

**CLINICAL DATA COLLECTION**

**Purpose**

Essential information for the application of predicted word scores is knowledge of the location, size, and shape of the intersubject distribution of obtained scores around individually predicted scores. This information was not available in the literature. To obtain an estimate of the distribution, the following study was undertaken.

The major goals of this study were to obtain and statistically describe the intersubject distribution of differences between rau transformed observed and predicted speech recognition scores (the rau difference score) for a large and diverse group of regular clinical subjects. Subgoals were to refine our procedures for data acquisition, transmission, and analysis, as it was assumed that these procedures would continue in use after the conclusion of this project.

**Method**

**Data**

Our starting data set consisted of 4064 hearing test results (ears) from 1997 individuals. This number included some persons who were tested more than once and a few who had results for only one ear. From this data set, 410 ears were excluded from the evaluation for the following reasons.

Our goal was to evaluate the distribution of rau difference scores among subjects with sensorineural hearing loss, excluding those with possible retrocochlear pathology. Thus, ears with conductive hearing losses and the poorer ear of persons with asymmetric rau difference scores (defined as a rau difference score in one ear more than 30 rau worse than that in the better ear) would be excluded. These rules resulted in the exclusion of 215 ears with conductive losses and 17 ears with asymmetric rau difference scores.

Also excluded were results that were likely clinical or clerical errors and those where there was a suspicion of subject exaggeration. Thus, two ears were excluded for apparent masking problems and 11 ears were excluded because of "strange or missing data." Most of these 11 exclusions were the result of a missing or an obviously incorrectly recorded speech test presentation level. Nineteen ears were deleted because the results appeared "questionable." In most of these cases, this meant that there was poor agreement between the SRT and the pure-tone average. Finally, 146 tests were excluded because they were earlier tests on the same ear. Thereby, only the single latest result for each ear of each subject was used for this analysis.

---

360
After these exclusions, the data set included results for 3654 ears. Of these, 2609 results were based on 50-word tests and 1045 were based on 25-word tests. All but one of the analyses carried out for this project were based solely on the 50-word test results.

Procedures

Data were collected in three different audiology clinics: a community speech and hearing center (Memphis Speech and Hearing Center [MSHCl]), a veterans hospital audiology clinic (Memphis Veterans Administration Hospital [MVAH]), and a medical center clinic (University of Maryland Medical System-Baltimore [UMMS]). All speech tests were administered using the CD version of the Auditec NU-6 monosyllabic word test. The CD players were under computer control. The outputs of the CD players were delivered to the subjects via standard audiometers, in quiet, using supra-aural audiometric earphones (TDH-49 or TDH-50). Certified audiologists who were regular employees of the clinics carried out most tests. Students under supervision carried out some tests. All tests were conducted as part of the standard audiological evaluation of each subject.

In all clinics, the STEPS presentation and evaluation system was used in conjunction with an audiogram data retrieval program. STEPS reads each subject's speech recognition threshold (SRT) from the data file produced by these programs.

The clinician was required only to enter either the sensation level (SL) or the presentation level (PL) in dB HTL into STEPS via the computer keyboard. If SL was entered, the PL was calculated or vice versa. We were not involved in any way in determining the PLs used by the clinicians. STEPS took this information into account in determining audibility and, thereby, predicted score.

The STEPS software is a subset of routines derived from a larger, more general-purpose speech score prediction program that had been stripped of those components not needed for this trial (e.g., masking spread, reverberation). To this subset were added modules to (1) automatically read the subject's pure-tone audiogram and SRT from a computer data file, (2) control word presentation from a CD player, and (3) provide the examiner with summary information and graphic results via a computer monitor.

The subject's verbal responses were assessed by the examiner listening to the subject through the audiometer's talk-back system. The examiner scored the responses by pushing the left (correct) or right (incorrect) mouse button. The program tallied each button click and triggered the next word to be presented immediately. Thus, the test proceeded as fast as the subject's responses allowed, and the interitem presentation interval could be adjusted to subject needs for each item, as with live-voice testing.

After a speech test was completed, STEPS automatically calculated the subject's score, read the subject's audiogram file, and calculated predicted scores for the subject for all speech test presentation levels from 0 to 105 dB HL in 1-dB or larger steps, depending on computer speed. Based on these results, and on information read from the subject's audiogram file, STEPS generated a computer screen like that shown in Figure 1. Included were the patient's identifying information, the hearing test results, results of the statistical evaluation, and a list of relevant test conditions.

Using preliminary estimates of variability, the program provided numeric and verbal descriptions of the statistical likelihood that the obtained result came from the target population (i.e., the population of persons with pure-tone audiograms like that of the subject). As shown in Figure 1, two graphs were also provided to the audiologist. The graph on the left depicts the predicted performance-intensity (PI) function for the subject. The subject's obtained score was plotted as a diamond at the PL used and a vertical 95 percent confidence interval bar around the predicted score was plotted at that same level. If previous results are available in the record for this ear, they, and the associated error bars, are shown also in a subdued color.

The second graph shown was on the right side of the screen, as depicted in Figure 1. It showed the subject's audiogram for the test ear plotted in HL and the auditory area occupied by the speech signal, also in HL, at the presentation level used in testing the subject. The speech area was depicted as divided by the audiogram into audible and inaudible parts. Also reported on the screen were the WAI and the predicted and observed scores expressed in both rau and percentages.

The graphs shown in Figure 1 are similar to some that have been used for a number of years at the Massachusetts Eye and Ear Infirmary (MEEI). Examples of portions of those graphs may be found in papers by Halpin et al (1994, 1996). In spite of this similarity, it should be noted that significant differences exist.
between the methods used at the MEEI and those described here. Most notable are those that relate to the nature of the person with whom the subject is compared (normal vs hearing impaired), how the results are reported (rau vs percentage), and the statistical assessment of the results (likelihood statements). Other relatively minor differences include the speech test used (W-22 vs NU-6), the FIF used, the IIF used, etc. (For an overview of the methods used at the MEEI, see Halpin et al, 1996.)

As noted, the probability values shown to the examiner at the time of the test were based on preliminary estimates. Those results were not used in the analyses reported in this paper. Because the exact predicted scores reported to the audiologists depended to a degree on some options that can be changed in a set-up screen, the data were all reprocessed for this report using a single set of prediction constants and set-up. This ruled out the possibility that some of the differences between the subjects, and/or the clinics, might reflect variations in these settings.

In this connection, it should be noted that while it would be possible to adjust these optional settings to obtain the best possible results, this was not done. All of the procedures and constants used in predicting the scores of the subjects included in this analysis were determined prior to this project, as discussed earlier in this paper.

The results from the batch-processed data were transferred into a Paradox database file for sorting and preliminary evaluation. Additional analyses were then carried out using Lotus-123 (Lotus Development Corporation, 1994) and TableCurve 2D. Except where noted otherwise, all analyses used only the 50-word test scores.

Results

Subject Characteristics

Table 1 reports the numbers of subjects and ears, mean age, age range, gender, range of hearing losses, mean, median, range, and standard deviation of the pure-tone hearing losses.
Table 1  Numbers of Subjects and Ears, Gender Breakdown, Age Ranges, Mean and Median Ages, Hearing Loss Range, Mean and Median Hearing Loss for Each of the Three Participating Clinics and for the Combined Data Set

<table>
<thead>
<tr>
<th>Clinic</th>
<th>UMMS</th>
<th>MSHC</th>
<th>MVAH</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number: subjects/ears</td>
<td>271/514</td>
<td>268/490</td>
<td>858/1605</td>
<td>1397/2609</td>
</tr>
<tr>
<td>Proportion male</td>
<td>54.9%</td>
<td>53.1%</td>
<td>97.6%</td>
<td>80.7%</td>
</tr>
<tr>
<td>Age range (yr)</td>
<td>9-88</td>
<td>8-92</td>
<td>19-90</td>
<td>8-92</td>
</tr>
<tr>
<td>Mean/median age (yr)</td>
<td>55.1/59.0</td>
<td>61.4/69.0</td>
<td>63.2/65.0</td>
<td>61.2/65.0</td>
</tr>
<tr>
<td>SD age (yr)</td>
<td>18.24</td>
<td>21.61</td>
<td>12.01</td>
<td>15.86</td>
</tr>
<tr>
<td>IWHL range (dB)</td>
<td>0.4-86.1</td>
<td>0.9-97.5</td>
<td>0.9-97.6</td>
<td>0.4-97.6</td>
</tr>
<tr>
<td>Mean/median IWHL (dB)</td>
<td>28.9/26.3</td>
<td>42.6/45.0</td>
<td>38.0/38.3</td>
<td>37.1/37.7</td>
</tr>
<tr>
<td>SD IWHL (dB)</td>
<td>18.59</td>
<td>18.56</td>
<td>16.08</td>
<td>17.65</td>
</tr>
</tbody>
</table>

These results are for the 50-word tests.

IWHL = importance weighted hearing loss. See footnote 3.
UMMS = University of Maryland Medical System, MSHC = Memphis Speech and Hearing Center, MVAH = Memphis Veterans Administration Hospital.

included in the subject sample for the 50-word tests from each clinic and for the combined data. It can be seen that the MVAH clinic provided the most data (62%) and that this part of the data set came predominantly from male subjects. The other two clinics were more similar to each other in that (1) they each had about equal numbers of male and female subjects, (2) the average subject was somewhat younger than at the MVAH, and (3) they included a broader range of subject ages.

The extent of the pure-tone hearing threshold loss for each subject was described using the importance-weighted hearing loss (IWHL)\(^3\). The mean IWHL values reported in Table 1 indicate that the average hearing loss was greatest at the community speech and hearing center (MSHC) and least at the medical center clinic (UMMS). The intersubject variability of the IWHL was about equal at these two clinics and, in both cases, somewhat larger than at the MVAH.

\(^3\)The IWHL is the sum of the products of the pure-tone hearing thresholds in each 1/3-octave band within the normal audiometric test range (some of which are interpolated) multiplied by the importance of that band for recognition of the test material. The IWHL takes into account the entire audiogram while giving greatest weight to those parts of the spectrum that are of greatest importance for speech recognition. (The classical three-frequency average, in effect, assigns weights of 0.333... to the thresholds at 500, 1000, and 2000 Hz and 0.0 to those at all other frequencies.) For a flat hearing loss, the IWHL is exactly equal to the average pure-tone threshold. For a sloping loss, it can be greater or less than the arithmetic mean threshold, depending on the relative importance for speech recognition of the frequencies with the greatest hearing loss.

Absolute Percentage Scores

Table 2 and Figure 2 report on the characteristics of the frequency distribution of the 2609 absolute percentage scores for the 50-word tests. Table 2 gives the means, medians, standard deviations, and skewness values of these scores by clinic for each ear separately, for both ears combined, and for the data combined across clinics. There were very nearly equal numbers of right and left ears in the case of each clinic. In all cases, the results for the two individual ears were nearly identical to each other and to those for the two ears combined. The results for the three clinics differed somewhat from each other in that the average score was about 10 percent lower at the MSHC than at the UMMS and in that the intersubject variability was larger at the MSHC than at the other clinics. Also, the skewness of the MSHC data was less than for the other clinics.

Based on these findings, the absolute score data were combined across ears to produce frequency distribution histograms for each of the three clinics and for the combined data. They are shown in Figure 2. To improve the clarity of the overall trends, the data were grouped into 4 percent intervals. Thus, the bar labeled “99” represents the scores of 100 percent and 98 percent, combined, etc. The ordinates in each panel are scaled as a percentage to assist in the comparison across clinics and with the combined results. Each panel also includes the N (ears) upon which that distribution is based.

The most prominent feature of these distributions is that they are strongly negatively
Table 2  Summary Statistics Describing the Absolute Percentage Scores Obtained for Each Ear (R and L) at the Three Clinics and for the Data Combined across Clinics and for Both Ears (B)

<table>
<thead>
<tr>
<th></th>
<th>UMMS</th>
<th>MSHC</th>
<th>MVAH</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>L</td>
<td>B</td>
<td>R</td>
</tr>
<tr>
<td>Mean</td>
<td>88.6</td>
<td>87.3</td>
<td>87.9</td>
<td>76.4</td>
</tr>
<tr>
<td>Median</td>
<td>94.0</td>
<td>94.0</td>
<td>94.0</td>
<td>84.0</td>
</tr>
<tr>
<td>SD</td>
<td>16.5</td>
<td>16.5</td>
<td>16.5</td>
<td>21.4</td>
</tr>
<tr>
<td>Skewness</td>
<td>-2.4</td>
<td>-1.9</td>
<td>-2.1</td>
<td>-1.2</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>5.7</td>
<td>3.1</td>
<td>4.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

These results are for 50-word tests. Subject numbers are reported in Table 1. Right and left ears were approximately equal in number.

UMMS = University of Maryland Medical System, MSHC = Memphis Speech and Hearing Center, MVAH = Memphis Veterans Administration Hospital.

Figure 2  Proportionate frequency distributions of the absolute speech recognition scores (in % correct) obtained at three clinics and for the data combined across clinics. For these figures, the results were combined across ears.

The data was skewed. This appearance is consistent with the results reported in Table 2, where it is recorded that for all clinics, median scores were higher than the mean scores and all produced negative skewness values. Skewness for the combined data was -1.65. The distributions also suggest that the likely origins of the large intersubject variability values reported in Table 2 are the relatively small number of subjects with speech scores that extend far below the mean value.
Also, it is noted that there is a substantial positive kurtosis in the data of two clinics but little or none in the data of the third.4

**Rau Difference Scores**

The frequency distributions shown in Figure 3 are based on the same data as in Figure 2. However, in this case, the scores were processed to optimize the comparison between predicted and observed scores, as discussed in this paper. First, each observed speech test score was transformed into a rau value. Then, an expected score (in rau) for that subject was calculated, taking into consideration the test material, subject's audiogram, PL, and earphone employed. The expected score was then subtracted from the observed score to produce the rau difference value. The frequency distributions of the rau difference scores for each clinic and for the combined clinics are reported in Figure 3. To facilitate evaluation, each figure also includes a normal distribution function with the same means, standard deviations, and area under the curve as the data functions. As for Figure 2, the ordinates are scaled in percentage and the N (ears) values are reported in each panel.

Table 3 reports the descriptive statistics for the distributions in Figure 3. As for the percentage scores in Table 2, values are reported individually for the two ears and the combined ears for each of the three clinics, as well as for the data combined across clinics. Again, the two

---

4 Due to the large N, all but one of the skewness and kurtosis values reported in Tables 2 and 3 are statistically significant at the 1 per cent level of confidence.
Table 3  Summary Statistics Describing the Rau Difference Data Obtained for Each Ear (R and L) at the Three Clinics and for the Data Combined across Clinics and for Both Ears (B)

<table>
<thead>
<tr>
<th></th>
<th>UMMS</th>
<th></th>
<th>MSHC</th>
<th></th>
<th>MVAH</th>
<th></th>
<th>Combined</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>L</td>
<td>B</td>
<td>R</td>
<td>L</td>
<td>B</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>Mean</td>
<td>-2.4</td>
<td>-2.0</td>
<td>-2.2</td>
<td>-5.7</td>
<td>-4.7</td>
<td>-5.2</td>
<td>-4.1</td>
<td>-4.2</td>
</tr>
<tr>
<td>Median</td>
<td>-1.3</td>
<td>-1.3</td>
<td>-1.3</td>
<td>-4.3</td>
<td>-3.6</td>
<td>-3.9</td>
<td>-3.6</td>
<td>-3.7</td>
</tr>
<tr>
<td>SD</td>
<td>12.2</td>
<td>13.6</td>
<td>12.9</td>
<td>15.6</td>
<td>15.7</td>
<td>15.7</td>
<td>12.1</td>
<td>12.7</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.84</td>
<td>-0.95</td>
<td>-0.90</td>
<td>-0.41</td>
<td>-0.46</td>
<td>-0.44</td>
<td>-0.09</td>
<td>-0.42</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.4</td>
<td>2.3</td>
<td>2.4</td>
<td>0.9</td>
<td>1.2</td>
<td>1.1</td>
<td>2.2</td>
<td>1.9</td>
</tr>
</tbody>
</table>

These results are for 50-word tests. Subject numbers are reported in Table 1. Right and left ears were approximately equal in number.

UMMS = University of Maryland Medical System, MSHC = Memphis Speech and Hearing Center, MVAH = Memphis Veterans Administration Hospital.

ears produced results nearly identical to each other and to those for the ears combined. Thus, only the data for the combined ears are shown graphically in Figure 3.

The distributions of the rau differences for two of the clinics (UMMS and MSHC) contain a prominent artifact. This artifact is the relative excess of rau difference values in the +8 interval and an associated dearth of values in the +4 interval. When evaluating this observation, it was noted that the size of this effect was related to the number of 100 percent scores in the subject data set. The reason for this pattern then became clear. It occurred, in part, because scores between 98 percent and 100 percent are not possible with a 50-item test and, in part, because, under the conditions assumed in this trial, STEPS did not predict performance higher than 98.7 percent. With predicted scores in this region and an observed score of 98 percent, the rau difference fell into the zero-difference bin. However, when a subject produced the next possible higher score (100%), the rau difference fell within the +8 bin. Thus, under this particular set of conditions, a rau difference value in the +4 bin (in the range +2 to +6) was not possible. This artifact quickly disappears as the number of test items contributing to a score is increased above 50. Further, the effect on results is unlikely to be significant in individual cases because it occurs only when excellent performance is predicted (about 98%) and the subject scores 100%. For these reasons, it is not considered a significant practical problem.

Table 3 reveals that the rau differences across the three clinics differed less than the absolute percentage scores. Also, the skewness values are smaller and more consistent across clinics. Finally, the intersubject standard deviations of the rau difference scores are smaller than those of the absolute percentage scores, although, obviously, they are still not small.

**Age Effects**

Recall that the prediction scheme used in this project did not include a correction for subject age. Therefore, the rau difference values provide a basis for assessing the effects of subject age on speech recognition while controlling for the effects of audibility loss. Figure 4 shows the mean rau difference scores for the combined clinics data obtained in this study plotted as a function of age (filled circles). To produce this plot, the subject data were first grouped into decade ranges of age. The mean rau difference and mean age of the subjects in each group were then calculated. In addition to the mean values, Figure 4 reports the N (ears) contributing to each plotted point. Results for subjects under 10 years and over 90 years of age were not included because subject numbers in these decades were very small and all available ages were near the edge of the age interval. (The data represented by the open circles are from a different study that...
will be compared to the present results in the discussion section.)

Effects of Test Length

During approximately the second half of the trial period, one clinic (MVAH) tested essentially all of its clients using 25-word tests. Selection criteria and all other procedural aspects appeared to be the same during the two periods. Table 4 descriptively compares the characteristics of the absolute percentage and rau difference distributions obtained with 25-word and 50-word test lengths in this one clinic. The results in Table 4 are based on 1605 50-word tests and 932 25-word tests on different subjects. (A few 25-word tests were carried out in the other clinics, but too few for analysis.)

Table 4 reveals that the mean and median values were nearly equal across trial periods, suggesting that the average characteristics of the subjects during the two test periods were essentially the same. Of interest is that the standard deviations associated with the 25-word tests are only slightly larger than for the 50-word tests. This will be discussed in the next section.

Discussion

Score Distributions

There are a number of difficulties in using the distributions of absolute percentage scores, such as those shown in Figure 2, for evaluating speech recognition scores. The commonly used normal distribution function would be unsuited as a basis for probability statements because the actual distribution is highly asymmetric. Also, as noted earlier, percentage score variability is highly correlated with mean score (Thornton and Raffin, 1978; Studebaker et al, 1995). This is a particularly important problem when many scores are at or above 90 percent, as is clearly the case with the usual run of clinical subjects.

It might be possible to develop a reasonably valid speech score probability function, based on curves like those in Figure 2, and/or a similar one for normal-hearing subjects, but significant difficulties would remain. One of the most important would be that it would not be possible even then to compare a subject's performance with that of an average hearing-impaired subject with the same hearing loss.

The frequency distributions of rau difference scores, as shown in Figure 3, are substantially more symmetric and normal in appearance than those for the percentage scores. It can be seen more clearly by comparisons with the normal curves superimposed on the data in Figure 3 that the distributions are somewhat negatively skewed and that they are somewhat leptokurtic as well. Both observations are consistent with the kurtosis and skewness values reported in Tables 3 and 4. These observations are consistent with those of Schum et al (1991) for their subjects listening in quiet, as here.

Figure 3 indicates that the skewness values are likely the result of the larger than normal number of scores appearing in the far left-hand tail of the distribution. In general, this seems an expected outcome for such samples. In fact, the figure indicates a degree of negative skewness that is less than we expected. When evaluating this result, it should be recalled that no data were eliminated simply because the score was low, with the exception of the 17 ears that were eliminated from the sample because the rau difference score for each of those ears was more than 30 raus below that of the subject's opposite ear. Inclusion of these few cases might have increased the negative skewness, but probably only slightly.

Although positive kurtosis is a feature of the rau difference data, it is not quite so large or so variable across clinics as it was for the percentage data. The MVAH data suggest that kurtosis was somewhat more positive with 50-word tests. Inspection indicates that the positive kurtosis is the result of a relative abundance of results in the region between -12.5 and 2.5 rau difference and a relative dearth of results both just above and just below these intervals.

In comparing the data with the normal curve, we concluded that the normal distribution provides a reasonable basis for evaluating rau difference scores for the following reasons. First, the modest excess of score differences in the negative tail likely represents the effects of
unusual pathology. Using a symmetric distribution assumption in the calculation of the statistical likelihood makes these results stand out, as they should. Second, the observed positive kurtosis normally reflects a relatively reduced number of scores about one standard deviation from the mean and a relative abundance of scores nearer the mean, as in this case. However, because the effect is no more than moderate (see Fig. 3) and the distribution of scores relatively near to the mean usually is of little interest, it is unlikely that the observed leptokurtosis would have significant practical consequences.

**Average Performance**

As reported in Table 3, average scores were about 3 (median) or 4 (mean) rau below predictions. Recall that no age correction was applied in making the predictions. Figure 4 shows the relationship between age and average rau difference scores in this sample. Note that the average difference scores for the younger subjects are near zero and that the rau difference scores became gradually more negative with increasing age. This result suggests that the average prediction provided by STEPS is very close to the correct value for young hearing-impaired subjects, as intended, and that a negative difference for the group as a whole is expected in the absence of an age correction.

**Age Effects**

Inclusion of an age correction would reduce the downward tilt of the rau difference function over age, move the negative mean rau difference for the group as a whole closer to zero, and reduce the intersubject variability. However, because the relationship of word recognition scores to age is small, these effects would also be small. In any case, such a correction should be made optional to allow the clinician to compare a subject's result with that of a young hearing-impaired person (no age correction) or with a hearing-impaired person of the same age (with age correction), whichever the clinician deems appropriate for a particular application.

It should be noted that the age effect seen in this study, although not large, is nonetheless larger than that observed earlier by Studebaker et al (1997) (open circles, Fig. 4). In the earlier study, the subjects were confined to those with pure-tone thresholds of 25 dB or less over the frequency region of the speech signal used in the tests. The results from that study are compared to those of the present study in Figure 4. The difference between the two studies may reflect an interaction between hearing loss status, speech recognition loss, and the effects of age. However, it is apparent from Figure 4 that downward trend over age is similar in the two studies and that the differences are not large in absolute terms. The present results support the conclusion of the earlier study that the effects of age on monosyllabic word recognition scores are small (<5 rau), after audibility effects are taken into account, at least into the sixth decade of life.

**Intersubject Variability**

It is evident that the intersubject standard deviations of the rau differences observed in this sample are large. In some cases, they were only modestly smaller than those of the original absolute percentage score values. There are several factors that should be borne in mind about this. First, except for the exclusion of conductive hearing losses, 17 highly asymmetric losses, and a few probably invalid results, the data represent a nearly unrestricted sample of persons with sensorineural hearing loss who present themselves at audiology clinics. Thus, smaller values would be anticipated for most subgroups, including, for example, groups categorized by age and/or etiology.

Second, the larger than normal number of very negative rau difference scores contributed substantially to the obtained rau difference standard deviation values. This last point is demonstrated by the fact that if the distribution of the combined rau difference data to the left of the distribution of scores relatively near to the mean usually is of little interest, it is unlikely that the observed leptokurtosis would have significant practical consequences.

---

*5 A post hoc analysis of the data revealed that the mean of the age-corrected rau differences was essentially zero. This means little, of course, as the correction was based on the same data. However, it is noteworthy that the intersubject variability values were decreased only very slightly (≈0.2 rau) by this correction or the one reported by Studebaker et al (1997). This suggests that the intersubject variability introduced by a range of ages is small compared to other factors that cause subjects to differ from each other and that the distributions and variability values reported in Figures 2 and 3 and Tables 2 and 3 would not have been noticeably different if an age correction had been used.

6 We estimated an even larger standard deviation value (≈17 rau) for a comparable distribution (Fig. 7 "In Quiet") reported by Schum et al (1991). However, due to the relatively small subject numbers (101), this result was dramatically influenced by a few, very negative rau difference values.
mean were the mirror image of that actually observed to the right of the mean, the resulting standard deviation of the distribution would be 9.4 rau. Thus, it can be argued that the intersubject standard deviation of hearing-impaired subjects is in the region of 9 to 10 rau, except for that added variability created by the occurrence of a relatively small number of very negative rau difference scores. Values in this range are still somewhat larger than comparable values observed for normal-hearing subjects (Studebaker et al., 1997).

Table 4 reveals that the intersubject standard deviations were only slightly larger when 25-word tests were used. This observation might be taken as evidence favoring the use of 25-word tests. This, however, would be an incorrect conclusion. The major effect of increasing test length is to increase the stability of the individual test score. It does not decrease true intersubject differences, the differences that are likely the major factor controlling the standard deviation values reported here. When multiple sources of variability are combined, the largest source of variability dominates the overall result. Therefore, the small list length effect on the standard deviation observed in this study is an expected outcome. It does not contradict the findings of Thornton and Raffin (1978) and Dillon (1982), for example, which indicate that shortened test lengths can have substantial negative consequences for reliability.

Choosing a Standard Deviation

The question of what standard deviation to use when making a statistical evaluation of a particular clinical result is an important one, but one for which there is no definitive answer. The question is a variation on the more general “Who do you compare to?” question discussed earlier. Some theoretical arguments would seem to favor using the standard deviation value for young normal-hearing subjects, that is, about 7 rau. However, practical considerations favor using values based on the variability of hearing-impaired subjects. This is because most clinicians are concerned with whether a patient is unusual in his or her category. As noted earlier, the audiogram already tells them that the subject does not have normal hearing.

Also, if the intersubject standard deviation for normal-hearing subjects was used, the resulting probability values associated with clinical results would indicate that relatively large proportions of hearing-impaired subjects are very unusual. It would soon become evident to the clinician that such results were actually not that unusual and so would soon be ignored. Thus, the use of a standard deviation that identifies a subject as unusual, when indeed the subject’s speech score is unusual among those with similar hearing losses, would seem to be a more useful approach.

Another defensible conclusion is to use a standard deviation value around 9 or 10 rau on the grounds that this value represents the intersubject variability of the hearing-impaired subjects, excluding only those with conditions that produce particularly negative rau difference scores. As these are the very subjects we often want to identify for further study, the smaller standard deviation value and the use of the symmetric normal curve would help ensure that these low scoring individuals would be identified as unusual. But, whatever value is used in the calculation of the probability value, clinicians would soon learn which values are actually worthy of further attention. For this reason, the exact standard deviation value used is less important than a system that provides clinicians with a consistent and objective framework within which to better evaluate their speech recognition test results. It is concluded that a system similar to that demonstrated in this project can meet that goal.

Acknowledgment. This work was supported in part by research grant R01 DC00154 from the National Institute on Deafness and Other Communication Disorders and by the Center for Research Initiatives and Strategies for the Communicatively Impaired. Portions of this paper were presented at the 1997 meeting of the American Academy of Audiology. The authors want to thank Craig C. Formby, Sarah Ferguson, LaGuinn P. Sherlock, Kay M. Pusakulich, Izell M. Rivera, Allison W. Lane, Karen C. Brown, and Virginia A. Hull for their help in the data collection; Robert L. Sherbecoe and Catherine A. Gwaltney for their help in preparation of the manuscript; and Joseph Matesich and Edward Brainerd for their help in writing the computer programs.

REFERENCES

Speech Intelligibility Index. (ANSI S3.5-1997.) New York: ANSI.


